Naturalistic Speech Misperception

Kevin Tang

A dissertation submitted for the degree of

Doctor of Philosophy

of

University College London (UCL).

Department of Linguistics

University College London

2015

Declaration

I, Kevin Tang confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis presents a new corpus containing ≈ 5,000 instances of naturally occurring misperception of conversational English, which is the result of a standardised format for the orthographic and phonetic transcriptions and meta-data of existing naturalistic corpora.

I examined top-down phonetic/phonological factors and bottom-up lexical factors for their contributions in naturalistic settings. On the feature level, voicing/place/manner confusions were best explained using sonority, featural underspecification (Lahiri and Reetz, 2002) and markedness (Lombardi, 2002), and vowel height/backness confusions using perceived similarity (Steriade, 2001) and chain shifts (Labov, 1994a).

On the segment level, I found that confusions can be explained with acoustic/featural distances, and extreme signal-to-noise ratio and narrow bandwidth were less ecologically valid. Furthermore, three well-known sound changes (TH-fronting, velar nasal fronting and back vowel fronting) were consistently found in naturalistic and experimental data.

On the syllable level, codas are more likely to be misperceived than nuclei/onsets for monosyllables, but onsets are more likely to be misperceived for polysyllables. Fewer errors occur in the stressed syllables than in unstressed syllables in polysyllabic words, but not monosyllables. Initial syllables are more likely to be misperceived than medial syllables, which in turn are more prone to misperception than final syllables.

On the word level, listeners were found to perceive a word of similar frequency as the intended word in a misperception – but crucially not a more frequent word. This supports the graceful degradation account of a malfunctioning system (Vitevitch, 2002). On the utterance level, listeners were sensitive to the predictability of a word, suggesting that less predictable words are more likely to be misperceived.

Together, these analyses establish the naturalistic corpus as an ecologically valid resource and a benchmark of misperception, bridge the gap between experimental and naturalistic studies, and highlight the need of examining misperception with units larger than nonsense syllables.

Contents

List of Figures

List of Tables

Acknowledgements

I am eternally grateful to my supervisor, Andrew Nevins, for his constant support and ideas with this work, above and beyond his duty as a supervisor. Over the years, Andrew has trained me to be an independent researcher. Of all the skills that he has bestowed upon me, I particularly appreciate the ability to constantly expand the scope of one's research, the importance of thinking big and stepping outside one's comfort zone. Not any less important, my secondary supervisor, Stuart Rosen, has given me help with experimental techniques used in speech processing, constantly looking out for relevant papers and articles as well as introducing me other researchers, such as Martin Cooke and Gaston Hilkhuysen, whom I could learn from. The work presented here has benefited from their input. I am indebted to them for their help and advice.

I must thank Anne Cutler who has inspired me to work on naturalistic misperception data and whose work has showed me the wonders of psycholinguistics; Mark Huckvale for the idea of using a language model to predict errors on a sentence-level; Kyle Gorman for his generous advice and discussion on the Philadelphia lexical set that was used to transcribe part of Labov's naturalistic corpus; Jyrki Tuomainen and Michael Becker for sharing with me the joy of statistical analyses.

Over the year, I have benefited greatly as a researcher from extensively interacting with Paweł Mandera, Emmanuel Keuleers, John Harris, Michael Becker, Cilene Rodrigues, Judit Druks, Jyrki Tuomainen, Outi Tuomainen, Peter Howell, Brent de

Chene, Sam Green, Ye Tian, Shanti Ulfsbjorninn, Jieun Bark, Giorgos Markopoulos and many others.

I would like to thank those who have shared their data with me – William Labov for his misunderstanding corpus; Paweł Mandera and Emmanuel Keuleers for the SUBTLEX data; Peter Graff for the tabulated versions of the confusion matrices in Miller and Nicely (1955) and Wang and Bilger (1973); Jont B. Allen for the data in Phatak and Allen (2007); and Wafaa Alshangiti, Tim Schoof and Melanie Pinet for their speech-in-noise stimuli. Special thanks goes to my Chinese collaborators – Yan Lou, Shi Zeng, Henry Lam, Huang Puming and many others – for their collaborative work on collecting naturalistic and experimental misperception data of Mandarin Chinese.

Furthermore, I would like to thank the audiences at the 43rd Meeting of the North East Linguistic Society, 20th Manchester Phonology Meeting and Old World Conference in Phonology XI; the regular attendants of Professor Andrew Nevins's lab group (LLSD); Moira Yip and Katrin Skoruppa for their comments, suggestions, criticisms on earlier versions of my thesis.

I would like to thank all those from the department of Linguistics (especially Ye Tian, Sam Green, Nick Neasom, Thanasis Soultatis, Ya-fang Lu, Diana Mazzarella, Maria Varkanitsa and Matthew Gotham) and the department of Speech, Hearing and Phonetic Sciences (especially Mauricio Figueroa, Joaquin Atria, Melanie Pinet, Sonia Granlund, Cristiane Hsu, Yasuaki Shinohara and Tim Schoof) at University College London for their support both as researchers and as friends. I also would like to thank Steve Nevard for his technical expertise and for being so generous with his time.

Last but not least, I would like to thank Anny King and Christopher Hicks from the University of Cambridge for their encouragement during my undergraduate degree, my friends and family for their support, and the Art and Humanities Research

Council in the UK for funding my PhD.

Chapter 1

Introduction

This thesis presents a new corpus of naturalistic misperception, the largest corpus of its kind to the author's knowledge. Using this new corpus, I present a series of analyses showing how naturalistic misperception is influenced by factors from the lowest level of features (e.g. more sonorous manners are more robust in noise) to the highest level of conditional probabilities of words in an utterance (e.g. unpredictable words are harder to perceive correctly). Crucially, most of these factors are above and beyond what classic experimental confusion studies, which focus on the level of features, such as Miller and Nicely (1955) and Wang and Bilger (1973) can identify. In addition, I compare segmental confusions from naturalistic data with those from experimental data in order to establish the ecological validity of experimental controls.

These analyses have multiple intended outcomes. The first is to provide a naturalistic corpus that could serve as an ecologically valid resource and benchmark of misperception in order to bridge the gap between experimental and naturalistic studies. This could be accomplished by identifying similarities within the new corpus to existing experimental misperception data and by examining whether the confusion patterns can be explained using theories from psycholinguistics, phonetics and

phonology. Secondly, they highlight the need for examining misperception with units larger than nonsense syllables (cf. Miller and Nicely, 1955; Wang and Bilger, 1973). This can be accomplished by identifying theoretically informed confusion patterns, in units larger than nonsense syllables.

Overall, this thesis highlights the importance of understanding a common speech misperception phenomenon that has a direct impact on everyday communication. Most of the time, speakers and listeners can utilise their knowledge of language with minimal effort; however, speakers sometimes would make errors in production and/or listeners make errors in perception. This study focuses on the errors of perception, that is, when the speaker's utterance does not match what the listener hears. When communication operates smoothly, one can only assume that the listener's understanding is identical to the utterance of the speaker, but it is only when communication breaks down that we are able to disentangle what was said from what was received, as outlined in Laver (1970, pp. 61): "The strategy of inferring properties of control systems from their output can be applied not only to the efficient operation of these systems, but also to their output when malfunctions occur. The evidence from characteristic malfunctions is more penetrating than that obtained when the system is operating efficiently." Therefore speech *mis*perception can provide us with new insights about the mechanisms that underlie speech perception.

The key question is where should we look for these errors. Perceptual errors can be found both in the laboratory (henceforth experimental misperception) and in our everyday life (henceforth naturalistic misperception).

Data of experimental misperception has so far been collected by presenting participants with stimuli (syllables, words, sentences) that are artificially degraded, and the participants are then asked to repeat what they have heard (Miller and Nicely, 1955; Felty et al., 2013). Data of naturalistic misperception are collected by people who are usually the interlocutors (the listener or the speaker) of a conversation in

which the errors occur. These errors are simply noted down and later reported to the researchers who are collecting them, raising the question if this approach can accurately and adequately reveal the perceptual mechanisms behind misperceptions.

In speech production, lab speech has often been criticised as being unnatural, overly clear and over planned, compared to spontaneous speech which is rich in various patterns but lacks experimental controls (Xu, 2010). Similar arguments can be made for speech perception. On the one hand, experimental studies of speech perception can be highly unnatural. As stimuli are often presented to participants artificially masked with noise and other distractors, and participants are then required to give a response to what they thought they heard. In real life, when listeners are uncertain about what they heard, they often have the option of asking for clarification, and are not required to understand immediately or give an instant response. Secondly, listeners rarely need to tell others what they heard in real life. Thirdly, in an experimental situation, listeners are not engaged in a conversation and there is no real communicative need. These factors contribute to overgenerating errors. Without prior contexts or communicative need, utterances in isolation are much harder to perceive as there are no top-down factors (such as lexical priming) to aid the perceptual process. If participants are required to give a response when they have none to give, then these guesses may be no more than "noise" and the unnaturalness immediately casts doubt about the ecological validity of the experimental data.

That isn't to say that experimental speech under laboratory conditions cannot contribute to further understanding of speech perception. The stimuli and, to a lesser extent, the listeners in experimental conditions can be carefully selected and controlled for and consistency can be applied to the listening conditions. In contrast, we have no control over the "stimuli" in everyday life as utterances can be of any length and masked with any kind of noise. For research purposes, the demographics of the listeners in the naturalistic data are not always available.

More importantly, there are top-down factors that can influence naturalistic misperception. Studies have shown that the listeners' age (Mirenda and Beukelman, 1987; Nábělek, 1988; Munson et al., 2003), accent (Bradlow and Pisoni, 1999; Pinet and Iverson, 2010), the listening conditions, and the length and linguistic structure of the utterances (Wingfield et al., 1985) can all have an effect on perception. Therefore, the lack of experimental rigour that may arise when using naturalistic data can cast doubt on whether meaningful perception patterns can be identified and are not masked by the interactions between top-down factors. Despite these drawbacks, under *Cognitive Ethology*, a research approach to modelling cognition, I argue that naturalistic misperception is nonetheless preferable, and should serve as a benchmark for speech misperception.

The primary goal of Cognitive Ethology is to understand the functioning of human cognition in the *real* world and is based on the four assumptions listed below *verbatim* (the texts in italics are my own emphasis).

Invariance The dynamics of cognition are, at least in part, contextualized. *Variability in cognitive processing that arises from contextual differences is important to understand.* Only by explaining such variability will meaningful and stable cognitive processes be discovered.

Control Important insights into cognition will be gained when individuals behave in an unconstrained and uncontrolled manner in their natural environments. *The goal is to measure naturally occurring variance rather than the variance that emerges from controlling the system.*

Cognition as a distributed system Cognition is a non-linear systemic process. *Important aspects of cognition will only emerge when embodied individuals are considered as a part of a system that involves their natural environment* (including other individuals).

Subjective reports *Subjective reports provide a direct measure of people's conscious experiences, goals, intentions, and beliefs pertaining to their attentional behaviour in everyday environments.*

These four assumptions of cognitive ethology underline the need to study speech misperception that occurs in real life. Firstly, it is true that there are many factors that can affect naturalistic misperception, and these factors cannot be controlled for. However, it is by understanding these variabilities in context that we may be able to understand the robustness of perceptual processes that are found in the laboratory. Secondly, the initial research focus should be on naturally occurring variability, rather than variability that emerges from controlling the cognitive system, such as adding noise to the stimuli. Thirdly, listeners and their natural environment (with different contexts, noise types etc.) should be considered as part of a system, and this cannot be replaced by putting listeners in laboratory environments. Fourthly, subjective reports (which are an essential part of naturalistic data) of misperception can serve as a direct measure of people's conscious experiences and perceptual behaviour in everyday life. Finally, naturalistic data can be used to generate hypotheses (Moisl, 2007), which can be tested under controlled experimental settings.

The pros and cons of speech misperception within and outside of a laboratory setting were discussed in the previous sections. Let us move on to the following sections. Section 1.1 provides a summary of some of the classic studies of speech misperception in a laboratory setting and Section 1.2 describes important studies wich have taken place outside of a laboratory setting. To better understand whether experimental and naturalistic data can complement each other, Section 1.3 presents arguments for and against naturalistic data. Section 1.4 summarises the review of speech misperception in this chapter. Finally, Section 1.5 outlines how this thesis is organised.

1.1 Speech misperception in the laboratory

The literature of laboratory based speech misperception is extensive, and this section does not aim to summarise everything but rather the key work that is particularly relevant to this thesis. Two particular experimental studies are summarised because they were the earliest large scale misperception studies in laboratory. Given that they are the cornerstone of laboratory misperception, both their methodologies and insights are particularly valuable.

1.1.1 Classic confusion experiments

Two large scale confusion experiments in particular have had a major influence on work in this field, Miller and Nicely (1955) and Wang and Bilger (1973).

Miller and Nicely (1955) examined sixteen English consonants which are about three quarters of the consonant phonemes. The consonants were embedded in a nonsense CV syllable with the /a/ vowel in *father*. The stimuli were frequency distorted by applying low-pass filters with different cut-off levels and masked with noise at different signal-to-noise ratios. The stimuli were then presented to listeners in a forced-choice task with the output in confusion matrices of 16 by 16 tables, showing the successful and unsuccessful perception counts of the consonants. Wang and Bilger (1973) tested a wider range of consonants and also syllable forms – CV and VC nonsense syllables. In their analyses, both studies, and many other researchers after them, utilised the idea of distinctive features and information transfer as a way of explaining how listeners discriminate between different phonemes.

Traditionally, the perception of phonemes was assumed to rely on the identification of distinctive features by the auditory system (Kollmeier, Brand, and Meyer, 2008). These distinctive features are perceptually, acoustically and articulatorily defined and have unary or binary values, such that all consonants can be uniquely

defined by a distinctive set of features. Each of these features is transmitted via its own transmission channel to the listener's ear. The signals were then decoded by the central auditory system. The listener would recognise and combine these features to determine the identity of the phoneme.

Both studies employed the transinformation analysis, which is based on information theory (Attneave, 1959; Shepard, 1972). It quantifies the amount of information by particular features which are successfully transferred to the listeners' perceptual system. Miller and Nicely (1955) analysed five articulatory features, voicing, nasality, affrication, duration, and place of articulation, by breaking down the confusion matrix into five smaller matrices (one for each feature) which represent five transmission channels. They concluded that voicing and nasality were transferred the most, while place was most affected by low pass filters and noise; furthermore, Miller and Nicely (1955) found that these features/channels were relatively independent, which supports the idea that the perception system consists of individual channels rather than a single complex one.

Wang and Bilger (1973), on the other hand, arrived at a different conclusion with transinformation analyses by developing and using a sequential method of partitioning the transmitted information, called sequential information analysis (SINFA). This method aims to partial out the internal redundancy of a feature amongst a set of given features. It first identifies the feature that gives the highest fraction of transfer, and then selects the most important feature from the remaining features, given the previously selected feature(s). The process is then applied iteratively, providing an estimate of the importance of each feature independently. Wang and Bilger (1973) concluded that different sets of features perform equally well in capturing the information transferred and the perceptually important features were inconsistently identified, casting doubt on the idea of natural perceptual features.

1.1.2 Later work on speech misperception

There is no doubt that speech perception studies could benefit from examining the robustness of speech in a wide range of adverse conditions caused by external factors such as reverberation (e.g. depends on the size, shape and materials of the room), environmental noise (e.g. factory noise), competing talkers (e.g. in a cocktail party scenario) and channel distortions (e.g. the transmission system in an auditorium) (Assmann and Summerfield, 2004).

Moreover, a range of internal factors has been extensively investigated, for instance, perceptual adaptation when encountering unusual accents and the effect of talker-listener accent similarity (Pinet, Iverson, and Evans, 2011; Pinet, Iverson, and Huckvale, 2011; Iverson and Pinet, In press) as well as speech perception in typical and atypical populations (Hazan, Messaoud-Galusi, and Rosen, 2013; Green, Faulkner, and Rosen, 2012; Rosen, Adlard, and Lely, 2009).

Much of this work has primarily used speech in noise tests, along with other tests such as text comprehension, connected discourse tracking and sentence verification tasks, to measure global performances. These studies provide an indication of the real world performance of listeners but do not provide detailed analyses of *how* do people make errors.

In fact, many models have been developed to predict overall speech intelligibility in various adverse distortion conditions using different signal parameters, including: 1) Articulation Index (AI) (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Rhebergen, Versfeld, and Dreschler, 2006) - macroscopic models that considers the influence of the frequency content of speech on intelligibility; 2) Speech Intelligibility Index (SII) (ANSI, 1997) which considers the spread of masking, standard speech spectra and the relative importance of frequency bands; 3) Speech transmission index (STI) (International Electrotechnical Commission, 2003) which uses the modulation transfer function of a speech transmission system (Kollmeier, Brand, and

Meyer, 2008).

These models have shown success in predicting the average recognition rate for a given degraded speech sample; however, they do little to inform on the detailed process of speech perception into the detailed process of speech perception. One of the challenges would be to predict listener's responses to individual degraded speech tokens. To meet this challenge, microscopic models were developed, for example 1) Holube and Kollmeier's (1996) microscopic model which is based on a model of the human auditory system and an automatic speech recognition system and 2) Cooke's (2006) glimpsing model which identifies spectro-temporal regions of speech which would be most likely to survive energetic masking (which is when the noise interferes with the speech signal in the acoustic environment (Lidestam, Holgersson, and Moradi, 2014) and others such as Régnier and Allen (2008). These models have shown limited success and are restricted to very small domains, such as nonsense VCV syllables. Perhaps the first to step outside these highly restricted domains was Cooke (2009) which explored the idea of constructing a corpus of consistent confusions on a word level.

1.2 Speech misperception beyond the laboratory

The basic information of naturalistic misperception studies will first be introduced in this section. The backgrounds and research findings of four existing naturalistic misperception studies of conversational speech are then summarised. The last section summarises the nature of naturalistic misperceptions of sung speech (also called *Mondegreens*) and how they may differ from misperceptions of conversational speech.

When trying to gain a better understanding of a phenomenon, the first place to look would be where the phenomenon naturally occurs. No study of speech misperception is complete without an understanding of how inferences are made

by the listener in our everyday life. In fact, the concept of mishearing surfaces in our lives in many ways, such as in misheard song lyrics such as hearing "kiss this guy" instead of "kiss the sky". Not unlike Freud's belief that slips of the tongue can reveal our subconscious thoughts, "slips of the ear" have the potential of revealing our speech perceptual processes.

In fact, naturalistic corpora for production errors do exist. Naturalistic error corpora of slips of the tongue (Fromkin, 1973) and tongue twisters (Shattuck-Hufnagel and Klatt, 1979) have provided insights about the processes underlying fluent speech production. The largest corpus of this type is Fromkin's Speech Error Database (Fromkin, 2000) containing about 9,000 instances of speech errors across English, French, German and Italian. The database was developed by combining multiple speech error corpora from independent researchers. However in speech misperception (sometimes called "slips of the ear"), existing naturalistic data and their analyses are scarce.

The first corpus of naturalistic misperception was compiled by Meringer and Mayer (1895) and Meringer (1908). The corpus consists of 47 instances of German misperception, e.g. *Durst oder Hunger* perceived as *Verdruss oder Kummer*. Celce-Murcia (1980) later added to these first 47 instances. Even with such a small sample, some observations emerged. Firstly, Meringer (1908) found that consonants are misperceived more frequently than vowels. Secondly, Celce-Murcia (1980) observed that many of the instances involved proper names, and the perceived utterances tended to be grammatically correct, but they were inappropriate to the context (in terms of semantics and pragmatics).

The research potential of analysing naturalistic misperception is apparent even from Meringer's small corpus and further motivated error researchers to collect and analyse naturally occurring (naturalistic) errors made by us everyday (Browman, 1980; Cutler and Butterfield, 1992; Bond, 1999; Labov, 1994b; Labov, 2010b).

Speech misperception is generally agreed to be the misperception of an intended speech signal (Bond, 1999). The word "intended" is important here, as speech misperception is not speech misproduction in which the mismatch lies between the intended utterance (which is not actually produced) and the actual utterance (which is produced). In speech misperception, the mismatch lies between the produced utterance (by a speaker) and the perceived utterance (by a listener), whereas in speech misperception there is no mismatch between the intended utterance and the utterance produced by the speaker. The errors come from the perceiver; for example, a speaker intended to produce *doll*, and he successfully produced *doll*, but a hearer might perceive *doll* as *dog*. In general, there are two types of speech misperceptions, those from conversational speech and those from music. In this thesis, unless stated otherwise, all misperceptions are from conversational speech. The nature of Mondegreens is summarised in the last section.

The data used in this thesis were sometimes collected by the researchers themselves, or in some cases by a team of trained phoneticians, but in most cases they are collected by a third party, e.g. friends, family, students etc. The data collected by people with knowledge of phonetic transcription would usually provide the transcriptions of the intended utterance and the perceived utterance. In some cases, detailed demographic information and contexts were also documented.

In the following sections, four existing studies of naturalistic misperception, and the nature of Mondegreens are discussed.

1.2.1 Browman (1980)

Work on speech perception has so far focused on lower level units such as features and phonemes (Miller and Nicely, 1955; Wang and Bilger, 1973), but the higher level units such as syllables, words, and phrases have been less explored. It is this research gap that motivated Browman (1980) to focus on the interaction between higher level

processes such as lexical decisions and lower level processes such as acoustic analysis.

To explore the interaction between lexical decisions and acoustic analysis in speech perception, the author collected 222 misperception instances of American English in order to relate patterns of segmental misperceptions to syllable structures. The analyses first calculated the rate of segmental errors by syllable constituency (onset, nucleus and coda), and stress (stressed and unstressed) for both monosyllabic and polysyllabic words, and also syllable positions (word initial, medial and final) for polysyllabic words. The author then separated the segmental errors into two categories, those containing one featural difference between the intended segment and the perceived segment (e.g. *van* $\rightarrow$ *fan*, "$\rightarrow$" denotes perceived as), and those containing multiple featural differences (e.g. *clean teeth by tonight* $\rightarrow$ *my tea butter knife*). The author assumed that the errors containing one featural difference were the result of acoustic errors, and gross errors were the result of lexical errors.

The distribution of lexical errors in unstressed polysyllabic words correlates with the positional saliency on lexical retrieval. In lexical retrieval, initial and final phonemes are more salient cues than medial phonemes (Horowitz, White, and Atwood, 1968). Fewer lexical errors were found word-initially and word-finally than word medially. The distribution of lexical errors therefore supports the experimental findings that initial and final phonemes are more salient cues for lexical retrieval because they are less susceptible to lexical errors. However the distribution of lexical errors in stressed polysyllabic words and monosyllabic words had an entirely different pattern, and the author did not fully explain this mismatch.

Different distributions of acoustic errors were found. Stressed syllables in polysyllabic words were found to be less frequently misheard than unstressed syllables. This can simply be explained by the fact that stressed syllables are acoustically more salient. However, the reverse pattern was found in monosyllabic words, with stressed syllables being misheard more often. The author explained by referring to a report-

ing bias in the naturalistic data, such that people are less likely to report unstressed monosyllabic words in a misperception because these unstressed syllables are more likely to be function words (e.g. *of*) which have low information content and are therefore less noticeable even when they are misheard. A syllable constituency effect was found with all but stressed monosyllables, such that codas were less erroneous than onsets. The author explained this with reference to acoustic differences between onsets and codas.

Based on these findings, the author proposed that there are three modules involved in speech perception, a lexical decision module, an acoustic analyser, a segment recognizer which overlaps with a lexical decision module. The acoustic analyser and the segment recognizer make different contributions according to different positions within a word, correlating with the error rates at these positions. Therefore, the segment recognizer focuses more on word final and word initial positions than on word medial positions, and the acoustic analysers focuses more on coda and word final positions.

Acoustic information was first fed into the acoustic analyser. The output was then fed into the segmental analyser. The new output was then fed into the lexical decision module, which in turn affected the segment recognizer. The proposed model therefore links acoustic analysis, which is a "low level" process, with lexical decision making which is a "high level" process.

1.2.2 Bird (1998)

Bird (1998) collected 300 hearing errors of British English from naturally occurring conversation. Four analyses were conducted using the data in order to replicate or extend previous analyses of naturalistic misperception.

The first analysis was conducted to replicate the analyses of naturalistic misperception by Garnes and Bond (1980), which found that the stress pattern is almost

always retained, except when unstressed syllables are deleted or inserted.

Of the 300 instances of misperceptions in this study, the stress pattern was retained in 62% of the data, e.g. *bleach it* → *leaches*. 35% of these were instances when the number of stressed syllables was retained but weak syllables were inserted or deleted, e.g. *Have you applied?* → *are you blind?*. Only 3% showed a change in stress pattern, e.g. *I've walked a long way today* → *I've put on a lot of weight*. Bird's (1998) findings using the independent data of misperceptions therefore agreed with Garnes and Bond's (1980) analyses.

The second analysis assessed the conclusion of Miller and Nicely (1955) and Garnes and Bond (1980) that place confusions are more frequent than manner confusions, which in turn are more frequent than voicing confusion. This was again confirmed using the author's data. Counting only the substitution errors (ignoring insertion, deletion and the correctly perceived segments), 40% were place errors, 35% were manner errors, and 25% were voicing errors.

The third analysis examined whether the frequency of a phone as the intended segment in a substitution error correlates with the frequency of the same phone found in the language. Say that /t/ is frequently the intended segment of a substitution error. Could this be because /t/ is a frequent segment in the language? In other words, is there a target bias due to frequency? Similarly, the question arises whether the frequency of a phone being the perceived segment of a substitution error correlates with the frequency of the same phone in the language; in other words, is there a response bias due to frequency? Both correlations were found to be highly significant (R = 0.80 – 0.93) for both consonants and vowels.

The fourth analysis attempted to replicate Cutler and Butterfield's (1992) findings of juncture misperception using naturalistic misperception data. Cutler and Butterfield (1992) found that listeners tended to insert word boundaries before strong syllables, and delete word boundaries before weak syllables. The resultant words

from inserting word boundaries before strong syllables were more likely to be content words (e.g. *analogy* → *and allergy*), and similarly the words with word boundaries before weak syllables tended to be function words (e.g. *effective* → *effect of*). These findings were again confirmed in Bird (1998).

In summary, the author managed to replicate the previous findings of studies that also used naturalistic misperception data. This reinforces the robustness of the patterns examined as well as the stability of the naturalistic misperception data.

1.2.3 Bond (1999)

Bond (1999) presented a corpus with almost 900 instances of misperceptions of mainly American English with 783 instances produced by adults and a subcorpus of 105 misperception instances by children. Overall, the author classified three main types of errors, simple consonant misperceptions, simple vowel misperceptions (both with one segmental difference) and multiple complex misperceptions of more than one segmental difference. About one-third of the instances consisted of simple consonant misperceptions, two-thirds consisted of complex misperceptions and less than 50 instances were simple vowel misperceptions.

Errors were identified by manual categorisation, e.g. deletions, substitutions, and reorderings. The confused segments were not extracted or explicitly analysed; instead, the studies provided an overview of the distribution of different classes of errors at different linguistic levels. The analyses focused on 1) vowel misperceptions and their interaction with stress; 2) consonant misperceptions in simple (one segmental change) and complex cases (multiple segmental changes); 3) misperception of syllable shapes when syllable insertions/deletions occurred; 4) lexicon, such as the role of frequency, nonwords and morphology; 5) syntax, such as grammaticality, constituent boundaries. The percentages and discussions of certain instances of the errors were provided per category but they lacked statistical analyses.

The descriptive analyses (as mentioned above) were conducted separately using the adult misperception data and children misperception data. The confusion patterns with the children's data were on the whole similar to those found in the adult data (Bond, 1999, pp. 83–98). However, children produced errors less constrained by the language. For instance, they were more likely to perceive non-words, which suggests that they are less constrained by the lexicon, e.g. *to missions* was perceived as *temissions*; and they were more likely to perceive ungrammatical phrases/sentences, suggesting that they were less constraint by syntax, e.g. *cuff him* was perceived as *cough him*, and *Mayor McCheese* was perceived as *Mayor get cheese*. Together this suggests that children use the knowledge of their language less often than the adults in perception, presumably because they are still acquiring the language.

In summary, the author extensively summarised the patterns of misperceptions at different linguistic levels and demonstrated how naturalistic misperception data can be used to generate hypotheses of speech misperception that can be subsequently tested in laboratory conditions. This was indeed the case with Cutler and Butterfield (1992) and Vitevitch (2002). Both studies relied on Bond's (1999) data to generate hypotheses which were tested experimentally.

1.2.4 Labov (1994b) and Labov (2010b)

Labov and his colleagues collected 872 misperception instances over the course of fourteen years. Unlike Browman (1980), Bird (1998) and Bond (1999), Labov (1994b) and Labov (2010b) focused on how misperceptions motivate sound change rather than linguistic structures.

Labov (2010b) classified the instances of misperception by assigning a tertiary scoring scheme relating to how each of the five linguistic factors (lexicon, dialect, phonology, pragmatics and syntax) was inhibiting, promoting or neutral to the misperceptions. This classification reflected the relative effect of these factors to the

misperception instances. It was found that phonology was the main promoter, while pragmatics was the main inhibitor. Dialect came second with 27% as the promoter of misperceptions.

When studying sound change, one crucial factor is to examine the frequency of confusions, such that more frequently confused pairs of segments are more likely to merge. In relation to sound change, Labov (1994b) examined the relative rates of vowel misperceptions to support the concept of subsystem. Subsystems are defined by Labov (1994b) as a set of vowels that maximally subjects to pairwise confusion. Labov (1994b) predicted that confusions across subsystems should be less frequent than those within subsystems.

Labov (1994b) classified the misperception instances into two types, "global" and "local" in order to test this prediction, and focused only on the *local* type. Global misperceptions (or misunderstandings as he called them) are defined as cases in which the phonetic conditions are degraded and the perceived utterance has very little in common with the intended utterance, e.g. ***The mayor*** *found an answer for the Eagles.* $\rightarrow$ ***Ralph Nader*** *found an answer for the needles.*. While local misperceptions are defined to be dependent on the misperception of a particular segment, which he suggested was due to the phonetic realisation of the segment, e.g. *I'd go to the Acme and b**a**g* $\rightarrow$ *I'd go to the Acme and b**e**g*. Focusing on vowel confusions, the author examined four segmental environments – post-vocalic nasals (e.g. *thi/n/k*), /l/ (e.g. *goa/l/*), /r/ (e.g. *fea/r/*) and pre-vocalic obstruent-liquid clusters (e.g. */bl/ack*) – that tended to reduce the phonetic distance between vowels.

154 local misperceptions of these types were examined. Globally, it is not true that there are more misperceptions within subsystems (78 out of 154) than across subsystems (74 out of 154). However, a closer examination of their distribution across the four segmental environments showed radical differences between the misperceptions within and across subsystems.

It was found that 40% of the misperceptions within subsystems occurred in one of these four segmental environments, while 80% of the misperceptions across subsystems occurred in these same segmental environments. This distribution suggests that after excluding misperceptions triggered by phonetically reduced environments, misperceptions are more frequent within subsystems than across subsystems as predicted. More generally, the author used the frequency of local misperceptions to strengthen the conception of phonological subsystems and phonological hierarchies.

1.2.5 Mondegreens

Mondegreens are the naturalistic misperception of sung speech. The term Mondegreen is, in fact, a Mondegreen itself. The writer Sylvia Wright published her own misperception of a Scottish ballad in her magazine article (Wright, 1954), titled "The Death of Lady Mondegreen". The ballad is called "The Bonny Earl of Murray":

> Ye highlands, and ye lawlands, Oh! whair hae ye been? They hae slaine the Earl of Murray, and layd him on the green.

This was misheard as:

> Ye highlands, and ye lawlands, Oh! whair hae ye been? They hae slaine the Earl of Murray, and **Lady Mondegreen**.

Two main differences between Mondegreens and misperceptions of conversational speech are that Mondegreens involve sung speech and are masked with music. Sung speech is different from spoken speech in a number of ways, such as duration, vowel formant (Sundberg, 1970), intonation (Zatorre and Baum, 2012), pronunciation (using full vowels instead of schwas, e.g. *evil* could be pronounced as [i**ː**v**iː**l]) etc. While sung speech is almost always masked with music, spoken speech may be masked with a wider range of noise types.

Furthermore, words and melodies are processed interactively (Gordon et al., 2010; Asaridou and McQueen, 2013). Listening to sung speech is a very different activity compared to listening to a conversation. When listening to conversational speech, the listener is often involved in the conversation, which provides a great deal of contextual information. When listening to sung speech, the nature of the context is more complex; it involves not only the listener's immediate surrounding and the activity that the listener is doing while listening to sung speech, but also the general expectation of the artist and the genre of the song containing the sung speech.

In summary, it is clear that Mondegreens and misperception of conversational speech differ in the production of language, the listening environment, the perception mechanism, and the available context. Given these differences, analyses of naturalistic misperception should consider Mondegreens separately from those of conversational speech, and Mondegreens are therefore not analysed in this thesis.

1.3 Complementarity of laboratory and naturalistic studies

1.3.1 Arguments against naturalistic data

The naturalistic corpora have largely been observational data from uncontrolled sampling, leading to many researchers questioning their value.

Cutler (1982) provided a detailed discussion on the reliability of this kind of data with a focus on misproduction. From a theoretical point of view, the discussion concentrated on the issue of detectability of different types of speech misproduction. Using research on speech perception, she suggested that detectability is dependent on factors including hearing errors, and perceptual confusions. The paper concluded that since the level of detectability varies with the kinds of errors, the misproduction

data from everyday life are confounded with the problem of detectability. Although these confounds mostly apply to misproduction, a potential confound for speech misperception is that they could in fact be misproduction. Another relevant confound the paper pointed out is the issue of reconstruction (see also Cutler et al., 2000). Consider that when the hearer perceives an implausible word, the hearer could reconstruct a plausible word (possibly even the same word as the intended word) as a repair strategy. These misperception instances would not be recorded by the reporters, and there is no way of knowing how often this occurs, thus biasing the data.

Ferber (1991) further argued against naturalistic speech misproduction data particularly highlighting the reliability of the collection process itself. The study gave a list of possible factors that could affect the quality of the data, which is reproduced below:

1. Failure to record
    (a) Because slips were not recognized as such
    (b) Because they were not remembered
    (c) Because the decision - slip of the tongue or other speech error - was not taken quickly enough or was wrong
    (d) Because slips occur so frequently that there is not time to record them all
2. Erroneous recording
    (a) Because slips were misheard
    (b) Because of faulty recall on the part of the recorder
    (c) Because slips were transcribed incorrectly
3. Erroneous classification

(a) Because slips were transcribed incorrectly

(b) Because the context was transcribed incorrectly

(c) Because the context was insufficient

The study examined the consistency of "on-line" collection of speech misproduction. Three people were asked to collect instances of speech misproduction from a recording of people making speech production errors. The results suggested that the consistency between collectors was poor, there were not particular kinds of speech misproduction that were more detectable than others as suggested by Cutler (1982), and that production and perception cannot be separated when collecting instances of speech misproduction. Again, the same arguments can be made for the consistency of collecting instances of speech misperception.

1.3.2 Arguments for naturalistic data

Having cast doubts on naturalistic corpora, Cutler (1982) concluded that experimentally induced speech misperception data could complement naturalistic data. Many researchers did in fact use this approach when utilising naturalistic data, for example in Cutler and Butterfield (1992), Bond (1999), Vitevitch (2002) and Labov (2010a) and Labov (2010c). These studies all showed that their experimental findings agreed with the naturalistic findings, and some of these studies are discussed below.

Cutler (1982) set out to test the rhythmic segmentation hypothesis, which predicts that English listeners operate on the assumption that strong syllables are likely to be the initial syllable, while weak syllables are either not word-initial and if they are, they are more likely to be grammatical words. The study started with an analysis of 246 juncture misperceptions from Bond's (1999) corpus, and found that there is indeed more juncture insertions before strong syllables than weak syllables, and more juncture deletions before weak syllables than strong syllables. In an experiment

where low amplitude speech was used, the same error pattern was found in listeners' misperception.

Vitevitch (2002) examined the role of lexical frequency and neighbour frequency in speech misperception. First, using a subset of Bond's (1999) corpus, it was found the words involved in naturalistic misperception tended to have a higher frequency than words randomly sampled from the lexicon which contradicted previous research such as Brown and Rubenstein (1961). To elucidate this contradiction, two experiments were conducted by manipulating the duration of words in auditory stimuli. In one case, the duration was reduced by 25% for all words, and in the other case, all words were made to be the same duration. It was demonstrated that the perceptual consequences would be different for words varying in word frequency, depending on their duration.

Overall, the author concluded that the advantage of high frequency words in perceptual processing could be attenuated when there are other variations present (as in naturalistic settings) such as phonological similarity (Luce and Pisoni, 1998) and word duration (Wright, 1979), thus resulting in the contradictory findings in the corpus analyses.

Finally, in Music Information Research, the misheard lyric matching problem has been known to be a challenge for internet search engines. This is when users enter misheard lyrics in a search engine hoping to find the intended lyrics. One model using naturalistic corpora was introduced by Hirjee and Brown (2010) who utilised naturalistic corpora of speech misperception but of lyrics rather than of conversational speech. The data were 20,788 instances from misheard lyrics websites that collect instances from the public. Hirjee and Brown (2010) introduced a probabilistic model of mishearing by calculating phoneme confusion frequencies of the misheard lyrics. The model was able to find up to 8% (of 146 misheard lyrics queries) more correct lyrics than other methods that did not use these naturalistic data, such as phoneme edit

distance (Levenshtein, 1966; Luce, 1986b). This study suggests that the use of naturalistic corpora of misperception provides better predictions of people's perceptual errors of lyrics, thus strengthening the validity of the data.

1.4 Conclusion

Earlier work on large scale confusion experiments by Miller and Nicely (1955) and Wang and Bilger (1973) provided insights into the overall pattern of perceptual confusions in the lowest domain, auditory/phonetics, i.e. without any interactions with any of the linguistic levels higher up. Despite the domain restriction, their confusion matrices have been used widely to predict perceptual processes in many fields of research (e.g. speech sciences, audiology, linguistics and computer sciences). Their novel methodologies with information transfer shed light into the nature of distinctive features in perception, and have remained a prominent method for analysing confusion matrices. Later perception research with speech in noise investigated both external factors such as environmental noise and speech transmission systems, as well as internal factors, including the perception of unusual accents, and perceptual difference between typical and atypical populations. However, much of these studies only considered the global performances of the listeners in adverse conditions rather than the nature of their errors. Furthermore, the existing macroscopic models (e.g. Articulation Index (AI)) that predict how distorted speech will be perceived have only focused on predicting the overall speech intelligibility as a numerical value. The few microscopic models, (Holube and Kollmeier, 1996; Cooke, 2006), which aim to predict listener's responses to individual degraded speech tokens, are still in their infancy and can only operate in very restricted domains, such as nonsense VCV syllables.

Many researchers have seen the potential of using naturalistic data in error

research as a testbed for theories and findings from laboratory studies, and thus spawned a trend of naturalistic corpora collected by error researchers. Unlike naturalistic speech misproduction, existing data and analyses for misperception were scarce. The ones that did exist either have limited data (Browman, 1980), or the analyses were not quantitative enough (Bond, 1999), or are restricted in their dialectal interactions (Labov, 2010b). Although the data of the naturalistic corpora have potential reliability issues, as argued extensively by Cutler (1982) and Ferber (1991), the counter arguments and successes in using the naturalistic corpora to support experimental data have been overwhelming (Cutler and Butterfield, 1992; Bond, 1999; Vitevitch, 2002; Labov, 2010a; Labov, 2010c; Hirjee and Brown, 2010).

1.5 Research aim and organization of the thesis

Having reviewed the research on speech misperception in the laboratory and beyond the laboratory, it is clear that there is a gap between experimental work which has controlled for and explored a limited number of factors at one time, and the research that uses naturalistic corpora involving multiple levels of linguistic interaction. This thesis aims to 1) bridge this gap by comparing naturalistic and experimental data and 2) establish the naturalistic corpus as an ecologically valid resource and a benchmark of misperception by identifying the effect of both top-down and bottom-up factors on misperception. The thesis is organised as follows.

First, both the quality and quantity of naturalistic misperception data need to be improved. This thesis aims to achieve this by further collecting naturalistic data with meta-data (e.g. the demographics of the interlocutors) and recompiling all of the existing corpora into a standardized format for both orthographic and phonetic transcriptions. Such a mega corpus will allow more comprehensive analyses to be done, especially in cases where it had not previously been possible due to data

sparsity and because of the possibility of cross-validation of a phenomenon using subsets of the corpus. Chapter 2 presents a detailed documentation of the compilation process. The result of this approach is a mega corpus of naturalistic English misperception with complete orthographic and phonetic transcriptions, alongside the available meta-data. In addition, the corpus is phonetically aligned and syllabified using data-driven methods to avoid issues of circularity.

Second, naturalistic misperceptions are affected by both low level factors (such as acoustic/featural similarity) and high level factors (such as stress, syllable positions and segmental/lexical frequencies). A number of questions arise as a result. How much of an impact do these low level factors and high level factors make? How similar are naturalistic misperceptions and experimental misperceptions? Are they more similar in specific experimental conditions, such as signal to noise ratio (SNR)? In other words, are there specific experimental conditions that are more ecologically valid than others? In order to answer these questions with a focus on bottom-up low level factors, Chapter 3 presents analyses of naturalistic misperceptions on a featural level (place confusion, manner confusion and voicing confusion), as well as on a segmental level. Low level phonetic factors were found to play a surprisingly important role in consonant and vowel confusions, despite the potential effects of high level factors. Extreme experimental conditions such as too high or too low SNRs, or narrow bandwidths tend to be least ecologically valid. Asymmetrical patterns of confusions were consistently found in both naturalistic and experimental data. Together the results highlight the fact that naturalistic data are affected by low level phonetic factors and are complementary to the experimental data.

Third, Chapter 4 examines whether naturalistic misperceptions are affected by top-down high level factors from the lexicon. Top-down effects were found to play a role across multiple linguistic levels of an utterance – segments, syllables, and words. These factors include the segmental frequency, syllable constituencies and syllable

positions, stress, word frequency and the conditional probabilities of words given other words in utterances. The findings suggest that naturalistic misperceptions are affected by high level factors. Many of these factors were not possible to identify in experimental misperception studies that focus on nonsense CV and VC syllables, thus highlighting the need for examining larger units than nonsense syllables.

Finally, Chapter 5 discusses the limitations of the thesis across the levels of segments, syllables and words and also perspectives for future work with a focus on the individual analyses in Chapter 3 and 4. Potential solutions are suggested, such as by doing context-sensitive analyses using the naturalistic corpus, conducting laboratory experiments and creating a computational model that incorporates the bottom-up and top-down effects found in misperception.

Chapter 2

Corpus compilation

This chapter documents the core data that are used in this thesis. The subsequent chapters in the thesis contain analyses that are based on these core data; therefore, close attention was paid to how the data was processed as well as the reasoning and assumptions that were made in the process. Given that arbitrary decisions have to be made when processing the data, the extensive documentation that is provided in this chapter will allow future researchers to more accurately replicate the findings that were made using this data, and to question the assumptions involved. Researchers can then make precise modifications and the resultant findings can be compared with each other. Crucially, such comparisons could inform us of the validity/appropriateness of specific modifications and the robustness of the findings.

This chapter is broken down into five sections and structured as follows: Section 2.1 will document the naturalistic English misperception corpora, detailing the backgrounds and formats of the existing corpora, as well as the steps that I took to compile them into a single mega-corpus. Section 2.2 will document the phonetic transcription that was performed on the mega-corpus, detailing the transcription heuristics that I adhered to, my source of pronunciation (the choice of pronunciation dictionaries), the broadness of transcriptions (e.g. stress and allophones), the

syllabification method, the dialect classification system based on geographical information, and finally the vowel sets for 14 dialects of English. Section 2.3 will document the compilation and the processing of the mega written text corpus that was used as a control corpus from which psycholinguistic variables (such as token frequencies, neighbourhood density and others) were derived. Having established the compilation and processing of the misperception corpora and the control corpus, I will switch the focus to how the phonetic transcriptions from the misperception corpora can be used to detect misperception. Section 2.4 will document a standard approach for detecting differences between transcriptions, which is commonly called *pairwise alignment* (Durbin et al., 1998), starting with a review of some of the existing alignment methods and finishing with an adapted version of an existing method that I will use for the alignments in this thesis. Finally, Section 2.5 will summarise the contribution that this chapter has made to the field of linguistics.

2.1 English naturalistic corpora

Only five naturalistic speech misperception corpora of English exist. There are many advantages of compiling all five corpora into one mega corpus. Different corpora have been collected by different individuals in different parts of the English-speaking world (predominantly the US and the UK). Any agreement between the analyses using different corpora would provide support for the results not simply being a product of the following factors.

a) Geographical differences: the collectors involved in each corpus are located in different places.

b) Collectors' bias: since the researchers themselves are often also the collectors, they might be inadvertently biased to collect/detect particularly kinds of misperception that are related to their research question.

c) Small sample size: small samples are less likely to be representative of the population. They are also limited to low-power statistical techniques, such as the chi-squared test.

Similarly any disagreement between findings in different corpora would open up areas of further research thus allowing us to triangulate our findings derived from each of the corpora (Browman, 1978, Ch. 2). In this section, these six corpora will each be documented in turn below, in terms of the means of collection, the inclusion of any meta-data (such as demographics of the interlocutors), any categorisation of the corpus, and other corpus-specific details.

2.1.1 Background

2.1.1.1 Browman

Browman (1978) collected 222 naturally occurring hearing errors from spontaneous casual English conversation. The details of the collection process, the data structure of the published corpus and the meta-data are described below.

2.1.1.1.1 Collection process The collectors were the author and her friends in Los Angeles, as well as three other linguists, Zinny Bond and Sara Garnes in Columbus, Ohio, and Stefanie Shattuck-Hufnagel in Boston, Massachusetts. The interlocutors (utterers and perceivers) were academics.

2.1.1.1.2 Data structure The published corpus was provided in two tables in the appendix of the paper; see Figure 2.1 for a snapshot of the corpus.

2.1.1.1.2.1 Orthographic transcriptions In the first published table (see Figure 2.1a), the full utterance was reported in orthographic form, including portions that were not correctly perceived, with the erroneous portion of the utterance isolated

between slashes. If there were no slashes in the intended utterance, this meant that there was no correctly perceived portion in that utterance. The perceived utterance includes only the erroneous portion, since the full perceived utterance can be reconstructed from the intended utterance. If the correctly perceived portion is identical for two consecutive data points, the correctly perceived portion of the second data point is indicated with "[ditto]".

1	LOOKS VERY /WIDE/	1	BLIND
2	YOU WONT HAVE TO PUT /APPLES/ DOWN.	2	UP WITH
3	FEATURE DETECTORS	3	FRIGIDITY
4	THATS /BROAD/ ON MY NOSE	4	BOARD
5	[DITTO] BROAD	5	POOR
6	[DITTO] BROAD	6	RAW
7	YOU MAY HAVE THE /REST/ OF IT	7	BEST
8	REAL HYPOCRITE	8	RAH@@PRACHAT
9	/DOLLIE/ TOMORROW	9	DAWN OF
10	MAJORCA	10	MALAGA

(a) Orthographic transcriptions

1	'WAYD	1	'BLAYND
2	'AEP$AXL+Z	2	'AHP#WIHDH
3	'FIYCHER#DIH'TEHK$TER+Z	3	FRIH'JHIHDIHTIY
4	'BRAOD	4	'BAORD
5	'BRAOD	5	'PUWR
6	'BRAOD	6	'RAA
7	'REHST	7	'BEHST
8	RIHL#'HHIHPAXKRIHT	8	RAX'HHUWPRAXCHAXT
9	'DAALIH	9	'DAAN#AHV
10	MAY'YOWR$KAX	10	MAX'LAAGAX

(b) Phonetic transcriptions

Figure 2.1: Browman's (1978) corpus: intended utterance (left) and perceived utterance (right) (please note that these are the original transcriptions in Browman's (1978) corpus and not the transcriptions in the combined corpus)

2.1.1.1.2.2 Phonetic transcriptions In the second published table (see Figure 2.1b), the phonetic transcriptions of the erroneous portion were listed in APRAbet, a transcription coding scheme using ASCII characters, developed by the Advanced Research Projects Agency (ARPA). This means the correctly perceived portions of the utterances were not phonetically transcribed. Finally, a table of the IPA–ARPAbet mapping was also published; see Figure 2.2.

There was no mention of when the phonetic transcriptions were made. It is not

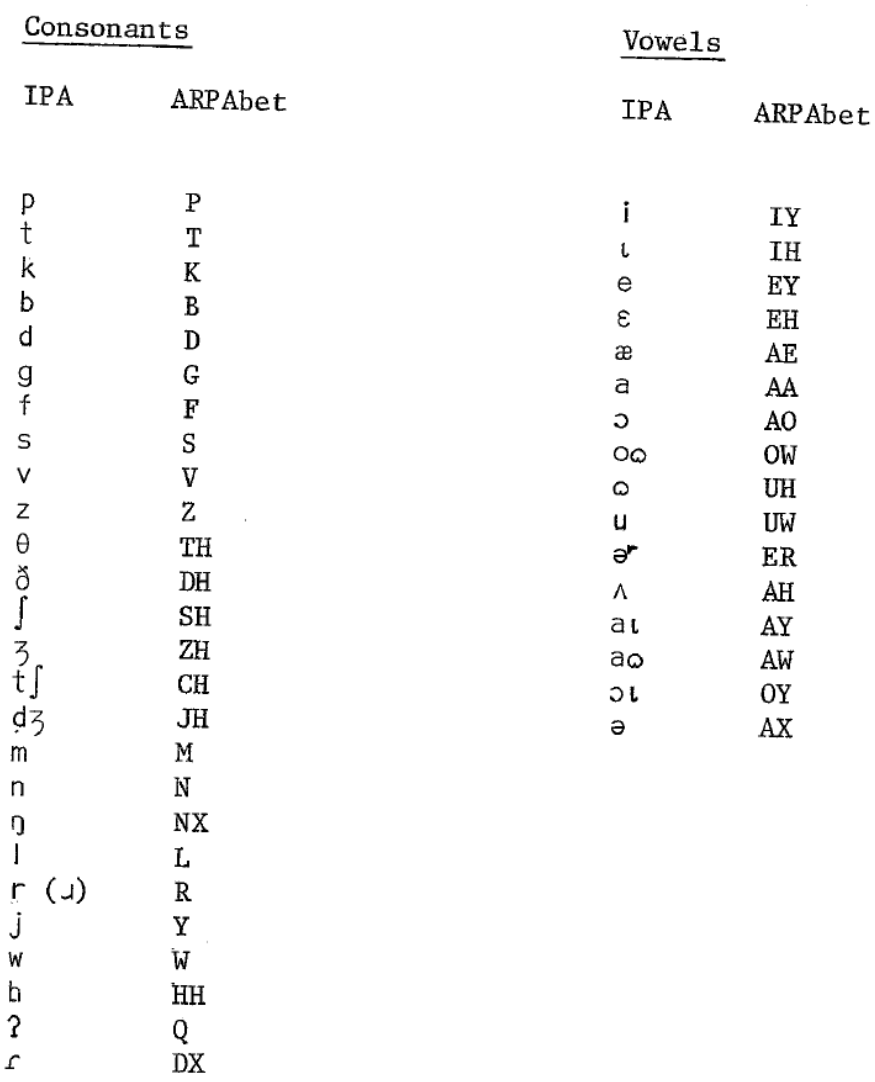
possible to know if the transcriptions were done contemporaneously by the collectors – immediately after the misperception was realised – or if they were done much later. The author did not report if the transcriptions were transcribed to a particular accent or not. However, judging from the IPA–APRAbet mapping, it is reasonably clear that the transcriptions were done on a broad level, but since the mapping included cases of taps and glottal stops, the transcriptions were not so broad that they were phonemic. Since only one set of vowels were listed, it is reasonably clear that the transcribers did not take into account accent variations, because otherwise there would have been one set of vowels per accent group.

2.1.1.1.2.3 Meta-data No mentioned of meta-data, such as the demographics of the utterers and perceivers, was included. The author had already passed away at the at the time of writing this thesis and I was therefore unable to follow up on any meta-data for each reported hearing error.

2.1.1.2 Bird

Bird (1998) collected 300 naturally occurring hearing errors from natural English conversation. The details of the collection process, the data structure of the published corpus and the meta-data are described below.

2.1.1.2.1 Collection process No information about the collectors was mentioned; therefore, information about where the collectors were from, and how many collectors were involved is unknown. The author reported that for each instance of misperception, the intended and perceived utterances were first transcribed orthographically, and then other factors were noted such as noise, context, regional accent and world knowledge of the listener etc.



Consonants

IPA	ARPAbet
p	P
t	T
k	K
b	B
d	D
g	G
f	F
s	S
v	V
z	Z
θ	TH
ð	DH
ʃ	SH
ʒ	ZH
tʃ	CH
dʒ	JH
m	M
n	N
ŋ	NX
l	L
r (ɹ)	R
j	Y
w	W
h	HH
ʔ	Q
ɾ	DX

Vowels

IPA	ARPAbet
i	IY
ɪ	IH
e	EY
ɛ	EH
æ	AE
a	AA
ɔ	AO
oʊ	OW
ʊ	UH
u	UW
ɚ	ER
ʌ	AH
aɪ	AY
aʊ	AW
ɔɪ	OY
ə	AX

Figure 2.2: IPA–ARPAbet mapping for Browman's (1978) corpus

2.1.1.2.2 Data structure The published corpus was in the format of a list in the appendix of the paper; see Figure 2.3 for a snapshot of the corpus.

2.1.1.2.2.1 Orthographic transcriptions The author published the intended and perceived utterances in orthographic form. In the list, for each data point the intended utterance comes first, and the perceived utterance comes second. The full utterances were listed, including both correctly and incorrectly perceived portions.

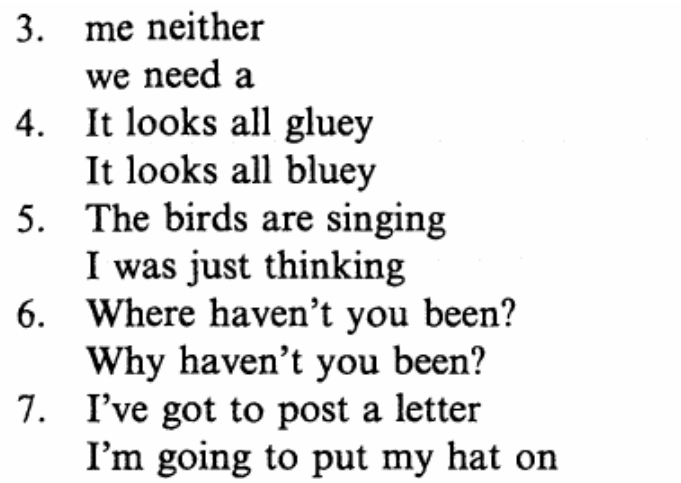


Figure 2.3: Bird's (1998) corpus: orthographic transcriptions (please note that these are the original transcriptions in Bird's (1998) corpus and not the transcriptions in the combined corpus)

2.1.1.2.2.2 Phonetic transcriptions The author did not publish any phonetic transcriptions in the appendix. No details were provided for how the transcriptions were done. The author referred to the segments of analyses as phonemes, and we can therefore could assume that the transcriptions were phonemic.

2.1.1.2.2.3 Meta-data The author did not publish any meta-data, such as noise, context, regional accent and world knowledge of the listener.

2.1.1.3 Labov

Labov (2010b) collected ≈ 870 naturally occurring misperceptions in English. The details of the collection process, the data structure of the published corpus and the meta-data are described below. The author used the term *misunderstandings*, instead of *misperception*. For consistency purposes, I will refer to them as misperceptions throughout.

2.1.1.3.1 Collection process The collection was part of a project on Cross Dialectal Comprehension. The collectors were linguists and linguistic students, and

they were asked to note down any instances of misperception on a pad of printed collection cards. Figure 2.4 shows a collection card that was used.

The collection process covered 16 years between 1984 to 2000. The majority of the misperceptions ($\approx$ 60%) were collected between 1986 to 1988; fewer than 60 misperceptions were collected in each of the remaining twelve years. During 1986–1988, frequent reminders were sent to the collectors to encourage them to collect, and this yielded on average two to four instances per week per collector.

There were nine collectors who were linguists with phonetic training, and were belonged to the following places: Long Island City, Montreal, Northern New Jersey, Connecticut, Chicago, California, Edmonton, and New York City. These linguists provided the majority of the corpus (76%). The author mentioned that they made every effort to avoid collection bias that would favour the collection/detection of dialectally-motivated instances. The author argued that the period during which a majority of the instances was collected would be less biased than the years when fewer misperceptions were collected.

MISUNDERSTANDINGS Date _____
Speaker _____ Hearer _____
Dialect area _____ _____
Speaker said [continue on back for full setting]:

Hearer heard:

Hearer corrected mishearing after ___sec ____min
____ before utterance was over
____ by speaker's response to look or query
____ by inference from further utterances
____ by accidental events that followed

Figure 2.4: Labov's (2010) corpus: collection card

2.1.1.3.2 Data structure The collection cards (Figure 2.4) created a basis for the data structure of corpus. The cards encouraged the collectors to complete the information asked for, as well as making the collection process more amiable. Although the corpus was not published, Professor William Labov has kindly provided

Professor Andrew Nevins and me with the corpus for research purposes.

The corpus was encoded in FileMaker format. It was converted into an excel spreadsheet for simpler data extraction and viewing. The content of the corpus is discussed below.

2.1.1.3.2.1 Orthographic transcriptions Figure 2.5 shows an example of the spreadsheet. The orthographic transcriptions shown in the first column were tabulated in a specific format. The identity of the speaker is indicated before the intended utterance and separated by a colon. The symbols "=>" appearing before the perceived utterance are used to indicate "heard as". The perceived utterance follows a similar format as the intended utterance with the identity of the perceiver listed before the utterance. Notes about the instance itself are enclosed by square brackets "[]".

Text	Phonemes
Text as heard with notes => 'heard as'	
Gillian: Appendixes are welcome => Ruth Herold: Tendencies are welcome.	*/0 p/t n/0 ks/s
Eduardo Llach: Do you know any place where I can get some coffee? => RS: . . get some copies.	f/p 0/s
Robin: Some days I feel like dung. Sherry: You're not dumb.	ng/m

Figure 2.5: Labov's (2010) corpus: orthographic and phonetic transcriptions (please note that these are the original transcriptions in Labov's (2010) corpus and not the transcriptions in the combined corpus)

2.1.1.3.2.2 Phonetic transcriptions Figure 2.5 shows an example of the spreadsheet. The broad phonemic transcriptions are shown in the second column. The transcription notation followed Labov's own system and was written in ASCII rather than IPA or SAMPA. Only the phonemes that were different between the

intended utterance and the perceived utterance were transcribed. The intended and perceived phonemes were separated by a slash. It was not documented whether the transcriptions were done contemporaneously or later prior to submission. While later transcriptions could lower the level of accuracy, any possible reduction in accuracy may have been lessened by the broader, phonemic nature of the transcriptions.

2.1.1.3.2.3 Meta-data The meta-data in the corpus are richer than those in Browman's (1978), Bird's (1998), and Bond's (1999) corpora. Figure 2.6 shows an example of the meta-data section of the spreadsheet.

The only information provided about the utterers and perceivers was their place of origin. For US places, abbreviations were used to indicate the names of the states. Other demographics, such as age and gender, were not explicitly specified, but the names of the collectors were reported and the exact date, month and year when each instance occurred was available. The corpus contains temporal information about how many seconds after the spoken utterance was perceived before the perceiver noticed that he/she misperceived the utterance.

Date	Speaker	Hearer	Observer	NOTICE TIME
	Geographic background	Geographic background		
12/17/1986	Phila	NNJ	WL	3
11/26/1986	Cleveland	NY	RSabino	1
11/26/1986	NYC..	Chicago	SAsh	5
11/26/1986	NYC	Chicago	SAsh	1
12/02/1986	Phila	NY	R Sabino	10

Figure 2.6: Labov's (2010) corpus: meta-data

2.1.1.4 Bond

Bond (1999) collected ≈ 900 hearing errors from ordinary conversational speech. The details of the collection process, the data structure of the published corpus and the meta-data are described below.

2.1.1.4.1 Collection process The author herself was either the utterer, the perceiver, or the observer for some of the errors. A sizable portion of the corpus consisted of errors reported by students, friends and colleagues of the author. Unlike other naturalistic corpora, this corpus has a subcorpus of 105 misperception instances by children. However, the majority of the corpus, 783 instances, was provided by adults. The authors mentioned that the instances were collected over the course of many years. The first mention of the corpus was in an earlier paper, Garnes and Bond (1980). This paper stated that the corpus contained around 900 instances which is consistent with the figure reported by Bond (1999); therefore, we can infer that the collection process stopped before 1980, and the collection is likely to have started several years before then.

The vast majority of the collected instances occurred during face-to-face conversations. The exact background details of each instance varied considerably, and included anything from being in a car to ordering food in a restaurant. 2% of the corpus consisted of telephone conversations and 5% consisted of cases where the speakers were not addressing the listeners directly, such as via television and radio. Demographics such as the accent of the utterers and perceivers were rarely available, especially when the reporter did not know the interlocutors that were involved in misperceptions.

2.1.1.4.2 Data structure The published corpus was in the format of a list in the appendix of Bond (1999); see Figure 2.7 for an example of the corpus.

2.1.1.4.2.1 Orthographic transcriptions The author published the intended and perceived utterances in orthographic form. For each data point in the list, the intended utterance was listed first, followed by an arrow, and then the perceived utterance was listed. The intended utterance was listed in its full form (including both correctly and incorrectly perceived portions), while only the incorrectly misper-

ceived portion was listed for the perceived utterance. However, it is obvious in almost all of the cases that the full form of the perceived utterance can be reconstructed from the full form of the intended utterance.

1. What's wrong with her bike? → her back
2. Wattsville → Whitesville (North Carolina to Ohio)
3. It really turned wet out → white out
4. It's like a math problem → mouth problem
5. Did I ever tell you about this usher? → this esher
6. where we went to the horse show → horse shoe
7. a lot of nude beaches → nude bitches
8. I don't know if we have any more "trecks" left → tracks left
9. You know that soil can be → swail can be
10. Fleischmann's → Flashman's

Figure 2.7: Bond's (1999) corpus: orthographic transcriptions (please note that these are the original transcriptions in Bond's (1999) corpus and not the transcriptions in the combined corpus)

2.1.1.4.2.2 Phonetic transcriptions The author did not publish any phonetic transcriptions in the appendix. The author stated that accurate phonetic representation of the speakers' utterances was not available, which implies that the transcription was not done contemporaneously. The transcriptions used by the author were essentially phonemic, or in the author's description, "a distinct pronunciation". The author argued that this best captures the utterer's intent, even with words that are often reduced such as *and*, and the full form /ænd/ would still be used in the transcription. The drawback is that the transcription might therefore be less realistic and more arbitrary. Furthermore, by assuming that the phonemic forms were used, the perceivers would have to bear more "responsibility" for the misperception, which might be overestimated, because it is possible that some of misperceptions were due to pronunciations that were reduced.

2.1.1.4.2.3 Meta-data The author did not publish any meta-data. As mentioned previously, the demographics of the utterers and perceivers were rarely available, so little was available to publish.

2.1.1.5 Nevins

Prof. Andrew Nevins collected ≈ 2,900 instances of misperceptions from conversational English speech. The details of the collection process, the data structure of the published corpus and the meta-data are described below.

2.1.1.5.1 Collection process Prof. Nevins recruited 24 linguistics students who were attending a course on speech misperception at Harvard University for one semester (14 weeks) per year in 2009 and 2010.

Over the 14 weeks, they were made aware of various kinds of misperception errors using Bond (1999) and other papers as course materials, starting from phonetic factors through to pragmatic factors. The students were instructed to report 5-10 misperception instances in their ordinary daily life per week. Some of the errors contributed by the students were analysed in each class. Specifically, they were instructed to record the intended and the perceived utterances in orthographic form, and if possible, provide phonetic transcriptions, the demographics of the utterers and perceivers (such as the age, gender, accent, native and non-native language(s) and hometowns), the context of each misperception, and any comments or corrections by interlocutors. At the end of the two years this collection yielded 2,857 instances of misperceptions of mostly American English speech, of which 1,523 instances were in 2009, and 1,334 instances were in 2010.

2.1.1.5.2 Data structure The corpus was made available to myself in an excel spreadsheet format by Professor Nevins. The content of the corpus is discussed below.

2.1.1.5.2.1 Orthographic transcriptions Figure 2.8 shows an example of the spreadsheet. The orthographic transcriptions are shown in the first two columns. The intended and perceived utterances were listed in two separate columns in their full form.

Intended	Perceived	IPA Intended	IPA Perceived
Shorty want a thug.	Shorty want a hug.	θʌg	hʌg
Like a G6	Lick a Cheesesteak	laɪk ʌ dʒi sɪks	lɪk ʌ tʃizsteɪk
I love seedless grapes.	I love cinnamon sticks.	sidlɛss greɪps	sInʌmʌn stIks
Shoot me now!	Shoot me cow!	næʊw	kæʊw

Figure 2.8: Nevins's corpus: orthographic and phonetic transcriptions (please note that these are the original transcriptions in Nevins's corpus and not the transcriptions in the combined corpus)

2.1.1.5.2.2 Phonetic transcriptions Figure 2.8 shows an example of the spreadsheet. The phonetic transcriptions are shown in the third and fourth column but the transcriptions were not always available and only the misperceived portions of the utterances are transcribed in most cases. Collectors transcribed in IPA, but there was a considerable amount of notational variations; for instance, [ə]-[ʌ], [ɛ]-[e], [eɪ]-[ej] and stress marks were not always provided. There are no indications of whether the perceived utterance was transcribed with the perceiver's accent or the utterer's accent (see Section 2.2.5 for why this is relevant); the students were simply told to transcribe the utterances in IPA, without specific instructions about whose accent to transcribe. Finally, the instructions to the students did not indicate whether the transcription should be done contemporaneously or immediately after

the misperception instances, so it is possible that the transcriptions were done much later, which would have lowered the accuracy of the transcriptions.

2.1.1.5.2.3 Meta-data The meta-data in the corpus are detailed compared to the other corpora. Figure 2.9 shows an example of the meta-data section of the spreadsheet.

The meta-data includes the utterers' and perceivers' locations, genders and ages. The level of detail varies, e.g. with the details ranging for location from giving the exact states from which the interlocutors came, to just simply stating the country.

The location in which each instance took place along with some descriptions of the locations were also documented under the column "Where", e.g. *Room, a few girls talking*, *Weld Laundry Room. Multiple washers and dryers running*. The names of the collectors were also noted, just the first name in most cases.

If a particular instance was from misheard lyrics (also called mondegreens), rather than conversational speech, this would be indicated under the column "Notes". For mondegreens, the name of the utterer and the song would both be specified in the demographics of the utterers, e.g. *Lil Wayne, "Lollipop"*.

The topic of conversation was also reported, e.g. *Meeting each other for the first time. Getting to know each other*. Lastly, any requests for clarifications by the perceivers were also reported, e.g. *Wait! What did you say?*.

2.1.2 Compilation

The existing corpora were summarised above. The following sections will describe how all five corpora were compiled into one mega corpus to cover the orthographic transcriptions, the phonetic transcriptions and the meta-data. Due to the high level of variability between corpora, one of the most difficult tasks was to normalise their format, and to extract as much as useful information as possible while making

Utterer	Perceiver	Where	Collected By	Notes	Topic of Conversation	Request for clarification
Lil Wayne "Lollipop"	NY, 18	Room, a few girls talking	Dan	Mondegreen		
Movement "Like a G6"	PA, 19	heavy music and commotion	Stephen	Mondegreen	n/a	n/a
Female, Singapore, 18	PA, 18	Study Break. Crowded. Lots of talking.	Lauren	n/a	n/a	"Wait! What did you say?"
Female, Florida, 18	PA, 18	Speaker mildly exasperated.	Lauren	n/a	n/a	no
Female	PA, 18	Weld Laundry Room. Multiple washers and dryers running.	Lauren	n/a	Meeting each other for the first time. Getting to know each other.	"How many sisters?"
Male	PA, 18	Class	Lauren	n/a	How to use slips chart.	psychological processes?"
Cherish "Killa"	PA, 18	Radio	Lauren	mondegreen	n/a	no
Romanian female, 23	Israeli male, 37	my office	Michael	n/a	student papers	no

Figure 2.9: Nevins's corpus: meta-data

reasonably justifiable assumptions.

2.1.2.1 Orthographic transcriptions

Any normalisation of the orthographic transcriptions in the misperception corpus needs to be consistently applied to a written English corpus which is used as a representative corpus of English. This written English corpus acts as a "control", and it is used to estimate psycholinguistic lexical variables, such as token frequencies (Brysbaert and New, 2009), and information content with language modelling (Chen and Goodman, 1999). The compilation of this written corpus will be described in Section 2.3. For the purpose of devising the normalisation of the orthographic transcriptions, the normalisations need to be scalable to a large written English corpus, meaning that they have to be fully automated and do not require manual corrections.

2.1.2.1.1 Capitalisation Orthographic words in English are capitalised at the beginning of each sentence (e.g. *This is my name*), at the beginning of proper names (e.g. *Kevin*), in acronyms (e.g. *HSBC*) as well as at the beginning of certain pronouns (e.g. *I*).

2.1.2.1.1.1 Misperception corpora In the misperception corpora, it is not always clear if the reported utterances are full sentences or just part of a sentence. If they are full sentences, then the first word should be capitalised according to English orthographic conventions. If they are not full sentences but part of a sentence, then the capitalisation of the first word is incorrectly assigned. In some instances, the reported orthographic transcriptions simply have one word. This may be due to a reporting bias, since the collectors (or those who noticed the occurrence of a misperception) might only remember the most noticeable difference between the intended utterance and the perceived utterance.

2.1.2.1.1.2 Written language corpus A written corpus was selected to serve as a "control" for subsequent analyses of the misperception corpus. This corpus also requires normalisation of its orthography like with the misperception corpus. The corpus consists of TV and film subtitle texts. The reason for selecting such a corpus will be documented later in Section 2.3. Like the misperception corpus, the controlled written corpus also suffers from inconsistent capitalisation. Firstly, the writers of the texts could create these capitalisation inconsistencies. Secondly, line breaks that do not indicate the start of a new sentence could be another cause. These kinds of line breaks are often used in subtitle texts since there is a space limit to how much text can be fitted on the screen and the first word in the portion of the text after the break is sometimes incorrectly capitalised. Thirdly, headings and titles are sometimes capitalised inconsistently, with variations including capitalising every single word, every letter in every word, the first word, or all words except those with

three or fewer letters.

These inconsistent capitalisations can cause natural-language-processing tools, such as Part-Of-Speech taggers, to yield poorer and more inconsistent performances (Halteren, 2000). When processing written texts, almost all natural language processing tools have to first tokenise the texts, i.e. identify words, with the capitalised and lower-case versions of a word treated as two distinct words. While there are cases where the two forms are indeed distinct, (e.g. *Don* – a proper name vs. *don* – a verb meaning to put on a item of clothing), more often the two forms are not in fact distinct lexical items but simply an error caused by conventions of capitalisation.

Even if the conventions of capitalisation are consistently applied, the capitalisation can still be a poor cue for whether a capitalised form is indeed a distinct lexical item from the lowercase form. This is apparent if we just consider the capitalisation of the first word in an utterance. There is no reason to believe that the forms "This" and "this" are two distinct lexical items, purely because the first form is found in a sentence initial position. If we were to treat capitalised and lowercased forms as distinct words then the size of the lexicon would be likely to double, since almost all words can be placed in a sentence initial position.

In sum, the treatment of capitalisation will influence the process of tokenisation. This will have an impact on any estimates derived from the control written corpus, since, for instance, this will affect the quality of the estimated token frequencies, since the count of a word would be "shared" between the two forms (capitalised and lowercased).

2.1.2.1.1.3 Normalisation by lowercasing Having discussed the issues of capitalisation in orthographic transcriptions, it is clear that there is a need for a solution. In this thesis, I will take a pragmatic approach to solving this issue, which is to simply make all orthographic words entirely lowercase. This will be applied to both the misperception corpus and the control written corpus. It is important to apply

the same treatment to both sets of corpora, since the lexical properties extracted from the control written corpus will be used in the analyses of the misperception corpora.

There are drawbacks to this approach, such as the distinctions that will be lost for lexical items that are distinguishable by their capitalisation (e.g. *Don* and *don*) and this approach will affect the identification of proper names, but this drawback could be mediated by using a Part-Of-Speech tagger and a Named-Entity Recognition tagger (Finkel, Grenager, and Manning, 2005) that are case-insensitive to identify proper names. However, neither of these taggers were used to enrich the corpora in this thesis, and this also means that I did not investigate the process of speech perception of proper names. Studies on proper names have shown that they are perceived differently from non-proper names such as content or function words (Valentine, Brennen, and Brédart, 1996), including in the retrieval process. Specifically, naturalistic and experimental studies on tip-of-the-tongue phenomenon have suggested that people are more likely to have retrieval difficulties with proper names than non-proper names (Valentine, Brennen, and Brédart, 1996, Ch. 5). Furthermore, the recognition of proper names has been shown to be related to memory of known individuals, and moreover the token frequency effect (more frequent words are processed faster) will only hold for proper names if the experimental task does not require access to memory of a known individual (Valentine, Brennen, and Brédart, 1996, Ch. 4). In sum, there is clear evidence that proper names are processed differently than non-proper names. Proper names are not tagged in the written corpus, but proper names are tagged manually in the misperception corpus.

Future work should be done to enrich both the misperception corpus and the control written corpus by tagging them with Part-of-Speech tags. In addition, the proper names in the misperception corpus should be further tagged with information regarding whether the processing of the proper name in each instance is likely to

require memory of a known individual by using meta-data in the corpus.

2.1.2.1.2 Punctuation marks This section examines the treatment of punctuation marks in the orthographic transcriptions, specifically full stops, apostrophes and hyphens, since they have an impact on the tokenisation process as discussed in the Capitalisation section.

2.1.2.1.2.1 Full stops Full stops are used to denote the end of a sentence, in ellipses (typically three full stops), separating the letters in initialisms (e.g. U.S.A.) and after abbreviations (e.g. etc.).

To normalise the use of full stops in the orthographic transcriptions, a simple treatment would be to replace them with spaces. However, special treatments are needed when full stops are used within an initialism. There are three possible options. The first option would be to not leave full stops within initialisms, e.g. *U.S.A.* would therefore be unchanged. The second option would be to replace them with spaces, such that the letters in the initialisms are treated as three orthographic words, e.g. *U.S.A.* would therefore become *U S A*. The third option would to be remove them, e.g. *U.S.A.* would become *USA*.

The first option may lead to potential issues with the consistency of using full stops in acronyms. For instance, the Massachusetts Institute of Technology is often represented through the three-letter acronym consisting of the letters, *M*, *I* and *T*, and it could be written as *M.I.T.* or *MIT*. While it is possible to manually check for inconsistencies in the misperception corpus, this is not possible for the control written corpus due to its size.

The second option may lead to potential problems with the identity of the acronym. By replacing the full stops with spaces, multiple words would be created from each acronym. If an initialism was treated as multiple words, then the identity of such acronyms would no longer be encoded.

The third option may lead to issues with creating mergers with non-acronyms. It is possible that by removing the full stops, together with ignoring cases (as mentioned in the Capitalisation section), homographs could be created. For instance, *Phase Impenetrability Condition* has the initialism *P.I.C*; after the treatment of removing the full stops and lowercasing, it would become *pic* which is a word for picture.

I have opted for the third option, which is to remove the full stops within initialisms. The reason is that the chances of creating homographs are low because they are restricted to initialisms that have legal English orthotactics. Furthermore, the third option is likely to ensure greater consistency than the third option.

In sum, full stops will be replaced by spaces if they are adjacent to at least one non-space unit or a line break, e.g. "My name is John. My name is John . My name is John." would become "My name is John My name is John My name is John ". All other full stops will be removed (without replacement), e.g. "U.S.A." will become "USA".

2.1.2.1.2.2 Apostrophes Apostrophes are primarily used for contractions (e.g. *can't – cannot*), possession (e.g. *the dog's bone*), single quotation marks (e.g. 'apple') as well as denoting an alternative pronunciation (e.g. rappin' – which is the word *rapping* with alveolarisation of *-ing*) (for an extensive discussion on the use of apostrophes, Quirk et al. (1985)).

The use of apostrophes for possession is particularly variable after a noun ending with an orthographic "s". The standard convention is to have the apostrophe before the "s" if the noun is singular (e.g. *the dog's bone*), but after the "s" if the noun is plural (e.g. *the dogs' bones*). When it comes to proper names that ends of an "s", there are more variations between writers. If the noun is plural, then the convention is to first pluralise the proper name (e.g. *Jones* would become *Joneses*), before appending an apostrophe (e.g. *Joneses' house*). If the noun is singular, then the convention is to append an apostrophe followed by an "s" (e.g. *Jones's house*).

However, these conventions are often broken in various ways: 1) missing apostrophes (e.g. *the dogs bone* for *the dog's bone* or *the dogs' bone*), putting an apostrophe before the "s" for a plural noun (e.g. *the dog's bone* for *the dogs' bone*), putting an apostrophe after the "s" for a singular noun (e.g. *the dogs' bone* for *the dog's bone*), adding an additional "s" for pluralised proper names (e.g. *Joneses's house* for *Joneses' house*) and many others (Hook, 1999).

Given the variations in the use of apostrophes, it is apparent that some kind of normalisation is needed. The first option is to remove any apostrophes that are between a) two space characters, or b) one space character and a line break and c) two line breaks. The first option is the most conservative and effectively only removes apostrophes that are adjacent to spaces/line breaks. This option relies on the assumption that the level of inconsistency is low in the misperception corpus and the control written corpus and does not remove the apostrophes that are functioning as single quotation marks.

The second option is similar to the first option, but in additional it uses regular expressions to remove pairs of single quotation marks. Concretely, a regular expression removes occurrences of two consecutive apostrophes such that the first apostrophe appears after a space or a line break and before a non-space character, and the second apostrophe appears after a non-space character and before a space/a line break. This option is problematic when applied to the control written corpus, which contains a large amount of typographical errors. For example if the first quotation mark is a double quotation mark and the second quotation mark is a single quotation mark (an apostrophe), then the regular expression will fail, and if the word before the second quotation mark is a noun that ends with an "s", then the second quotation mark will therefore indicate that that noun is a plural noun with an apostrophe denoting possession. Another way for the regular expression to fail is an accidental insertion of an apostrophe before a word; for instance, *'Kevin has the*

dogs' bones. where the first apostrophe is simply a typographical error. Just like the first option, this option relies on the assumption that the number of inconsistencies in the misperception corpus and the control written corpus is low, but this is unlikely to be true. In one study on written errors by college students (Haswell, 1988), the error rate for possessives was as high as 50 percent.

The third option is to keep any apostrophes that are between two non-space characters, with all other apostrophes removed or replaced by spaces. This option would remove the apostrophes in cases where the apostrophes come after the "s", e.g. *the dogs' bone* would become *the dogs bone* as well as the single quotation marks, but keep them when they are placed before the "s" and after a noun (e.g. *the dog's bone*). The drawback of this option is that the denoting of possession is maintained for singular nouns but not for plural nouns.

The fourth option is to keep any apostrophes that are a) between two non-space characters, or b) after an "s" and before a space or a line break. All other apostrophes are removed or replaced by spaces. This option would keep the apostrophes in cases where the apostrophes come after the "s", e.g. *the dogs' bone.* But it would also accidentally keep the single quotation marks that are placed after a word ending with an "s", e.g. *'apples'* would become *apples'.* While this option retains possession for both singular and plural nouns (unlike the third option), it over-generates the number of possessive plural nouns.

Having briefly discussed the possible options, it is clear they each bring their own problems. I opted for the third option in this thesis. This option has the drawback of maintaining the possession for singular nouns but not plural nouns. The main implication of this for subsequent analyses is that the distinction between a plural noun and a possessive plural noun is ignored, but given that none of the corpora will be part-of-speech tagged nor will syntactic aspects of misperception be considered in this thesis, the loss of this distinction is not expected to be problematic. Regarding

pronunciation, the two forms are in fact identical, with the exceptions of proper names ending in "s".

2.1.2.1.2.3 Hyphens One of the uses of hyphens in orthographic transcriptions is to indicate compounds, e.g. *bell-tower*. The issue here is how the hyphens should be treated – should they be kept, removed or replaced with spaces?

There is individual variability in the use of hyphens (Kuperman and Bertram, 2013). In this paper, the authors analysed three-way alternations of orthographic choices for two-constituent compounds – concatenated (e.g. *belltower*), spaced (e.g. *bell tower*) and hyphenated (e.g. *bell-tower*) variants. Using both diachronic and synchronic data (from behavourial tasks), they suggested that the orthographic choice is a function of orthographic, statistical, and semantic properties of the compounds' constituents. The concatenated form is taken to be lexicalised. The authors explored the precise nature of the route of lexicalisation, and found that the majority of the alternating compounds only consisted of the spaced and hyphenated forms, or only the spaced and concatenated forms. A substantial number of compounds consist of all three forms, and those that consist of only the hyphenated and concatenated forms are relatively rare.

Together, these distributional patterns suggest that the hyphenated forms and the concatenated forms co-occur only when the spaced forms are present, and that the alternations occur exclusively between the spaced forms and the two other forms, without the intermediate stage where the spaced forms are lexicalised via the hyphenated forms. The route of lexicalisation seems to suggest the hyphenated forms pattern with the concatenated forms, such that they are both single lexical units, supporting the decision that the hyphens should be removed, and that the hyphenated words should become concatenated.

However, considering only the alternating compounds that consist of all three forms, the route of lexicalisation appears to be different, such that the hyphenated

forms (being the least frequent) are firstly changed to the spaced forms, which in turn are changed to the concatenated forms. This route of lexicalisation conflicts with the aforementioned one, and therefore it casts doubt on the patterning of hyphenated forms and concatenated forms.

A different study by van Heuven et al. (2014) also faced with this issue of the treatment of hyphens in their complication of a word frequency corpus, SUBTLEX–UK, and specifically whether the hyphen should be removed or not – and they did not consider the concatenated form. In this study, they based their decision on behavourial data from the British Lexicon Project (Keuleers et al., 2012) which consist of 28 thousand lexical decision times. They correlated the lexical decision times with either the token frequencies of the dehyphenated forms (bi-gram frequencies) or those of the hyphenated forms. They found that the frequencies of dehyphenated forms captured significantly more variance in the data (5% more) than those of the hyphenated forms. Based on this result, they decided to replace the hyphens with spaces in their corpus.

Considering the conclusions one could draw from both studies, it seems the distributional patterning between the hyphenated forms and concatenated forms cannot provide substantial support that the two are treated as one and the same. In fact, behavourial data suggests that, at least in terms of the effect of token frequency on visual word recognition, the hyphenated forms should be treated as the spaced forms. Therefore for this thesis, I opted for the dehyphenation option, where hyphens are replaced with spaces.

2.1.2.1.3 Abbreviations Another common step in text normalisation is to expand numeric and symbolic abbreviations. By numeric abbreviations, I mean numbers that are written in digits as opposed to in alphabetic letters; for instance, *200* can be expanded into *two hundred*. By symbolic abbreviations, I mean signs such as dollars and pounds; for instance, $200 can be expanded into *two hundred dollars*.

Other title abbreviations such as *Mr.* and *Dr.* can be expanded into *Mister* and *Doctor*.

This process of expansion is useful for normalisation across texts, because sometimes these abbreviations are already expanded, so normalisation will ensure consistency. It can better reflect the number of spoken/perceived words, and furthermore it can yield a better correspondence between the word order in the orthographic transcriptions and the word order in the phonetic transcriptions, especially in cases of symbolic abbreviations where the currency signs are often spoken last, e.g. $200 is pronounced as*two hundred dollars*, not *dollars two hundred*.

This is in fact a non-trival task and is extremely challenging. For instance, the way numeric abbreviations are spoken can vary – years (e.g. in 1999) are often spoken differently from non-year contexts (e.g. 1999 cows). For an extensive overview for this text normalisation problem, see Sproat et al. (2001).

Due to the complexity of the problem, I employed an existing toolkit, *nsw expand*, which is part of the text normalization tools for the Festival speech synthesis system (Black, Sproat, and Chen, 2000). The tool has four domain models, each of which is particularly suitable for a specific genre of texts. The four domains are news, classified ads, an email-like technical mailing list and recipes. I chose the default model, news, for this thesis. This decision was made based on small scale testing with both the misperception corpus and the control written corpus and I found that the news model yielded better results. The output for the misperception corpus would be checked manually, while the output for the control written corpus would not be checked manually due to the size.

2.1.2.2 Phonetic transcriptions

Of the five corpora, two lack any phonetic transcriptions – they are the Bond's (1999) corpus and Bird's (1998) corpus. The remaining three corpora have reported

phonetic transcriptions that vary in multiple ways.

Firstly, the transcriptions differ in how broad/narrow they are, with Nevins's corpus being the most narrow, while Labov's (2010) and Browman's (1978) are more phonemic. Secondly, the amount of transcriptions also varies: in Labov's (2010) corpus, only the misperceived phones were transcribed; in Browman's (1978) corpus, only the words that were misperceived (intended and perceived words) were transcribed; in Nevins's corpus, some instances have the whole utterance transcribed, including the correctly perceived portions. Thirdly, the transcription system used by each corpus also varies – Nevins: IPA, Labov, 2010b: own system and Browman, 1978: ARPAbet (with IPA mapping information). Fourthly, the transcription details within each corpus can also differ. This is apparent in Nevins's corpus, because the collectors numbered around 20–30 each year (for two years) and more collectors typically lead to more inter-transcriber variations, e.g. the r-coloured NURSE vowel was transcribed in multiple ways – ɚ, ɜˠ, əɹ and ɜɹ.

Given the high within and across corpus variation, it is clear that the combined corpus needed to be re-transcribed using one transcription system with the same level of phonetic detail. The original transcriptions can be used as a guide, especially in cases when there could be multiple ways of transcribing a particular word, due to factors such as vowel reduction, homographs, rhoticity and others. Furthermore, the transcriptions would be richer if they are able to reflect the rich dialectal information we have, rather than simply being transcribed phonemically. In the following section, I will describe how the dialectal information was used and the assumptions made in instances where the dialectal information is missing or simply not available during the collection. The precise details of the phonetic transcriptions can be found in Section 2.2.

2.1.2.2.1 Accent classification

2.1.2.2.1.1 Browman We do not know the exact accent/dialect background of the interlocutors for each misperception instance in the corpus. However, since we know that the collectors were from Los Angeles, Columbus Ohio, and Boston, Massachusetts, the assumption is being made that the utterers and perceivers are speakers of American English. In terms of accent groups (see Table 2.5 for a classification of major regional speech areas in America), Los Angeles and Ohio are classified as *General American*, while Massachusetts is classified as *New England*. Given that the published corpus did not indicate which data points were collected by which team of collectors, it is not possible to estimate the accent group precisely. For the purpose of this thesis, I will assume that the utterers and perceivers had a General American accent.

2.1.2.2.1.2 Bird Although the author did not publish any dialectal background of the interlocutors of each misperception, by observing the individual data points, many British related words or proper names were provided. For instance, *Look North* is likely to be referring to *BBC Look North*, a regional television channel in Yorkshire, UK; *Blakeney Place* is a place in York, UK; *Haworth* is a village in West Yorkshire, UK; *80p*, which is a standard way of referring to 80 pence; *Biggles* was a popular UK series of youth-oriented adventure books written by W. E. Johns; *Walmgate* is a name used by establishments in York, UK (referring to *Walmgate Bar* which is a medieval gateway to the city of York, UK) and many others. Together with the fact that the author was affiliated with the University of Newcastle at the time when the paper was published, it is reasonably clear that the place names mentioned above are indeed those in Yorkshire, also in Northern England, like Newcastle.

While it seems likely that the utterers and perceivers may have a Northern English accent (given the multiple references to Northern England), for the purpose of this thesis, I will assume the utterers and perceivers involved in the data have a Standard Southern British English accent (see section 2.2.7.1). Given the widespread dialectal

levelling amongst British accents (Kerswill, 2003), the dialectal difference between Northern and Southern British English accents should not have a major impact on the overall analyses. However, caution should be taken when interpreting the results, especially with those involving North-South dialectal differences, such as the STRUT vowel, which is well-known to be realised differently.

2.1.2.2.1.3 Labov Most of the misperceptions the corpus included information about the utterers' and perceivers' places of origin. In cases where the information was available, the accent groups are assigned to the interlocutors according my classification system (see Table 2.5 for a classification of major regional speech areas in America).

There are numerous cases where background information is missing (either known but not reported or not available to the collectors). For the purpose of the thesis, I will assume they have a General American accent.

2.1.2.2.1.4 Bond The accent/dialect backgrounds of the interlocutors were not available. Although not explicitly mentioned, the corpus was most likely collected in Ohio where the author was based. This means that the collectors, the author, students, friends, and colleagues are most likely to be from Ohio or nearby in the North Central region of United States.

In terms of accent groups (see Table 2.5 for a classification of major regional speech areas in America), the North Central regions are classified as *General American*. For the purpose of this thesis, I will assume that the utterers and perceivers in all the data points (except for a few data that were indicated to have a British speaker) have a General American accent.

2.1.2.2.1.5 Nevins In this corpus, extensive information about the accent/dialect backgrounds of the interlocutors was available. Where the information

was available, the accent groups are assigned to the interlocutors according my classification system (see Table 2.5 for a classification of major regional speech areas in America). In cases where the background information is missing (either known but not reported or not available to the collectors), I have assumed they have a General American accent for the purpose of the thesis.

2.1.2.3 Meta-data

Having discussed how the orthographic and phonetic transcriptions were standardised, we now examine more closely how the normalisation was approached with the meta-data from the five corpora – specifically the geographic location, the age and the gender of the interlocutors.

2.1.2.3.1 Geographic location Labov's (2010) corpus and Nevins's corpus both contained geographic information about where the interlocutors were from. This information was used to identify the *Country* and *State*. The term *State* has been used since the majority of the interlocutors were from USA. For interlocutors from other countries, the equivalent regions are used instead e.g. the provinces of Canada.

2.1.2.3.2 Age Many instances from Nevins's corpus include the exact age of the interlocutor and these figures were used without further modifications. Sometimes the reported age was not exact value, e.g. *mid-thirties*, *40 something* and *50s*. In these fuzzy cases, I have simply take the middle value, e.g. 35, 45 and 55 for the examples above respectively.

Labov's (2010) corpus did not contain much age meta-data. However, there were 14 recurring interlocutors, and many of whom were also collectors or academics from the same universities as the collectors. The full name of the interlocutors was often provided. Using their full names, I would look for their date-of-birth from any information available on the internet, such as their CVs. If the date-of-birth was

not available for a recurring interlocutor, I would use their graduation years from their undergraduate degree, or doctorate degree as an estimate of the year of birth. Furthermore, the corpus reported the date of occurrence of each misperception. The age of the recurring interlocutors at the time of each misperception was calculated by using the year/date of birth and the date of occurrence of the misperception instances.

2.1.2.3.3 Gender Many instances from Nevins's corpus include the gender of the interlocutors reported and this information was used directly. Although the gender information of the interlocutors was not always specified explicitly, it could also be derived from the names and the pronouns reported for each instance.

2.1.2.3.4 Slip type Using any meta-data provided across these corpora, I further tagged each data point as being a Mondegreen (misheard lyrics) or not.

2.1.3 Combined Corpus

The outcome of the compilation of all the existing corpora is a combined corpus with orthographic transcriptions, phonetic transcriptions and meta-data. Figure 2.10 shows a snapshot of the combined mega corpus.

2.1.4 Summary

This section began with details for the collection process, data structure, orthographic transcriptions, phonetic transcriptions and meta-data of the existing naturalistic misperception corpora of English. I then proceeded to describe how I compiled these existing corpora into one mega corpus of misperception, focusing on the normalisation of the orthographic transcriptions, phonetic transcriptions and meta-data. Special attention was paid to the orthographic transcriptions to ensure the

Serial Number	Corpus Origin	Orthography Intended Entire Sentence	Orthography Perceived Entire Sentence	IPA Intended Entire Sentence (With Tapping and Aspiration)	IPA Perceived Entire Sentence (With Tapping and Aspiration)	Language
1	Nevins 2010	Shorty want a thug	Shorty want a hug	|| ˈʃɔɹ.ɾɪ ˈwɑnt ə ˈθəg ||	|| ˈʃɔɹ.ɾɪ ˈwɑnt ə ˈhəg ||	English
2	Nevins 2010	like a g six	lick a cheesesteak	|| ˈlajk ə ˌdʒij ˈsɪks ||	|| ˈlɪk ə ˈtʃijz.ˌstejk ||	English
3	Nevins 2010	i love seedless grapes	i love cinnamon sticks	|| aj ˈlʌv ˈsijd.ləs ˈgɹejps ||	|| aj ˈlʌv ˈsɪ.nə.mən stɪks ||	English
4	Nevins 2010	shoot me now	shoot me cow	|| ˈʃɯwt mi ˈnæw ||	|| ˈʃɯwt mi ˈkʰæw ||	English
5	Nevins 2010	i just have an older sister	i have seven older sisters	|| aj dʒʌst hæv ən ˈowl.dəɹ ˈsɪ.stəɹ ||	|| aj hæv ˈsɛ.vən ˈowl.dəɹ ˈsɪ.stəɹz ||	English
6	Nevins 2010	the phonological processes involved	the psychological processes involved	|| ðə ˌfow.nə.ˈlɑɑ.dʒɪ.kəl ˈpʰɹɑɑ.sɛ.sɪz ɪn.ˈvɑɑlvd ||	|| ðə ˌsaj.kə.ˈlɑɑ.dʒɪ.kəl ˈpʰɹɑɑ.sɛ.sɪz ɪn.ˈvɑɑlvd ||	English
7	Nevins 2010	Shawty is a killa	Shawty has a camera	|| ˈʃɑ.ɾɪ ɪz ə ˈkʰɪ.lə ||	|| ˈʃɑ.ɾɪ hæz ə ˈkʰæm.ɹə ||	English
8	Nevins 2010	is this paper for intro	is this paper pharyngeal	|| ɪz ðɪs ˈpʰej.pəɹ fəɹ ˈɪn.tɹow ||	|| ɪz ðɪs ˈpʰej.pəɹ fɔ.ˈɹɪn.dʒi.əl ||	English
9	Nevins 2010	have you ever seen the movie Intolerable Cruelty	have you ever seen the movie Charlie Saint Cloud	|| həv jʉ ˈɛ.vəɹ ˈsin ðə ˈmʉw.vɪ ˌɪn.ˈtʰɑ.lə.ɹə.bəl ˈkʰɹɘwl.ti ||	|| həv jʉ ˈɛ.vəɹ ˈsin ðə ˈmʉw.vɪ ˈtʃɑɹ.lɪ ˌsejnt ˈkʰlæwd ||	English
10	Nevins 2010	the collaborative learning center where people go to learn	the collaborative learning center where people go to burn	|| ðə kʰə.ˈlæ.bə.ɹə.ɾɪv ˈlɜɜɹ.nɪŋ ˌsɛn.təɹ wɛɹ ˈpʰij.pəl ˈgəw tʰə ˈlɜɜɹn ||	|| ðə kʰə.ˈlæ.bə.ɹə.ɾɪv ˈlɜɜɹ.nɪŋ ˌsɛn.təɹ wɛɹ ˈpʰij.pəl ˈgəw tʰə ˈbɜɜɹn ||	English

Figure 2.10: The combined corpus: orthographic and phonetic transcriptions and meta data (many of which are not shown here)

normalisation process could be automated with greater ease, such that our "control" corpus (a mega text corpus of English) can be processed consistently and be used in conjunction with the misperception corpus in future analyses. The overall process was documented. The following section, Section 2.2, will focus on the details of the phonetic transcriptions.

2.2 English naturalistic corpora – phonetic transcriptions

Transcription can be imprecise. Many have argued against the use of transcription or similar symbolic representation of speech sounds for linguistic analyses (Kerswill and Wright, 1990; Harrington, 2010). The arguments are often focused on the unreliability of auditory transcription even by trained phoneticians, the lack of fine-grained phonetic detail and the inherent notational limits of transcription systems. Concretely, Kerswill and Wright (1990) found that auditory transcriptions by phoneticians are unreliable when compared to articulatory data (electropalatography) and inconsistent within transcribers. Furthermore, Harrington (2010, Ch. 1) argued that "an auditory transcription is at best an essential initial hypothesis – never an

objective measure". While these arguments are well-founded, phonetic transcriptions are the next best option in cases when there are no acoustic recordings, and transcribing from orthography is still possible and has widely been used in education (Wells, 1996), speech synthesis (Black et al., 2002), and linguistic analyses. Furthermore, phonetic transcriptions have been used extensively by dialectologists and can capture phonetic details that are well reflected in acoustic signals (Wieling, Margaretha, and Nerbonne, 2012) and listeners' perceptions (Wieling et al., 2014).

In practice, phonetic transcriptions of misperception data are largely unavoidable. While we could, in theory, have the acoustic recordings of the intended utterances (although this is likely to be impractical in the case of naturalistic data, since one would need to carry a recording device at all times), it is impossible to acoustically record one's speech perception.

The following sections will document the transcription process used for the naturalistic misperception corpus. Section 2.2.1 will look at the databases of English pronunciation used as the source of pronunciation for the transcriptions and for how to decide amongst alternative pronunciations. Section 2.2.2 will describe the level of detail for the segmental transcription and the inventory of IPA symbols that was used. Section 2.2.3 will describe the level of detail for the prosodic transcription. Section 2.2.4 will review options for the syllabification of segmental transcriptions, justify the selection of one of these methods, and develop a principled way of selecting valid onsets for the selected method. The remaining sections will establish the dialectal transcription of the corpus. Starting with Section 2.2.5, the question of which accent should be transcribed for the intended and perceived utterances is addressed. Section 2.2.6 will move on to establishing the dialect classification for American English based on geographic locations. Finally, Section 2.2.7 will establish a vowel set for each major dialect in the corpus.

2.2.1 Choices of pronunciation

This section will first document the source of pronunciation and outline the transcription preferences when alternative interpretations are possible.

2.2.1.1 Databases of English pronunciation

The Longman Pronunciation Dictionary (henceforth LPD) (Wells, 2008) and Longman Dictionary of Contemporary English (Fox and Combley, 2009) were used extensively as a reference for pronunciation. While there are other pronunciation dictionaries such as CELEX (Baayen, Piepenbrock, and Gulikers, 1995) and CMUDICT (Weide, 2014), these only cover either British or American English and most of their entries contain only one pronunciation variant. LPD covers both British (Received Pronunciation) and American English (General American) as well as providing multiple variants for the majority of the entries, e.g. the word *chocolate* has the following pronunciations – in American English [ˈtʃɔːklət], [ˈtʃɑːklət] and in British English [ˈtʃɒklət], [ˈtʃɒklɪt], [ˈtʃɒkələt] and [ˈtʃɒkəlɪt]. The rich phonetic variations of LPD allow a better quality transcription, and therefore LPD was chosen as the core pronunciation database for this study.

In addition, the following online pronunciation dictionaries of English were consulted when words could not be found in LPD: *howjsay* (`http://www.howjsay.com/`) and *Forvo* (`http://www.forvo.com/`) contains audio recordings of words/phrases without phonetic transcriptions. The former is only British English, while the latter covers over 240 languages. In addition, four dictionaries were used for the phonetic transcriptions: *Cambridge Dictionaries Online* (`http://dictionary.cambridge.org/`), *Oxford Dictionaries* (`http://www.oxforddictionaries.com/`), *Merriam Webster* (`http://www.merriam-webster.com/`) and *The FreeDictionary.com* (`http://www.thefreedictionary.com/`). For an overview of these online dictionaries, please see Kyprianou (2009) and particularly for *Forvo*, see Grieser (2010).

2.2.1.2 Pronunciation preference

The transcriber is faced with many alternative pronunciations for each word and combinations of words for each sentence when performing phonetic transcriptions from orthographic texts.

A simple heuristic set was used for selecting between the alternative pronunciations. Firstly, weak forms were preferred in order to better capture the more conversational/informal nature of the majority of the corpus. Secondly, the intention was to minimise the number of segmental mismatches, in order to avoid creating mismatches that are due to alternative pronunciations. These two heuristics can be automated. First, all possible alternative pronunciations of the intended utterance and the perceived utterance were generated. Second, all alternative pronunciations of the intended utterance were then paired with those of the perceived utterance. Third, the number of segmental mismatches were calculated for each pair from the overall length of the sum of the two utterances, and the total number of schwas across both utterances. Finally, the optimal pair was chosen as the one with the minimal number of segmental mismatches; if more than one pair satisfied this criterion, the one with the shortest length of the sum of the two utterances was chosen; if more than one pair was still available, the one with the highest number of schwas was chosen; finally if more than one pair was still available, one would be arbitrarily selected to be the optimal pair. This automated method was not used for the naturalistic misperception corpus. Instead, during manual transcription, this set of simple heuristics was generally followed instead.

2.2.2 Segmental transcription

2.2.2.1 Levels of segmental transcription

The International Phonetic Alphabet (henceforth IPA) (International Phonetic Association, 1999) was chosen as the convention to follow for the transcription in this study. The level of transcription followed was neither entirely phonemic nor phonetic. Lexical stress was transcribed with the primary [ˈ] and secondary stress marks [ˌ]. Vowel length was either short or long, using only the length mark [ː] for long, and no additional mark for short, but as we will see later in Section 2.2.7.1.3, the length mark [ː] is replaced with the preceding segment and did not appear in the transcription. Syllabic consonants [n̩] were not transcribed, with schwas instead assumed to be fully realized.

Tapping was applied to the dialects that have this process. I assumed complete neutralisation of /d/-tapping and /t/-tapping and the IPA standard symbol for a tap [ɾ] was used. To identify the taps, a search was done to extract any words that contain /t/ or /d/ that were preceded by any vowels (including rhotic ones) and were followed by a set of vowels including a) the "weak vowels" HAPPY, COMMA and LETTER, b) any unstressed FLEECE, since in some cases (within and across accents), it is used for HAPPY, and c) any unstressed GOAT, since some of the GOAT vowels can behave like a weak vowel and often undergo weakening to become a schwa and therefore trigger /t/-/d/ tapping, e.g. *fell[ow]* as *fell[ə]*. They are essentially all the Germanic words spelt with letters "ow" (e.g. *tomorrow*) and loans spelt the letter "o" (e.g. *photo*).[1] The extracted word list was then manually checked by a linguist who is a native speaker of American English before being used in the tapping conversion.[2] Finally tapping across words was applied by converting word-final /t/s and /d/s that were preceded and followed by any vowels (including rhotic ones), e.g. *I nee[ɾ] a pen.*

[1] I thank Prof. John Harris for his expert input on formulating these rules as well as pointing out the existence of the "fake" [ow]s

[2] I thank Prof. Andrew Nevins for his input as a native speaker.

A three-way contrast for stops is encoded in the transcription – aspirated voiceless, unaspirated voiceless and voiced stops. Aspiration [ʰ] was transcribed for the aspirated voiceless stops [pʰ, tʰ, kʰ]. The unaspirated voiceless stops are [p, t, k]. The voiced stops are [b, d, g]. To identify the aspirated voiceless stops, I applied a rule-based conversion, stating that /p, t, k/ that are aspirated, if a) they are at the beginning of a word, or b) before a stress (primary or secondary) vowel but not after /s/, otherwise they are unaspirated voiceless stops.[3]

The following section will briefly describe the inventory of IPA symbols chosen for transcribing this corpus which was adjusted for handling multiple accents. The specific details on how these symbols are used to capture different vowel sets can be found in Section 2.2.7.

2.2.2.2 Inventory of IPA symbols

The 28 consonant phones are tabulated in Table 2.1. They are as follows: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, j, w, pʰ, tʰ, kʰ, ɾ].

The 16 vowel phones are tabulated in Table 2.2. They are as follows: [e, ɛ, a, ɑ, ɒ, ʌ, ɔ, o, u, ʉ, ə, ɜ, ɪ, ʊ, æ, i]. This set of vowels is typically used to transcribe General American and the RP pronunciations; for instance, in LPD (Wells, 1990), with the exception of the phone [ʉ] which is a centralised [u]. [ʉ] was included since many accents (Standard Southern British English, American Southern English, Philadelphia English and others) have undergone GOOSE-fronting, resulting in a centralised [u]. Its inclusion will increase the accuracy of the transcriptions, and since it exists in multiple accents that the corpus covers, it is less likely to lead to serious data-sparsity issues in later analyses.

[3]I thank Prof. John Harris for his expert input in formulating these rules.

	Bilabial	Lab. dent.	Dental	Alveolar	P-alveo.	Retroflex	Palatal	Velar	Uvular	Pharyng.	Glottal
Plosive	p(ʰ) b			t(ʰ) d				k(ʰ) g			
Nasal	m			n				ŋ			
Trill											
Tap/Flap				ɾ							
Affricate					tʃ dʒ						
Fricative		f v	θ ð	s z	ʃ ʒ						h
Lat. Fric.											
Approx				ɹ			j				
Lat. appr.				l							
Other phones:	w – Voiced labial-velar approximant										

Table 2.1: Consonant chart used in transcription

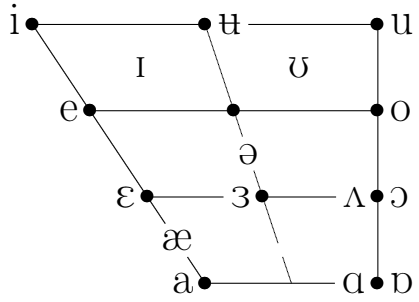


Table 2.2: Vowel chart used in transcription

2.2.3 Prosodic transcription

2.2.3.1 Levels of prosodic transcription

The placement of stress is particularly problematic when transcribing connected speech and this section will discuss the treatment of stress in the corpus. This issue was discussed by Wells (2011), which described different levels of prosodic transcription, and they are summarised in Table 2.3.

Level	Descriptions
1	Add stress marks to monosyllabic words, but only to content words and not function words. Polysyllabic words have stress marks.
2	Remove stress marks on the 2nd element of compounds and repeated words. Apply stress shift rules (e.g. *fif['']teen*, but *['']fifteen ['']people*. Add stress to function words that are used contrastively.
3	Divide the utterance into intonation phrases.
4	Underline nuclear tones.
5	Convert nuclear accent marks into tone marks.

Table 2.3: Levels of prosodic transcription

From the five levels of prosodic transcription outlined by Wells (2011), level 1 was chosen to be the prosodic transcription of this corpus. The reason for choosing to incorporate only the lowest level of prosodic details is as follows.

Level 4 and level 5 aim to transcribe tone marks, but it is clear that without an audio recording of the spoken speech the estimate of tone marks will be poor because a single orthographic sentence can be read with multiple intonational patterns (Wells, 2006). Furthermore it is also not possible to know how the intonational pattern of the perceived utterance should be transcribed, since the perceivers in most cases were not trained phoneticians, and even if they were (such as those in the Labov corpus), they did not transcribe them in their reports. Level 3 would divide the connected speech into intonation phrases, but since we are rejecting level 4 and level 5, level 3 is no longer essential.

Level 2 aims to handle a) the placement of stress on compounds, b) repeated

words should be unstressed, c) stress shift rules (Wells, 2006), and d) contrastive stress of function words. This level of transcription was attempted at an early stage of the corpus compilation, but soon it became clear that this involves a lot of ambiguities and arbitrary decisions.

Firstly, it is not always clear what constitutes a compound and what is the stress pattern. While compounds in English are often single-stressed, such as *ˈbedtime*, the complication arises with open compounds. Open compounds, such as *ˈrunning shoes*, can be confused with a phrase consisting of adjective plus noun such as *ˈrunning ˈwater*, which has lexical stress on both the adjective and noun. Open compounds can be also be double-stressed, and the general guidelines for their identification is that they are often a) proper names of people, roads, and places, such as *ˌNoam ˈChomsky*, b) compounds in which the first element names the place or time, such as *ˌtown ˈhall*, and c) compounds in which the first element names the material or ingredient (but not cakes and juices), such as *ˌapple ˈpie*. Furthermore, compounds can be nested, such as *[[credit card] bill]*, and depending on the nested structure, the stress placement differs (Wells, 2006, pp. 100–106). These complications were evident an attempt at level 2 transcription. In cases where a potential compound was not listed in databases of pronunciation dictionaries, I consulted multiple native speakers and found that they often had different intuitions of whether it is an open compound or an adjective-plus-noun phrase along with associated stress patterns. In fact there is considerable variability in compound stress in English that are determined by structural, semantic, and analogical factors (Plag, 2006).

Secondly, stress shift rules often result in multiple possible stress patterns, and some rules can be optionally applied. One such rule is the *rule of three*. This rule weakens the middle stress of three adjacent stresses, e.g. *ˈA ˈB ˈC* would become *ˈA B ˈC*. However, with longer strings of potential stresses, the number of possible patterns increases, e.g. *ˈA ˈB ˈC ˈD* can have *B* unstressed, *ˈA B ˈC ˈD* or *C* unstressed, *ˈA ˈB*

C ˈ*D* or a repeated application of the rule with both *B* and *C* unstressed ˈ*A B C* ˈ*D*. While I illustrated this rule with a string of letters, all of the above is applicable to normal sentences (Wells, 2006, pp. 228–229).

For these reasons, level 2 transcription was ultimately not adopted for the prosodic transcription. This leaves us with level 1, which is to add stress marks to monosyllabic words, but restricted to content words and not to function words, and polysyllabic words would carry lexical stress.

2.2.3.2 Function words

I manually selected a list of monosyllabic function words compiled by first extracting all monosyllabic words from CELEX (Baayen, Piepenbrock, and Gulikers, 1995) using *Leanlex* (Keuleers, 2006), and manually selecting the words that are *typically* unstressed according to my own intuition. On top of this CELEX list, I also added other monosyllabic words that are typically unstressed. The final list was checked by two native English linguists[4]. I admit fully that some of choices might be ambiguous, e.g. *just* and *such*, and that arbitrary decisions were made in some cases (discussed below). The full list of typically unstressed monosyllabic word is shown below in Table 2.4.

Some of the ambiguous cases involved words that can be both a content word and a function word, depending on their syntactic category. These words are *on, off, in*, which can be either a preposition (function word) or an adjective (content word). Similarly, words such as *such* and *just* are on the borderline between being a function word and a content word (Wells, 2011). While it is possible to subdivide these cases with respect to their syntactic categories, this was done only with the misperception corpus and not the control corpus (see Section 2.3) because the control corpus would have to have been automatically transcribed due to the large amount of data, and

[4]I thank John Harris and Andrew Nevins for their input.

therefore for ease of auto-transcription, this level of detail was sacrificed.

'd	at	he	in	she	there'll	we	who'll	your
'em	be	he'd	is	she'd	there's	we'd	who're	
'll	been	he'll	it	she'll	there've	we'll	who's	
're	but	he's	it's	she's	these	we're	who've	
's	by	her	its	should	they	we've	whose	
'tis	can	him	just	so	they'd	were	why	
'twas	could	his	like	some	they'll	what	why'd	
'twere	d'you	how	lo	such	they've	what's	why'll	
'twill	did	how'd	may	than	thine	when	why's	
'twould	do	how'll	me	that	this	when's	why've	
'un	does	how's	might	that's	those	where	will	
've	er	how've	my	the	thou	where'd	with	
a	for	i	nor	thee	though	where'll	would	
am	fro	i'd	o'	their	thy	where's	you	
an	from	i'll	of	them	till	which	you'd	
and	had	i'm	on	then	to	while	you'll	
are	has	i've	or	there	us	who	you're	
as	have	if	shall	there'd	was	who'd	you've	

Table 2.4: Monosyllabic typically unstressed function words in English

2.2.4 Syllabification

Having transcribed the utterances on a segmental level, the next step was syllabification, or assigning syllable boundaries such that a word can be divided into syllables. Doubts have been raised about the precise nature of a syllable and also the value of such a concept, but there is overwhelming evidence that supports the psychological reality of syllables. Examples include speech production errors (Fromkin, 1973), language games (Botne and Davis, 2000), the fact that young children can identify the number of syllables before they can identify the number of phonemes (Liberman et al., 1974), word segmentation from continuous speech signals (Cutler and Norris, 1988; Cutler and Butterfield, 1992; Cutler, 1997) and others. Furthermore, phonological theories (Hulst and Ritter, 1999) heavily rely on the notion of syllables.

Syllabification in the corpus would therefore allow us to address questions that

are relevant to syllabic and sub-syllabic units (onsets, nuclei, and codas). In the domain of perception, syllabification would enable us to answer questions such as – Are stressed syllables less likely to be misperceived than unstressed syllables? Is word mis-segmentation – that is, when one word is mis-segmented as two or more words (e.g. *atom* is misperceived as *at home*) – more likely to occur at a syllable boundary, or within a syllable (that is, not before an onset or after a coda)?

Together it is clear that syllabification will potentially enrich the transcription of the corpus and thus allow us to address a broader range of research questions. Practically speaking, for the purpose of this thesis, we therefore need an automatic procedure to segment each word into syllables. There are two general approaches, rule-based and data-driven. In the sections below, I will first provide a brief overview of some of these methods and then argue for the most appropriate for this thesis, and also make the parameters/specifications of the method explicit.

2.2.4.1 Rule-based

2.2.4.1.1 The Sonority Principle The principle assigns a numeric value for every phone and the value is determined by a scale which is based on the amount of acoustic intensity (Selkirk, 1984). After assigning the sonority values (a high value indicates high sonority, while a low value indicates low sonority), each word will then have a *sonority profile*. Finally, the syllable boundaries lie at the starting point (i.e. on the left-hand side of the consonants) of the troughs of the sonority profile.

This method has drawbacks. Firstly, it is possible that the trough contains multiple phones that have the same value, i.e. a trough with a flat surface. In these cases, several possible splits are possible. Secondly, some argue that the notion of sonority does not contribute to the phonological knowledge that is used by listeners and talkers to attach linguistic meaning to the speech signal (Harris, 2006). Furthermore, the supposed correlate of sonority, acoustic intensity, poorly captures the

perceptual distance between phones (Harris, 2006), which in turn casts doubt on its role in syllabification.

2.2.4.1.2 The Legality Principle Hooper (1972) states that the possible onsets and codas are those that are phonotactically possible at word-initial and word-final positions. This principle suffers from the same drawback as the Sonority Principle, in that there could be multiple splits for intervocalic consonant sequences. Furthermore, a split might not always be possible, and in these cases, one has to either create illegal codas or illegal onsets, and the Principle of Irregular Codas (Pulgram, 1970) states that the illegality should lie in the codas.

2.2.4.1.3 The Maximal Onset Principle The Maximal Onset Principle gives preference to the longest possible onset that can be found at word-initial positions (Kahn, 1976). When splitting the intervocalic consonants, it will take the longest possible onset to be the onset and anything that is remains before the onset is a coda. It is essentially the same as the sum of the Legality Principle and the Principle of Irregular Coda. Its advantage is that it will always yield a possible syllabification.

2.2.4.1.4 A modified Maximal Onset Principle This modified principle is proposed by Gorman (2013) for syllabifying RP English. A number of modifications were made to the principle. To make the syllabification process more explicit, the author discussed the treatment of ambiguous segments /r/ and the onglides /j/ and /w/ when separating sequences of segments into vowels and consonants. The heuristic is such that an ambiguous segment is vocalic if it imposes restrictions on adjacent vowels, and it is consonantal if it imposes restrictions on adjacent consonants.

The treatment for /r/ is that pre-consonantal *r* is assigned to the preceding nucleus (Harris, 1994; Pierrehumbert, 2006). This is motivated by the fact that /r/ is the only consonant that allows glottalisation in /r/-ful British dialects (Harris,

1994, p. 258), e.g. “certain” /sɜɹtən/ as [sɜɹʔən], but not “mister” /mɪstəɹ/ as /mɪsʔəɹ/; and /r/ is the only consonant that disallows deletion of word-final /t,d/ in American dialects e.g. “third” /θɜɹd/ as [θɜɹ] is not allowed, but /mɛnd/ as /mɛn/ is allowed. These phonological behaviours suggest that pre-consonantal *r* patterns with other vowels if it was considered to be part of the nucleus.

The treatment for the onglide /j/ is that a) /j/ is part of the onset word-initially or preceded by one consonant, e.g. “yes” /jɛs/ and “junior” /ʤɥwn.jɪə/; b) /j/ is part of the nucleus if preceded by two or more consonants; therefore, it is restricting the adjacent vowels. The argument is that an onset /j/ can be followed by any vowels, but a nucleus /j/ can only be followed by /ɥw/. The treatment for /w/ is that it is always an onset based on its co-occurrence with other consonants, such that the onset that can precede /w/ is almost always /k/ (see Figure 2.11 for the token frequencies of [Cw] onsets); therefore, it is restricting the adjacent consonants.

Another modification concerns the syllabification of intervocalic consonant clusters. The difference with the maximal onset principle is that there is an additional constraint, which is when an intervocalic consonant cluster is preceded by a stressed lax vowel, e.g. in words such as “whisper” /wɪspəɹ/, the first consonant should be assigned as the coda, giving /wɪs.pəɹ/, while the maximal onset principle would in fact syllabify the word as /wɪ.spəɹ/.

2.2.4.1.5 Ambisyllabicity Ambisyllabicity can be used to form an additional step in the process of syllabification. This additional step would assign the onset of an unstressed syllable as the coda of the preceding syllable, such that this consonant is both an onset and a coda belonging to both syllables simultaneously (Kahn, 1976, p. 33). The advantage of including the notion of ambisyllabicity in syllabification is particularly clear when operating with the Sonority Principle (Wells, 1990). Consider cases where the intervocalic consonant, e.g. in “city”, is at the trough of the sonority profile and it can be assigned to the left or the right peak, and the decision of

assigning it to either of the peaks seems arbitrary; therefore assigning it as both the onset and the coda becomes an attractive option, although Harris (2013) argues extensively against the use of ambisyllabicity in English. The details will not be repeated here. Readers are encouraged to go to the source article for a complete set of arguments.

2.2.4.1.6 Wells (1990) Wells (1990) has proposed an English-specific syllabification system. Similar to the Maximal Onset Principle, it seeks to maximise the coda and not the onset. The codas are those that are phonotactically possible at word-final positions. Intervocalic consonant clusters are therefore split by maximising the codas, and the remaining consonants belong to the onset of the next syllable. Furthermore, this system is sensitive to morpheme boundaries, such that the morpheme boundaries are always split, e.g. "re-print" would have the syllabification, /ri.pɹɪnt/, coinciding with the morpheme boundary. It relies on seven phonetic patterns (such as pre-fortis clipping, tapping, stop epenthesis, elision of /t, d/ and more) which can be captured using allophonic rules that make use of syllable boundaries as part of their triggering environments.

Harris (1994, p. 225) stated that the "coda" analyses of these phonetic patterns by Wells (1990) are founded on two assumptions, in light of the arguments in Harris (1994, Ch. 2, 4) on the topics of constituency and licensing. The two assumptions are repeated verbatim below – 1) A word-final consonant occupies a coda and 2) A consonant occupying an onset in core syllabification can under certain circumstances be captured into a preceding coda.

2.2.4.1.7 Interim conclusion So far, I have summarised some of the rule-based syllabification methods, and their pros and cons in terms of practical and theoretical reasons. In the next section, I will summarise some data-driven methods of syllabification, before finally comparing all the approaches and selecting one that is most

appropriate for my purposes.

2.2.4.2 Data-driven

Data-driven approaches to syllabification have been argued to yield better syllabification than rule-based approaches (See Marchand, Adsett, and Damper (2009) for a summary of the comparisons between rule-based and data-driven approaches). I will describe two sub-approaches that utilise the data to infer syllabification below. The first approach will be referred to as *human-dependent*, and the second approach as *inductive* for the reasons outlined below.

2.2.4.2.1 Human-dependent approach This approach is one that seeks to learn syllabification from pre-syllabified corpora. Numerous methods have been proposed, e.g. Daelemans and Bosch (1992) on Dutch; Bartlett, Kondrak, and Cherry (2009) on English; and Goldwater and Johnson (2005) on English and German. The pre-syllabified corpora are syllabified manually; for instance, those from dictionaries such as CELEX (Baayen, Piepenbrock, and Gulikers, 1995). Concretely, these pre-syllabified corpora are used to a) obtain a set of syllabification rules, and b) derive the weights (probabilities) of these rules.

One of the drawbacks with this approach is that it relies on human-judgement of syllabification, which is known to be inconsistent (Goslin and Frauenfelder, 2001). This approach assumes that a given set of syllabified words has been "correctly" syllabified. While native speakers can identify the number of syllables with ease, they have great difficulties identifying the syllable boundaries. The boundaries determined by native speakers have a high level of variability (Goslin and Frauenfelder, 2001). This variability can also been found in when comparing the syllabifications in pronunciation dictionaries – in one comparison between CELEX and Merriam-Webster Online, Bartlett, Kondrak, and Cherry (2009) found that there is only 84% consistency.

2.2.4.2.2 Inductive approach This approach requires no pre-syllabified corpora, and relies on the statistical distributions of each segment with its adjacent segment(s); for instance, using transitional probability or mutual information (Gambell and Yang, 2005). For word segmentation using transitional probability, the general idea is to calculate the transitional probabilities of a syllable given the previous syllable. Since the transitional probabilities at word boundaries tend to be lower than the transitional probabilities between word-internal syllables, word boundaries can therefore be found/estimated. The process is then applied iteratively until the change in transitional probabilities stabilizes (Gambell and Yang, 2005). This method for word segmentation can be extended to syllabification by using segments as the unit, instead of syllables. The disadvantage with this approach is that it is unclear how much data is needed to minimise the effect of data-sparsity. That is, the co-occurrences of some segments might not be found in the data or might have very low frequency, therefore yielding unreliable estimates of their probabilities.

2.2.4.3 Comparison

It is not easy to compare different syllabification methods, since there is no *gold standard*, i.e. a set of syllabified words that are assumed to be correct. As previously mentioned, human-judgements of syllabifications are unreliable and so are the ones in pronunciation dictionaries as they are done by humans. Furthermore, even if human judgements of syllabifications are consistent, they might not reflect psychological reality.

Given that an objective measure is not readily available, we shall consider the practicality and the drawbacks of each method before deciding which is more appropriate for this thesis.

Neither of the rule-based methods, the Sonority Principle and the Legality Principle, are always able to provide consistent syllabification, so there will not be con-

sidered further. The addition of ambisyllabicity to a given syllabification method, especially with the Sonority Principle, could resolve many cases of ambiguous syllabifications, but its application to English syllabification is highly questionable from a theoretical perspective (Harris, 2013), so this method is also ruled out. Wells's (1990) syllabification system is English-specific, and it is based on a particular method of analysing allophonic patterns in English, but given that alternative analyses can be made without relying on a specific method of syllabification (Harris, 1994, p. 225), this method is ruled out for being overly analysis-dependent as well as language-specific. The modified Maximal Onset Principle (Gorman, 2013) is theoretically grounded, which could potentially bias any results derived from the data to favour particular theories or mirror results obtained from similar data – one of the arguments for the front onglide /j/ being assigned to the nucleus after a consonantal onset is from the behaviour of [ju] in speech error (speech misproduction) data, and the data in this thesis are also error-based data (speech misperception) – which could therefore bias the results from speech misperception data towards being more similar to those from speech misproduction data. As a result, this modified Maximal Onset Principle is therefore ruled out. We are therefore left with the "original" Maximal Onset principle amongst the rule-based methods.

Moving to the data-driven methods, the human-dependent approach is ruled out on the basis of the need to rely on pre-syllabified data, which themselves might be unreliable. This leaves us with the inductive method. On the one hand, the inductive method is more attractive than the Maximal Onset Principle as it is entirely unsupervised – no knowledge of onset, nucleus, and coda is needed, while the Maximal Onset Principle requires one to specify the nucleus of each word in order to extract all the possible word-initial onsets, and what constitutes the nucleus is also debatable, as we have seen in the treatments of /r/ and onglides in English by Gorman (2013).

On the other hand, the inductive approach is not appropriate for this thesis

because the corpus is transcribed dialectally, and since some dialects have more data than others, the phones are not evenly distributed. This means that for the dialects that have little data, the inductive method would face the issue of data-sparsity, and would potentially yield less consistent syllabifications than those dialects that have more data. For this reason, the inductive approach is rejected, and the Maximal Onset Principle is adopted for the syllabification of the phonetic transcriptions in this thesis.

2.2.4.4 The specifications of Maximal Onset Principle

Having established which syllabification method to use, we must then provide the Maximal Onset Principle with three lists of segments – the possible consonants, nuclei and onsets. The consonant list is the same for all dialects. The onset list is also the same for most but not all dialects, because more conservative dialects, such as Southern British English, contain specific onsets that are being phased out in less conservative dialects. The nucleus list, however, requires a different list for each dialect, because different dialects have different vowel sets (see Section 2.2.7 for the different vowel sets). In the following paragraphs, I will specify these three lists, paying special attention to the possible onset list.

2.2.4.4.1 Possible consonants The list of possible consonants is as follows: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, j, w].

2.2.4.4.2 Possible nuclei The list of possible nuclei consists of vowels, which are those listed in vowel set tables in Section 2.2.7, with the exception of the rhotic vowels. Since the rhoticity of these rhotic vowels is denoted as a separate phone following the vowel (see Section 2.2.7.2.2 for the arguments supporting this analysis), it will not be included as part of the nucleus. Furthermore, this treatment would more accurately encode the re-syllabification of the rhotic consonant as part of the

onset when followed by a nucleus.

2.2.4.4.3 Possible onsets To determine what constitutes a possible onset, I have devised a principled method using frequency distributions of any potential onsets and this method is outlined below.

The most straightforward way to determine the set of possible onsets is to simply accept all word-initial consonant sequences as possible onsets. However, some word-initial consonant sequences are deemed as marginal cases; for instance, according to Harris (1994, p. 57), [sf] and [vr] are of marginal status, presumably because they are restricted to relatively fewer lexical items, e.g. the derivatives of *sphere* and *vroom*. Furthermore, Harris (1994, p. 58) argued that certain sequences are inadmissible, e.g. [pw], [bw], [tl], [dl] and [θl], and this is because of an intra-onset phonotactic constraint which states that homorganic clusters with [l] and [w] are not allowed. If we were to accept marginal cases such as [vr] and [sf] as possible onsets then the Maximal Onset Principle would syllabify them as onsets word-medially, e.g. *several* as [sɛ.vɹəl], and *Oxford* as [ɔk.sfəd]. If we were to accept these marginal consonant sequences to be valid onsets, then we would have to accept that these onsets are productive only word medially, and there are some constraints stopping them from occurring word-initially, which seems to be an illogical account of the pattern. Furthermore, under a parsimonious account of onset storage, these consonant sequences are best treated as invalid onsets, since word-medially they could be split as part of a coda and part of another valid onset; and only word initially, they are allowed to be onsets for restricted number of items. Following Harris (1994, p. 57), I conclude that it is unreasonable to blindly accept all word-initial consonant sequences as possible onsets, and therefore this option is ruled out.

While it is possible to adopt the analyses by Harris (1994, pp. 57–58) for determining possible onsets, some of the conclusions are less clear cut, specifically whether a particular consonant sequence is *marginal* or *inadmissible*. [θw] for example is said

to be a clearly admissible case (as much as uncontroversial cases like [pl] and [kl]), while [sf] is said to marginal (Harris, 1994, p. 57). However, [θw] is in fact restricted to a few lexical items such as *thwack*, *thwart* and *Thwaite*. One could therefore argue that [sf] and [θw] should share the same status – either both marginal or both admissible. Let us examine another two supposedly admissible sequences, [dw] and [gw]. [dw] can only be found in 26 words in the *Current British English* pronunciation dictionary (henceforth CUBE) (Lindsey and Szigetvári, 2014) and they can be reduced down to six lemma, *dwarf*, *dwell*, *Dwight*, *dwindle*, *Dworkin*, and *Dwyer*. Out of these six lemma, three are proper names, leaving three content words. [gw] can only be found in 44 words in CUBE. By inspecting these words, it was found that are mostly loanwords and proper names, from Spanish (e.g. *guano*, *guava* and *Guatemala*), and from Welsh (e.g. *Gwen* and *Gwersyllt*). Furthermore, the phonotactics of the English language could be affected by new words in the language, e.g. recently coined words such as *vlog* ("Video blog") have introduced the [vl] onset. Having considered these ambiguious cases in Harris (1994, pp. 57–58), it is clear that we need a more principled method.

My principled method considers two factors in the selection of the possible onsets. The first factor is the number of times a given consonant sequence occurs between the word-initial boundary and the first nucleus of a word. This factor can be seen as the amount of evidence that supports the onset status of a consonant sequence, such that the more frequently a consonant sequence is found word-initially, the more likely this sequence is a possible onset. The second factor is the number of times a given consonant sequence *could* form an onset word-medially. This factor can be seen as the amount of impact on the syllabification if a consonant sequence were a possible onset.

In order to calculate these two factors, we need an IPA-transcribed lexicon. A Southern British English lexicon, CUBE (Lindsey and Szigetvári, 2014), was chosen

over an American English lexicon because Southern British English is a conservative dialect (Harris, 1994, p. 61); e.g. it preserves specific onsets such as [tj] and [dj], which are often absent in General American English. This lexicon would allow us to establish a set of possible onsets for British English and for American English after removing specific onsets. The reason for choosing CUBE over another British lexicon such as CELEX (Baayen, Piepenbrock, and Gulikers, 1995) is that CUBE is being updated regularly, while CELEX has not been updated since 1995.

The lexicon underwent two simple pre-processing steps. First, all multi-word entries (e.g. *Christmas Eve*) were removed. Second, the remaining entries in CUBE were enriched with token frequency information from SUBTLEX-UK (van Heuven et al., 2014), a corpus compiled using British Broadcasting Corporation subtitles. The words with zero frequency were excluded. The final lexicon contains 70,181 word forms, with IPA transcriptions as well as token frequencies. This final lexicon was then used to compute the two factors. All potential onsets were identified by extracting all word-initial consonant sequences. In total, 100 types of consonant sequences were found. We then calculated the token frequencies of these potential onsets separately at word-initial positions and at word-medial positions.

Figure 2.11 plots the word-initial token frequency by the word-medial token frequency, both of them logarithmically (base 10) transformed. The plot shows that the 100 potential onsets cluster into two groups. One group is located at the top right portion of the plot starting from the value 4 (log10 initial/medial frequency) on both axes. The second group is located on the left hand side of the plot, in the range of 0 to 4 (log10 initial token). There is a clear linear relationship between the two frequency measures for the top right group, while this relationship is absent for the other group.

One explanation for the observed patterns is that the frequency of a valid onset word-initially and word-medially should be similar because, on average, syllables

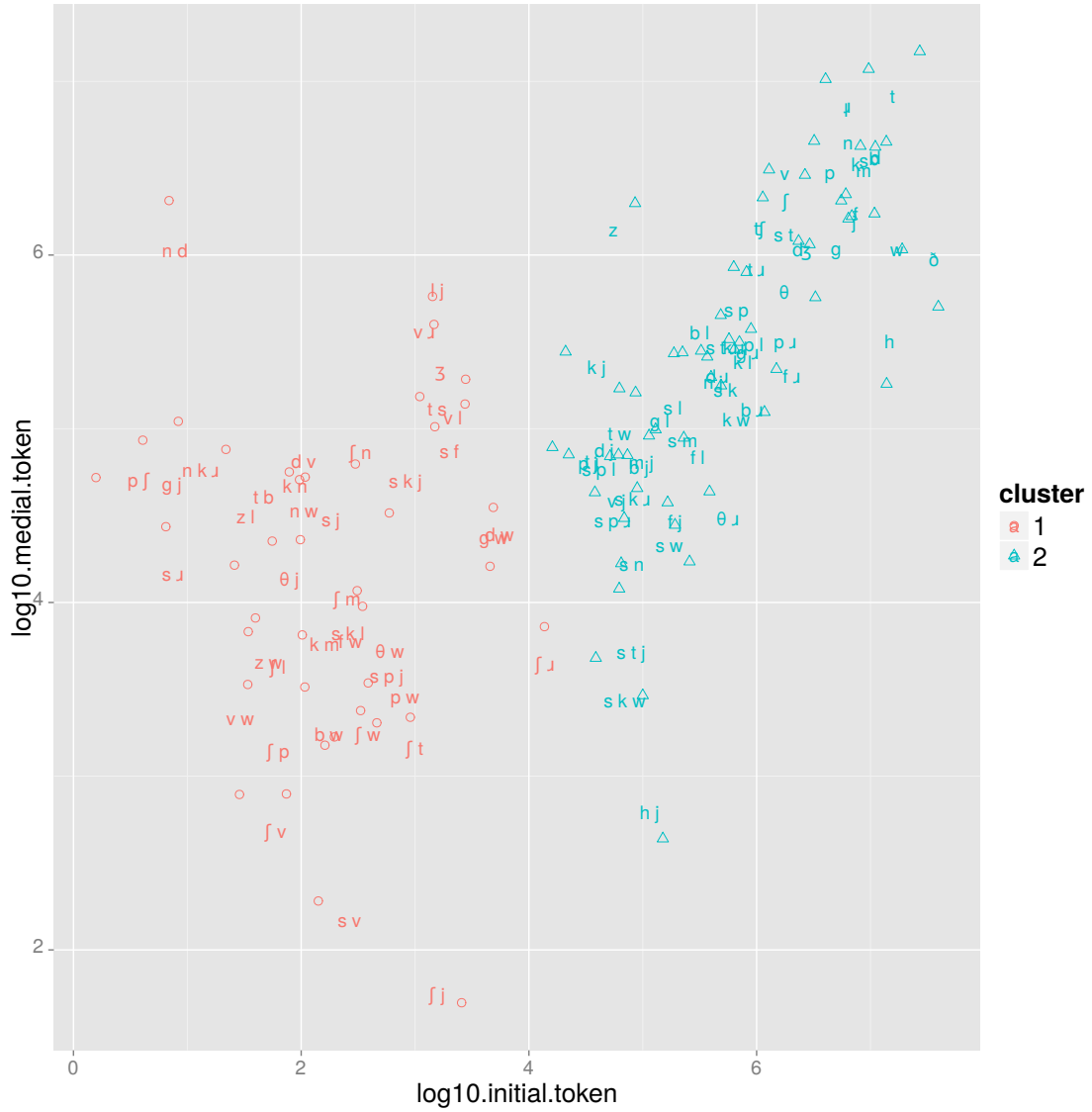


Figure 2.11: Frequencies of initial onsets vs. medial onsets with k-means clustering

tend to re-occur to form polysyllabic words and the position of re-occurrence should be relatively unrestricted. While a bogus onset need not follow this trend, consider the observations that [pʃ] is highly infrequent word-initially and only exists in one word, *pshaw*, but this sequence occurs much more often word-medially, e.g. in *option*. This is because word-medially the [p] can be a coda, and the [ʃ] can be an onset, and codas and onsets are relatively unrestricted in terms of their co-occurrences compared to onset clusters.

Looking more closely at the onsets in each group, it is immediately clear that the top right group consists of uncontroversial onsets, such as [fl], [t] and many others, while the group on the left consists of ambiguous/marginal onsets, such as [ʃm] (e.g. *schmuck*) and [pw] (e.g. *Poirot*). This suggests that we could use the clusters formed by the two frequency measures to help us select the set of valid onsets for the Maximal Onset Principle. To identify the clusters beyond visualisation, I employed the k-mean clustering algorithm (MacQueen, 1967). Clustering techniques seek to assign a set of objects into clusters by the similarity of the objects, which, in our case, is based on the two frequency measures mentioned above.

The first step when clustering is to determine the number of clusters (also commonly referred to as k). Although two clusters were visually identified, it is possible that the 100 potential onsets could be better explained with more clusters. The *NbClust* package (Charrad et al., 2014) was applied to determine k by 26 established indices in the clustering literature the optimal number of clusters. The chosen number of clusters was the one selected by the majority of the indices. In addition to NbClust, the chosen number of clusters was checked against a within-cluster sum of squares (WCSS) curve, to see if it lies on the "knee" of the curve. The "knee" of the WCSS curve is a common visualisation method for identifying the optimal number of clusters.

According to the majority rule by NbClust, the best number of clusters is two, voted by ten out of 26 indices. The number of clusters that came second was three, voted by only four out of 26 indices; therefore, the two-clusters solution is a clear winner. This is confirmed by the WCSS curve, see Figure 2.12, where the knee of the curves lies in the range of two to six clusters.

The function *kmeans* from the R (R Core Team, 2013) was used to perform the k-means clustering. The number of clusters (k) was set to two, with 1,000 random starting points. The two clusters are visualised in Figure 2.11 in two different colours

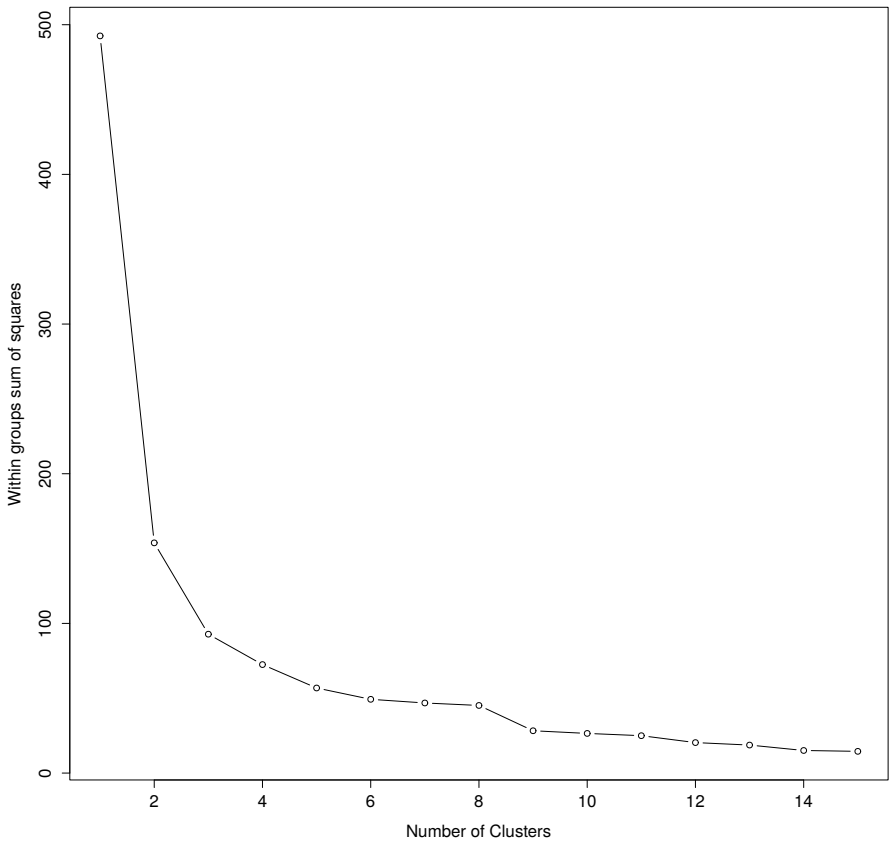


Figure 2.12: The within-cluster sum of squares (WCSS) curve for clustering potential onsets

and pointers. Cluster 1 is in red with the circle pointers, which contain the invalid onsets. Cluster 2 is in blue with the triangular pointers, which contain the valid onsets.

Cluster 1 consists of the following 40 invalid onsets – [ʃɹ, bw, dv, dw, fw, gj, gw, ʃj, km, kn, ʃl, lj, ʃm, ʃn, nd, nkɹ, nw, ʃp, pʃ, pw, sɹ, sf, sj, skj, skl, spj, sv, ʃt, tb, ts, ʃv, vɹ, vl, vw, ʃw, zl, zw, ʒ, θj, θw].

Cluster 2 consists of the following 60 valid onsets – [ɹ, dʒ, ʃ, tʃ, b, bɹ, bj, bl, d, dɹ, ð, dj, f, fɹ, fj, fl, g, gɹ, gl, h, hj, j, k, kɹ, kj, kl, kw, l, m, mj, n, nj, p, pɹ, pj, pl, s, sk, skɹ, skw, sl, sm, sn, sp, spɹ, spl, st, stɹ, stj, sw, t, tɹ, tj, tw, v, vj, w, z, θ, θɹ].

Finally, to evaluate the strength of clusters, I transformed the two frequency measures into two principal components, and plotted the outlines of the two clusters

found by the k-means algorithm, see Figure 2.13. This visualisation was done using the *clusplot* function from the *cluster* package (Maechler et al., 2013). The plot shows that the two outlines barely overlapped, which indicates that our clusters are stable and means the classification of the possible onsets to each cluster is not likely to change.

Looking more closely at Cluster 1, we see that it contains one consonant, [ʒ]. This clustering solution suggests that [ʒ] is an invalid onset, which is primarily due to its low frequency of occurrence word-initially. In fact, [ʒ] was a relatively new phoneme for English. It did not exist in Old English and was introduced through the process of palatalization, primarily from /zj/ and reinforced by French words that are familiar to English speakers (Fromkin, Rodman, and Hyams, 2003, p. 492). Word-initially, [ʒ] is more restricted and is more prone to variation with the affricate [dʒ], which contributes to its low frequency of occurrence. The implication of excluding [ʒ] as a valid onset is that instances of an empty-headed onset would be introduced, e.g. in *measure* [mɛʒ.əɹ]. Following the principle of avoiding empty-headed onsets,[ʒ] is classified as a valid onset, despite the clustering analyses.

In sum, I developed a principled method for selecting a set of possible onsets for the Maximal Onset Principle of syllabification. The 60 valid onsets in cluster 2 will be used for the syllabification in this thesis. For all but the British English accents, any coronal + [j] potential onsets were excluded from the set of possible onsets (Harris, 1994, p. 61), meaning that the following four onsets were removed – [tj, dj, nj, stj]. The method described above has plenty of room for further development; for instance, it could take into account of the prosodic shape of the words. Nonetheless, it is serviceable for the purpose of finding a set of valid onsets. One further remark is needed about the invalid onsets at word-initial positions. The invalid onsets are at word-initial positions, so they do not impact on the syllabification process. However, in terms of assigning each phone into either an onset, a nucleus or a coda, they do

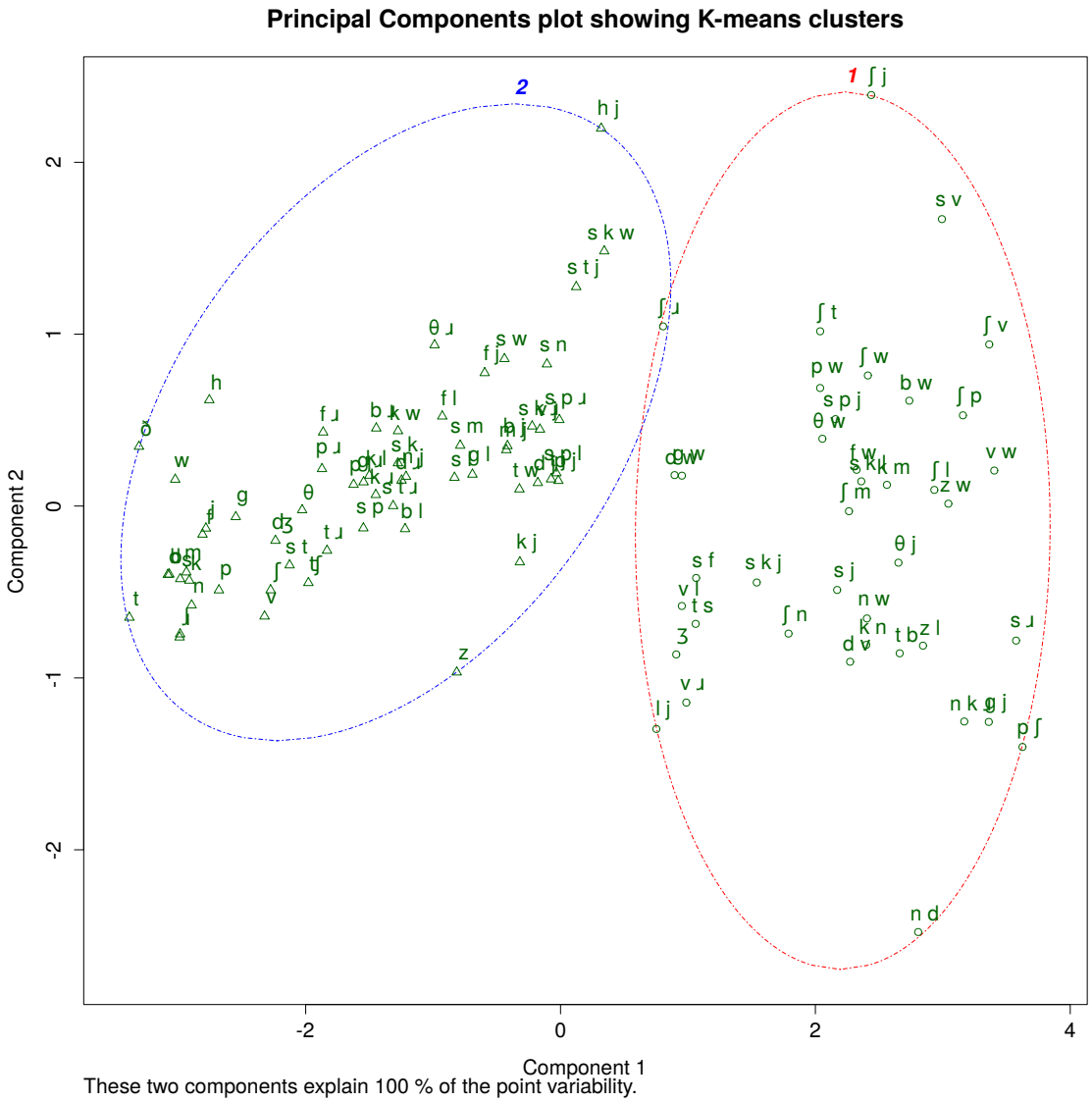


Figure 2.13: Principle components plot showing K-means clusters

matter. The simple solution I employed was to only assign these invalid onsets as being onsets in word-initial positions.

2.2.5 Dialectal transcription

In this section, I will discuss how to best represent the interlocutors' accents in the transcriptions. While we have dialectal information about the utterers and perceivers, both the intended and perceived utterances were transcribed with the dialectal group

of the utterers. In the following paragraphs, I will outline the rationale behind this decision. Let us consider the four possible options when deciding which accent to apply to the intended and perceived utterances.

1. Transcribe both intended and perceived utterances with the utterer's accent.
2. Transcribe both intended and perceived utterances with the perceiver's accent.
3. Transcribe the intended utterance with the utterer's accent and the perceived utterance with the perceiver's accent.
4. Transcribe the intended utterance with the perceiver's accent and the perceived utterance with the utterer's accent.

The first option is to assume that the perceiver perceives both the intended and perceived utterances with the utterer's accent. The rationale behind this option is that the traditional phonetic transcriptions represent the speech as produced by the utterers. Furthermore, in the perspective of the perceivers, when a perceiver perceives an utterance, he/she perceives it in the utterer's accent (be it correctly or incorrectly perceived), rather than his/her own, hence our ability to detect an accent that is differ from our own, and the same logic therefore applies to a misperceived version of the utterance.

The second option is to assume that the perceiver perceives both the intended and perceived utterances with his/her own accent. This is supported by the fact that adults are able to ignore indexical variations in speech for lexical retrieval, as demonstrated by word recognition models that rely on underlying representations (McClelland and Elman, 1986; Norris, 1994), or exemplars (McLennan and Luce, 2005). The use of the perceiver's accent for the transcription is therefore to represent the speech that has normalised to the perceiver's accent.

The third option is to transcribe the intended utterance with the utterer's accent and the perceived utterance with the perceiver's accent. This option takes into account the accents of the perceiver and the utterer, but if the intended speech were to be transcribed in the utterer's accent and the perceived speech in the perceiver's accent, then the difference between the two transcribed utterance would not only include perceptual errors (made by the perceiver) but also accent differences, mainly due to different vowel qualities. These differences in vowel qualities are not necessarily errors made by the perceiver, as the perceiver could in fact retrieve the correct lexical item despite differences in vowel qualities between the two accents. In other words, this option will overgenerate errors.

Finally, the fourth option is to use the perceiver's accent to transcribe the intended utterance and the utterer's accent to transcribe the perceived utterance. This option is entirely illogical; like the third option, it will overgenerate errors.

In sum, the first and second options seem to be most reasonably justified. The main difference between the two options is that the first option focuses on the incoming speech signal, and thus is similar to a surface/narrow representation, while the second option focuses on the processed speech signal, and thus is similar to a phonemic representation. While both are equally compelling, the first option is less susceptible to particular phonemic analyses under specific phonological frameworks, and for this reason, I will opt for the first option in this thesis – to transcribe both the intended and perceived utterances with the utterer's accent. Crucially, this decision does not necessarily imply that the perceiver's accent is entirely ignored; the perceiver's accent can be taken into account in subsequent analyses such as including them as a random effect in a mixed-effects model (Bates et al., 2014). In future work, one could devise a method of incorporating the perceivers' accents in the perceived transcription, while avoiding overgeneration of errors by comparing the orthography in the intended and perceived utterances, such that the mismatches in the phonetic

transcriptions are ignored if they belong to identical orthographic forms.[5] In this sense, they are correctly perceived in terms of lexical retrieval.

2.2.6 Dialect classification

Two of the subcorpora, Nevins and Labov, have extensive demographic information about where the interlocutors are from. I used this information to refine the transcription for different English accents. Wells's (1982) *Accents of English* was used extensively to transcribe the different English accents. Since a majority of data is related to North American English accents, it would be useful to also consult Labov, Ash, and Boberg's (2005) Atlas of North American English. However, Labov, Ash, and Boberg's (2005) use of IPA notation is limited and relies more on formant values and descriptions; for this reason, it was not used as the primary source of my dialectal transcription of North American English accents.

2.2.6.1 American English varieties

The demographic information was predominantly for American English varieties. The reporters provided the specific states, cities or regions, and I classified them into major regional speech areas with the dialectal map by Thomas (1958), reprinted in Wells (1982d, p. 472, Fig. 17). These major regional speech areas are Eastern New England, New York City, Middle Atlantic, Southern, Western Pennsylvania, Southern Mountain, Central Midland, Northwest, Southwest and North-Central and they are summarised in the first two columns of Table 2.5.

Guided by these major regional speech areas, each of the states, cities or regions were categorized into broad dialectal groups; see the last column of Table 2.5. These broad dialectal groups are Philadelphia, New York City, Southern, New England and General American (GenAm). The broadness of these varieties was motivated by the

[5] I thank Jamie White for suggesting this possibility.

scope of documentation by Wells (1982a). Since the lexical set for Chicago accents was not readily available, it was assumed to be GenAm, and this assumption is justified by the small portion (1%) of the data with speakers from Chicago; therefore, 1% is unlikely to have a major impact on the overall analyses. It is worth noting that the mapping between the major regional speech areas and the broad dialectal groups was not one to one, e.g. the major regional speech area for District of Columbia is Middle Atlantic and it is classified as the broad dialectal group, GenAm, while the major regional speech area for Delaware is also Middle Atlantic, but is classified as Philadelphia, not GenAm.

To ensure accuracy, both levels of classifications of the American English varieties were done in consultation with a native American English linguist[6].

2.2.6.2 Other varieties

The coverage of accents in this chapter was motivated by the scope of documentation by Wells (1982a). Other dialectal varieties in the corpora include Received Pronunciation/Southern Standard British English, Ireland, Scotland, South Africa, Canada, India, New Zealand, Caribbean and Australia. Apart from the accents mentioned above, there are another $\approx$ 40 accent groups covering 4% of the corpus ($\approx$ 200 data points) and are as follows – Africa[7], China, Singapore, Colombia, Israel, Vietnam, Czech Republic, Kenya, Turkey, Russia, France, Germany, Mexico, Korea, Italy, Guatemala, Brazil, Nigeria, Taiwan, Spain, Senegal, Norway, Japan, Zimbabwe, Somalia, Saudi Arabia, Romania, Poland, Peru, Madagascar, Lebanon, Greece, Finland, Ethiopia, Cameroon, Cambodia and Bahrain. The vowel sets of these 40 accents were not taken into account in the transcriptions of the corpus, and were transcribed as GenAm.

[6]I thank Prof. Andrew Nevins for his expert comments.

[7]Although African English is covered by Wells (1982a), only the Yoruba vowel set was documented. It is not appropriate to generalise the Yoruba variety to all African English varieties, and therefore the Africa accent group was not taken into account.

Regions	Abbrev.	Speech areas (Thomas, 1958)	Broad Dialectal Groups
Alabama	AL	Southern	Southern
Alaska	AK	Northwest	GenAm
Arizona	AZ	Southwest	GenAm
Arkansas	AR	Southern	Southern
California	CA	Southwest	GenAm
Colorado	CO	Central Midland	GenAm
Connecticut	CT	Eastern New England	New England
Delaware	DE	Middle Atlantic	Philadelphia
Florida	FL	Southern	Southern
Georgia	GA	Southern	Southern
Hawaii	HI	Southwest	GenAm
Idaho	ID	Northwest	GenAm
Illinois	IL	North-central	GenAm
Indiana	IN	Central Midland	GenAm
Iowa	IA	North-central	GenAm
Kansas	KS	Central Midland	GenAm
Kentucky	KY	Southern Mountain	Southern
Louisiana	LA	Southern	Southern
Maine	ME	Eastern New England	New England
Maryland	MD	Middle Atlantic	GenAm
Massachusetts	MA	Eastern New England	New England
Michigan	MI	North-central	GenAm
Minnesota	MN	North-central	GenAm
Mississippi	MS	Southern	Southern
Missouri	MO	Central Midland	GenAm
Montana	MT	Northwest	GenAm
Nebraska	NE	Central Midland	GenAm
Nevada	NV	Southwest	GenAm
New Hampshire	NH	Eastern New England	New England
New Jersey	NJ	Middle Atlantic	Philadelphia
New Mexico	NM	Central Midland	GenAm
New York[a]	NY	North-central	GenAm
North Carolina	NC	Southern	Southern
North Dakota	ND	North-central	GenAm
Ohio	OH	North-central	GenAm
Oklahoma	OK	Central Midland	GenAm
Oregon	OR	Northwest	GenAm
Pennsylvania[b]	PA	Western Pennsylvania	GenAm
Rhode Island	RI	Eastern New England	New England
South Carolina	SC	Southern	Southern
South Dakota	SD	North-central	GenAm
Tennessee	TN	Southern Mountain	Southern
Texas	TX	Southern	Southern
Utah	UT	Central Midland	GenAm
Vermont	VT	Eastern New England	New England
Virginia	VA	Southern	Southern
Washington[c]	WA	Northwest	GenAm
West Virginia	WV	Southern Mountain	Southern
Wisconsin	WI	North-central	GenAm
Wyoming	WY	Central Midland	GenAm
Subregions			
Puerto Rico	PR	–	GenAm
District of Columbia	DC	Middle Atlantic	GenAm
New York City	–	New York City	New York City
Long Island	–	New York City	New York City
Upstate NY	–	North-central	GenAm
Philadelphia	–	Middle Atlantic	Philadelphia

Table 2.5: Classification of major regional speech areas

[a]Only if the reporter stated New York with a specific city/region that is not New York City, otherwise it is assumed to be New York City.

[b]Only if the reporter did not state Philadelphia, otherwise it is assumed to be Philadelphia.

[c]Only if the reporter did not state that it is Washington DC, otherwise it is assumed to be District of Columbia.

2.2.7 Dialectal vowel sets

Wells (1982a) was used extensively to determine the surface realisation of the vowel set for each accent. Firstly the phonemic vowel sets were tabulated; secondly the detailed discussions of the surface realisation were utilised and tabulated; thirdly if any vowels not mentioned in the discussions of the surface realisation, they were extrapolated using the available surface realisations and phonemic representations; finally, the extrapolated vowel sets were simplified and normalised across all of the accents.

Vowel sets of multiple accents would invariably contain a wide range of phonetic realisations and therefore a wide range of phonetic symbols, which were dependent on the narrowness of the reported descriptions. When comparing transcriptions of different accented speech, it is preferable, if not essential, to reduce the number of different segments. This is because confusion matrices, which are used in multiple analytical techniques in this thesis, are negatively affected by sparsity – the zero frequency problem. Since confusion matrices contain frequencies of co-occurrences between any two segments that occurred even once, the segments with low frequencies of occurrence would therefore have many zero co-occurrences with other segments. Rather than removing segments that have low frequencies in the confusion matrices, they could be simplified, so that they are denoted by common segments. More specifically, this can be done by ignoring diacritics (Wieling et al., 2014, p. 261) or choosing phonetically similar segments as substitutes.

These steps are exemplified in detail in Section 2.2.7.1 and Section 2.2.7.2, where the General British and General American accents were tabulated respectively and subsequent accents were tabulated in the same way.

2.2.7.1 General British

The General British (henceforth GenBr) vowel set is described in this section and the complete table with different vowel sets is summarised in Table 2.6.

The first column "Keyword" is based on the standard lexical sets for English by Wells (1982b, pp. 118–124). Lexical sets are a group of words that have the same pronunciation for a certain sound for a particular accent. Wells's (1982) lexical sets were based on two 'reference' accents of English, namely *Received Pronounciation* (henceforth RP) and *General American* (henceforth GenAm). A few modifications were made to better capture the accent variations, specifically, the following pre-vocalic counterparts were added: NEAR, SQUARE, START, NORTH, FORCE, CURE, NURSE, COMMA, KIT and DRESS, some of which were described also in Wells (1982b, p. 124).

The keywords were divided into two major groups and each major group was sub-divided into two minor groups. The two major groups were the non-pre-vocalic group (from KIT to LETTER) and the pre-vocalic group (from MIRROR to LETTERING), e.g. the [ɪ] in KIT is non-pre-vocalic, while the [ɪɹ] in MIRROR is pre-vocalic. The two subgroups were vowels that do not contain a historical /r/ (from KIT to COMMA in the non-pre-vocalic group; MIRROR and MERRY in the pre-vocalic group) and those that contain one (from NURSE to LETTER in the non-pre-vocalic group; CURRENT to LETTERING in the pre-vocalic group).

To determine an appropriate vowel set for GenBr, I first examined two existing vowel sets, namely Wells (1982b, pp. 118–124), which is phonemic, and the Longman Pronunciation Dictionary by Wells (2008), which has a broad surface representation. These two sets were tabulated in the second and the third columns respectively. I then completed the vowel set by extrapolation, which was tabulated in the fourth column. Specific to the GenBr accent, I included another vowel set by Lindsey (2012b), which is a narrow surface representation and more contemporary, and this

set is tabulated in the fifth column. I identified and discussed the modifications I made to existing vowel sets below. The final vowel set for GenBr used in this thesis is listed in the final column of Table 2.6.

The final vowel set is in favour of a more contemporary SSBE accent, moving away from RP. While the use of a General British accent might overgeneralise the various British accents, for practical purposes its use is needed, similar to the General American accent. This is partially justified by the fact that dialect levelling is a widespread phenomenon in Britain and diversification is hard to find (Kerswill, 2003). Furthermore, it is well-documented that RP is fading with so-called Estuary English becoming more dominant (Kerswill, 2001; Przedlacka, 2001; Kerswill, 2006).

2.2.7.1.1 DRESS and HAPPY Wells (1982b, pp. 118–124) and Wells (2008) both aimed to describe a vowel set for RP, and the former contained more fine-grained details for the pre-vocalic counterparts. Amongst the lexical sets that they both shared, only the HAPPY vowel is different. Wells (1982b, pp. 118–124) was used as the base to make certain modifications, starting with the DRESS and HAPPY vowels which are discussed below.

The DRESS vowel was denoted with a tense vowel [e] in Wells (2008) and Wells (1982b, pp. 118–124). In the case of Wells (2008), the choice of the tense vowel was chosen on the basis of parsimony of symbols. The choice of the symbol for DRESS is closely linked to that of FACE, but since across dialects the FACE vowel is typically realised as a long monophthong [eː] or a diphthong, there will be no confusion even if the symbol for DRESS were to be the same as that of FACE without the length mark or the offglide (Wells, 2009).

In the case of Wells (1982b, pp. 118–124), this vowel was chosen on the basis of phonology. Phonetically, it is described by Wells (1982b, p. 128) as "a relatively short, lax, front mid vocoid". Given that the purpose of the vowel sets in this thesis is to transcribe (and therefore differentiate) differently accented speech, a phonetically

accurate symbol is preferred, namely the lax vowel [ɛ].

Regarding the HAPPY vowel, it was denoted with a short tense vowel [i] in Wells (2008) instead of the lax vowel [ɪ] in Wells (1982b, pp. 118–124). This difference is due to an on-going change in Southern British English and RP – with older speakers using the lax variant and with younger speakers using the tense variant (Fabricius, 2002). To capture a more contemporary GenBr use, the tense variant is preferred.

2.2.7.1.2 Offglides and GOOSE Lindsey (2012b) described a vowel set for *Standard Southern British* which differs significantly from RP and indeed from Wells (1982b, pp. 118–124); for instance, the vowel quality for GOOSE (from [u] to [ʉ]). Only four lexical vowels were identical between Lindsey (2012b) and Wells (1982b, pp. 118–124), namely the KIT, DRESS, COMMA and START. The modifications are based on the recordings of modern southern English speakers as well as those of the Royal family members who are traditionally regarded as the benchmark for RP (Wales, 1994).

While this vowel set is considerably more accurate, it lacks consistency with the vowel sets of other accents described by Wells (1982a); therefore, it would not be appropriate to blindly adopt the set by Lindsey (2012b). However, some aspects of Lindsey's (2012) vowel set could be adopted, including the treatment of offglides in diphthongs and long vowels as well as GOOSE-fronting.

The key modifications by Lindsey (2012b) on non-short-lax vowels are threefold. Firstly, the centring diphthongs (NEAR, SQUARE and CURE) are proposed to be long monophthongs. While this modification was well-motivated by an on-going change, this change is nonetheless incomplete, as the more conservative variants remain prevalent in less prestigious accents of Southern Britain (Lindsey, 2012a). Therefore this monophthongization was not adopted for this thesis, with the second vowel-unit left as [ə].

Secondly, two of the long vowels described in Wells (1982b, pp. 118–124), namely

FLEECE and GOOSE, are denoted to have the offglides [j] and [w] respectively instead of the length mark. Similarly, the diphthongs ending with a lax vowel [ɪ] and [ʊ] are denoted to have the offglides [j] and [w], with FACE, PRICE and CHOICE ending with [j] and GOAT and MOUTH ending with [w].

Finally, while most of the vowels differ in terms of quality from those in Wells (1982b, pp. 118–124), the modification of the vowel quality for GOOSE is arguably most needed (from [u] to [ʉ]), due to a change called GOOSE-fronting. This change in RP was first reported by Henton (1983) (Wells, 2010). In surveys by Przedlacka (2001), it was found that while sociolinguistic factors do influence the amount of fronting, the vowel is nonetheless fronted for both RP and Estuary English. Since the change is on the whole complete, the modification is therefore adopted for this thesis.

2.2.7.1.3 Length marks Since the transcription will be used in the alignment process for identifying errors, a natural question is how we should treat the length marks in the alignment process. This is ultimately related to the question of what the minimal alignment unit is and whether it be an IPA segment or a phoneme. This question will be examined in more detail in Section 2.4.3, but it is clear that the IPA symbol [ː] is a not a psychologically plausible unit of perception. The length mark denotes the lengthening of the preceding sound, and for the purpose of this thesis I will replace the length mark with the preceding symbol. This will be consistently applied to all the vowel sets of other accents in subsequent sections.

2.2.7.1.4 Extrapolation Finally, some vowels were absent across all of the vowel sets that were examined, especially CURRENT to LETTERING in the pre-vocalic group. These will be completed using the patterns observed in the non-pre-vocalic group. In GenBr, we simply assumed that [ɹ] is appended to their non-pre-vocalic counterparts, NURSE to LETTER. If the surface realisation of a lexical vowel was not documented

but its phonemic form was, then first the surface realisation of another lexical vowel with the same phonemic form was adopted. For instance, the surface realisation of PALM was not documented, but that of BATH, which has the same phonemic form as PALM, was; if this method of extrapolation was not available, then its phonemic form would be used as the surface realisation. Similar extrapolation processes were used for describing other accents in subsequent sections.

2.2.7.2 General American

Following the same format as Section 2.2.7.1 for describing the General British vowel set, I first examined two existing vowel sets in order to determine an appropriate vowel set for General American (henceforth GenAm), namely the Longman Pronunciation Dictionary by Wells (2008) and Wells (1982b, pp. 120–124). The final vowel set for GenAm for this thesis is listed in the final column in Table 2.7.

2.2.7.2.1 Length contrast and GOAT Wells (1982b, p. 120) suggested that vowel length is not as important in GenAm than it is in other accents, since the duration of all the vowels can vary depending on their phonetic environments. For this reason, the phonemic representation of the GenAm vowel set does not contain any length marks, e.g.vowels such as FLEECE and GOOSE are denoted as /i/ and /u/.

Wells (2008) on the other hand denoted length marks in its vowel set, highlighting the surface forms. For the purpose of this thesis, the analyses by Wells (1982b, pp. 120–124) on length contrasts will not be adopted, instead the length contrasts as in the case of Wells (2008) will be maintained, since our transcription aims to reflect the input listeners received, and thus a surface realisation is preferred. Related to the issue of length contrasts, Wells (1982b, pp. 121–124) denoted the phonemic representation of GOAT with a monophthong /o/ while Wells (2008) used a diphthong [oʊ]. The diphthong notation is preferred since it is the surface realisation.

Keyword	Phonemic Wells (1982b) (pp. 118–124)	Surface (Broad) Wells (2008)	Extrapolated	Surface (Narrow) Lindsey (2012b)	This thesis
KIT	/ɪ/	[ɪ]	[ɪ]	[ɪ]	[ɪ]
DRESS	/e/	[e]	[e]	[ɛ]	[ɛ]
TRAP	/æ/	[æ]	[æ]	[a]	[æ]
LOT	/ɒ/	[ɒ]	[ɒ]	[ɔ]	[ɒ]
STRUT	/ʌ/	[ʌ]	[ʌ]	[ə]	[ʌ]
FOOT	/ʊ/	[ʊ]	[ʊ]	[ɵ]	[ʊ]
BATH	/ɑː/	[ɑː]	[ɑː]	–	[ɑɑ]
CLOTH	/ɒ/	–	[ɒ]	–	[ɒ]
FLEECE	/iː/	[iː]	[iː]	[ɪj]	[ij]
FACE	/eɪ/	[eɪ]	[eɪ]	[ɛj]	[ej]
PALM	/ɑː/	–	[ɑː]	–	[ɑɑ]
THOUGHT	/ɔː/	[ɔː]	[ɔː]	–	[ɔɔ]
GOAT	/əʊ/	[əʊ]	[əʊ]	[əw]	[əw]
GOOSE	/uː/	[uː]	[uː]	[ʉw]	[ʉw]
PRICE	/aɪ/	[aɪ]	[aɪ]	[ɑj]	[aj]
CHOICE	/ɔɪ/	[ɔɪ]	[ɔɪ]	[oj]	[ɔj]
MOUTH	/aʊ/	[aʊ]	[aʊ]	[aw]	[aw]
HAPPY	/ɪ/	[i]	[i]	–	[i]
COMMA	/ə/	[ə]	[ə]	[ə]	[ə]
NURSE	/ɜː/	[ɜː]	[ɜː]	[əː]	[ɜɜ]
NEAR	/ɪə/	[ɪə]	[ɪə]	[ɪː]	[ɪə]
SQUARE	/ɛə/	[ɛə]	[ɛə]	[ɛː]	[ɛə]
START	/ɑː/	[ɑː]	[ɑː]	[ɑː]	[ɑɑ]
NORTH	/ɔː/	[ɔː]	[ɔː]	[oː]	[ɔɔ]
FORCE	/ɔː/	–	[ɔː]	–	[ɔɔ]
CURE	/ʊə/	[ʊə]	[ʊə]	[ɵː]	[ʊə]
LETTER	/ə/	–	[ə]	–	[ə]
MIRROR (KIT)	/ɪɪ/	–	[ɪɪ]	–	[ɪɪ]
MERRY (DRESS)	/ɛɪ/	–	[eɪ]	–	[ɛɪ]
CURRENT (NURSE)	–	–	[ɜːɪ]	–	[ɜɜɪ]
NEARER (NEAR)	/ɪəɪ/	–	[ɪəɪ]	–	[ɪəɪ]
MARY (SQUARE)	/ɛəɪ/	–	[ɛəɪ]	–	[ɛəɪ]
SAFARI (START)	–	–	[ɑːɪ]	–	[ɑɑɪ]
AURA (NORTH)	–	–	[ɔːɪ]	–	[ɔɔɪ]
ORAL (FORCE)	–	–	[ɔːɪ]	–	[ɔɔɪ]
CURIE (CURE)	–	–	[ʊəɪ]	–	[ʊəɪ]
LETTERING (LETTER)	–	–	[əɪ]	–	[əɪ]

Table 2.6: Vowel set: General British

2.2.7.2.2 NURSE and LETTER In GenAm, the NURSE and LETTER vowels are typically referred to as r-coloured vowels. Wells (2008) favoured the single symbol notation [ɜˠ], while Wells (1982b, pp. 120–124) favoured the dual symbol notation [ɜɪ][8]. Phonetically, the r-colouring is spread throughout the whole vowel (Wells,

[8]The vowel length difference and the type of the rhotic segment are irrelevant here.

1982b, p. 121), and this observation motivates the choice of [ɜˠ] which symbolises [ɜ] with r-colouring. The dual symbol notation preferred by Wells (1982b, pp. 121–124) was motivated by its parallelism with the START and NORTH vowels. Concretely, Wells (1982b, p. 121) argued that words such as *farm* and *form* often involve an r-coloured vowel which is the realisation of an underlying /ɹ/ segment as in /Vɹ/, and therefore in the case of GenBr and GenAm, the relationship between the two accent groups for the NURSE vowel, [ɜː] (GenBr) and [ɜɹ] (GenAm), will parallel those of the START and NORTH vowels – [ɑː] (GenBr) vs. [ɑɹ] (GenAm) and [ɔː] (GenBr) vs. [ɔɹ] (GenAm) respectively.

The dual symbol notation is preferred in this thesis. Firstly, since this vowel set will be used to transcribe dialectal speech and comparisons will be made between accent groups, the reasoning of Wells (1982b, pp. 120–124) on parallelism previously summarised is also applicable here. Secondly, while the r-colouring and the whole vowel are realised together in production, we do not have *a priori* any knowledge of how they would be processed in perception. By using a single symbol notation, we are assuming the r-colouring itself cannot be misperceived as an individual segment, i.e. it is inseparable from the [ɜ]. Having established the dual symbol notation for NURSE, the same should apply for LETTER and result in [əɹ] instead of [əˠ].

2.2.7.2.3 NORTH and FORCE Wells (1982b, pp. 120–124) maintained the difference between NORTH and FORCE vowels which are /ɔɹ/ and /oɹ/ respectively. Although not explicitly listed in Wells's (2008) vowel set table, a brief examination of the pronunciation used for keywords such as NORTH and FORCE showed that the NORTH-FORCE distinction is not maintained, and therefore the GenAm vowel set is assumed to have this merger. Since the Longman dictionary by Wells (2008) will be used as the key reference of GenAm pronunciation, it will be assumed that the GenAm vowel set in this thesis has the NORTH-FORCE merger.

2.2.7.2.4 Others The lax vowel [ɛ] is chosen for DRESS, and the short tense vowel [i] is chosen for HAPPY as discussed in Section 2.2.7.1.1. Similarly, the treatments of offglides, length marks and extrapolations of missing vowels are the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section 2.2.7.1.4.

2.2.7.3 New England

To determine an appropriate vowel set for the New England accent, and more specifically the Eastern New England accent, I examined Wells (1982d, pp. 518–527). I identified and discussed the modifications I made to the phonemic vowel set by Wells (1982d, pp. 518–527). I utilised Wells's detailed discussion of the surface realisation of the accent and made similar modifications to other accents for consistency. The final vowel set for the New England accent for this thesis is listed in the final column in Table 2.8.

2.2.7.3.1 Non-rhoticity Wells (1982d, pp. 520–522) discusses the status of rhoticity in the New England accent. In the early twentieth century, the New England accent is traditionally described as being non-rhotic. Its non-rhoticity is described as the loss of historical /r/ except before vowels. However, the accent is undergoing rhoticity and studies on the accent of Boston (the principal city of eastern New England) in the mid twentieth century showed that Boston speakers are more /r/-ful in formal speech (Labov, Ash, and Boberg, 2005). More recent quantitative analyses by Irwin and Nagy (2007) have shown that there is a high degree of variability, depending on the phoneme (e.g. NURSE being more r-ful than LETTER), as well as social factors including age, gender, and education. Given the high degree of variability, an accurate estimate of rhoticity is not possible. Variability can also be found in the transcriptions provided by the reporters of the naturalistic corpus in this thesis. However, this information is not sufficient nor reliable for determining the rhoticity of a given word in the corpus. Firstly the transcriptions usually only contain the key

Keyword	Phonemic Wells (120–124 1982b)	Surface Wells (2008)	Extrapolated	This thesis
KIT	/ɪ/	[ɪ]	[ɪ]	[ɪ]
DRESS	/ɛ/	[e]	[e]	[ɛ]
TRAP	/æ/	[æ]	[æ]	[æ]
LOT	/ɑ/	[ɑː]	[ɑː]	[ɑɑ]
STRUT	/ʌ/	[ʌ]	[ʌ]	[ʌ]
FOOT	/ʊ/	[ʊ]	[ʊ]	[ʊ]
BATH	/æ/	[æ]	[æ]	[æ]
CLOTH	/ɔ/	–	[ɔː]	[ɔɔ]
FLEECE	/i/	[iː]	[iː]	[ij]
FACE	/eɪ/	[eɪ]	[eɪ]	[ej]
PALM	/ɑ/	–	[ɑː]	[ɑɑ]
THOUGHT	/ɔ/	[ɔː]	[ɔː]	[ɔɔ]
GOAT	/o/	[oʊ]	[oʊ]	[ow]
GOOSE	/u/	[uː]	[uː]	[uw]
PRICE	/aɪ/	[aɪ]	[aɪ]	[aj]
CHOICE	/ɔɪ/	[ɔɪ]	[ɔɪ]	[ɔj]
MOUTH	/aʊ/	[aʊ]	[aʊ]	[aw]
HAPPY	/ɪ/	[i]	[i]	[i]
COMMA	/ə/	[ə]	[ə]	[ə]
NURSE	/ɜɹ/	[ɜˇː]	[ɜˇː]	[ɜɜɹ]
NEAR	/ɪɹ/	–	[ɪɹ]	[ɪɹ]
SQUARE	/ɛɹ/	–	[ɛɹ]	[ɛɹ]
START	/ɑɹ/	[ɑː]	[ɑː]	[ɑɑɹ]
NORTH	/ɔɹ/	[ɔː]	[ɔː]	[ɔɔɹ]
FORCE	/oɹ/	–	[ɔːɹ]	[ɔɔɹ]
CURE	/ʊɹ/	–	[ʊɹ]	[ʊɹ]
LETTER	/əɹ/	–	[əɹ]	[əɹ]
MIRROR (KIT)	/ɪɹ/	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	/ɛɹ/	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜˇː]	[ɜɜɹ]
NEARER (NEAR)	/ɪɹ/	–	[ɪɹ]	[ɪɹ]
MARY (SQUARE)	/ɛɹ/	–	[ɛɹ]	[ɛɹ]
SAFARI (START)	–	–	[ɑːɹ]	[ɑɑɹ]
AURA (NORTH)	–	–	[ɔːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔːɹ]	[ɔɔɹ]
CURIE (CURE)	–	–	[ʊɹ]	[ʊɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.7: Vowel set: General American

words that are different rather than the entire sentence, meaning the information is incomplete. Secondly the transcriptions by some reporters were inconsistent and incorrect; for instance, [ɔˇ] is sometimes used for cases that are clearly [ə][9]. A practical decision is made to assume non-rhoticity for the vowel set used for the New England accent in this thesis.

2.2.7.3.2 Vowel lengths As asserted by Wells (1982d, p. 519), length marks could be used for the monophthongs in what the author called *part-systems B, C and D* (a system for subdividing English vowel systems) which has a similar structure to RP (Wells, 1982b, p. 182). Part B consists of traditional long vowels that are front and diphthongs which have endpoints that are front. Part C consists of traditional long vowels that are back and diphthongs which have endpoints that are back. Part D consists of traditional long vowels that are open and diphthongs which have endpoints that are open. This essentially translates as having the following monophthongs carrying a length mark (i.e. long) – /i/ (FLEECE) as [iː], /u/ (GOOSE) as [uː], /ɜ/ (NURSE) as [ɜː] and /a/ (START) as [aː]. Regarding the monophthong in CLOTH and THOUGHT /ɒ/, since LOT is commonly merged with CLOTH-THOUGHT (Wells, 1982d, p. 524), all three vowels in the lexical set are indicated as having a length mark, [ɒː] (Wells, 1982d, p. 519).

2.2.7.3.3 BATH Wells (1982d, pp. 518–527) suggested two possible symbols for bath, /a/ and /æ/. Wells (1982d, p. 523) reported that the use of the [a] variant was in sharp decline in New England, as suggested by multiple surveys (Miller, 1953; Thomas, 1961) that [æ] was more dominant.

From a small survey of 14 Boston informants by Laferriere (1977), BATH–raising, [æ] to [ɛə], was in force and it was more productive with younger speakers. From

[9]Although one could argue these are accurate transcriptions of the results of hyper-rhotic pronunciation, e.g. *cough* [kɒːɹf], Wells (1982d, p. 522) suggested that these pronunciations are unlikely to become established.

this, Wells (1982d, p. 523) suggested that "as the older [a] declines, the new [ɛə] takes over, often without a stop at the hitherto standard GenAm [æ] type.".

However, a more recent meta-study by Nagy and Roberts (2004) on the phonology of the New England accent came to the conclusion that the Eastern New England accent has no BATH/TRAP/DANCE-raising, except for Boston as reported by Laferriere (1977) (Nagy and Roberts, 2004, p. 260). In conclusion, there is a lack of strong evidence for BATH-raising in the Eastern England accent, given that the small survey by Laferriere (1977) on Boston might be undersampled and might not be generalised to the rest of Eastern New England. Therefore BATH will be denoted with [æ] in our vowel set.

2.2.7.3.4 Others The treatments of offglides, length marks and extrapolations of missing vowels are the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section 2.2.7.1.4. For GOAT, the underlying monophthong /o/ is treated as being realised by a diphthong [oʊ], following the same argument as in Section 2.2.7.2.1. The short tense vowel [i] is chosen for HAPPY as discussed in Section 2.2.7.1.1.

2.2.7.4 Southern American

I examined Wells (1982d, pp. 527–553) to determine an appropriate vowel set for the Southern American accent. I identified the modifications I made to the phonemic vowel set by Wells (1982d, pp. 531, 550). I utilised Wells's detailed discussion of the surface realisation of the accent and made similar modifications to other accents for consistency. Wells (1982d, pp. 542–545) examined two varieties of the Southern accent, rhotic and non-rhotic. The main difference between them is the realisation of vowels followed by a historical /r/, namely NURSE to LETTER in the non-pre-vocalic group. In the following paragraphs, the other vowels will be discussed since they are identical for both varieties. Finally, the historical /r/ vowels will be discussed. The reported transcription will be used to choose between the two varieties during

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 518–527)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/	–	[æ]	[æ]
LOT	/ɒ/	[ɒː]	[ɒː]	[ɒɒ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/a/, /æ/	[a], [eə]	[æ]	[æ]
CLOTH	/ɒ/	[ɒː]	[ɒː]	[ɒɒ]
FLEECE	/i/	[iː]	[iː]	[ij]
FACE	/eɪ/	–	[eɪ]	[ej]
PALM	/a/	–	[aː]	[aa]
THOUGHT	/ɒ/	[ɒː]	[ɒː]	[ɒɒ]
GOAT	/o/	–	[oʊ]	[ow]
GOOSE	/u/	[uː]	[uː]	[uw]
PRICE	/aɪ/	–	[aɪ]	[aj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/aʊ/	–	[aʊ]	[aw]
HAPPY	/ɪ/, /i/	–	[ɪ], [i]	[i]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜ/	–	[ɜː]	[ɜɜ]
NEAR	/iə/	–	[iə]	[iə]
SQUARE	/æə/	–	[æə]	[æə]
START	/a/	[aː]	[aː]	[aa]
NORTH	/ɒ/	–	[ɒː]	[ɒɒ]
FORCE	/oə/	–	[oə]	[oə]
CURE	/uə/	–	[uə]	[uə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜɹ]	[ɜɹ]
NEARER (NEAR)	–	–	[iəɹ]	[iəɹ]
MARY (SQUARE)	–	–	[æəɹ]	[æəɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[ɒːɹ]	[ɒɒɹ]
ORAL (FORCE)	–	–	[oəɹ]	[oəɹ]
CURIE (CURE)	–	–	[uəɹ]	[uəɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.8: Vowel set: New England

re-transcription. If the reported transcription did not indicate that the speaker was a rhotic speaker, then I assume the speaker was rhotic. The final vowel set for the Southern accent used in this thesis is listed in the final column in Table 2.9 (non-rhotic) and Table 2.10 (rhotic).

2.2.7.4.1 Umlaut and Shading The six vowels /ɪ ɛ æ ʊ ʌ ɑ/ were classified as being lax in the Southern accent and two processes were reported to affect them (Wells, 1982d, pp. 533–537): *Umlaut* and *Shading*. Umlaut is a process in which the frontness of a vowel is affected by the vowel in the following syllable, e.g. the /ɪ/ vowel is realised as being successively less front in r**i**pping (the most front), r**i**p and r**i**pper (the least front). Shading is a process in which the frontness of a vowel (particularly /ɪ/) is affected by the following consonant. /ɪ/ is realised faithfully as [ɪ] only when followed by a velar consonant; before a labial consonant, it is realised more centrally [ɪ̟] or in the notation of Wells (1982d, p. 534) [ɨ].

For both Umlaut and Shading, Wells (1982d, pp. 533–537) used the IPA diacritics, [ẍ] and [x̟] (*x* being a placeholder for the diacritics), to denote backing and fronting. For the purposes of this thesis, these phonetic differences are ignored to avoid introducing new types of segments and therefore will not be adopted.

2.2.7.4.2 Schwa offglides In a stressed monosyllable, the three lax vowels /ɪ ɛ æ/ were reported to have a prominent schwa offglide when followed by a labial consonant, and are instead realised as [ɪə ɛə æə], while they would be lengthened when followed by a non-labial consonant (Wells, 1982d, p. 535) and are realised as [ɪː ɛː æː]. For example, *lip* is realised as [lɪəp], while *bid*, *bed* and *bad* are realised as [bɪːd], [bɛːd] and [bæːd]. This rule will be applied during transcription.

2.2.7.4.3 FOOT The FOOT vowel is often not rounded but more central and unrounded as in [ɪ̟] or in the notation of Wells (1982d, p. 534) [ɨ]. However, this

varies with different regions. In tidewater accents, it remains back with varying degrees of roundness [ʊ̟] (or in the notation of Wells (1982d, p. 536) [ω]) or [ʊ]. Due to its variability in both roundness and backness, a practical decision was made for the vowel to be realised as [ʊ] in our vowel set, assuming it is realised faithfully with respect to its underlying form.

2.2.7.4.4 TRAP and BATH The TRAP–BATH vowel was proposed to have /æ/ as the underlying form and /æɪ/ as a marginally phonemic/underlying form, since there are a small number of minimal pairs, e.g. [kæɪm] *can* "container" and [kæn] *can* "be able" Wells (1982d, pp. 531-532). However, I rejected this dual-phoneme analysis, and instead I follow McMillan's (1946) rule-based analysis, where /æ/ is realised as [æɪ] before a following /g/, voiceless fricatives /f θ s ʃ/, /v/ or /n/ (and possibly /d/), in monosyllables. This rule was applied throughout the transcription.

2.2.7.4.5 STRUT The STRUT vowel was suggested to have several different qualities (Wells, 1982d, p. 536): [ɜ] (the most typical), [ɤ̟] and [ʌ] (in the tidewater area); some speakers have an allophonic alternation with [ɜ] elsewhere and [ʌ] before a labial. As [ɜ] is the most typical realisation, this was chosen for my vowel set and the other realisations were rejected on the basis of their high variability.

2.2.7.4.6 LOT The LOT vowel varies along the front-back dimension, from [ɑ] to [ɑ̟]. These two realisations can have an allophonic relationship for many southerners, e.g. [ɑ] before a velar and /w/, and [ɑ̟] elsewhere (Wells, 1982d, p. 536). For simplicity, [ɑ] is chosen for my vowel set, over the fronted variant.

2.2.7.4.7 PRICE and MOUTH The PRICE vowel is commonly realised as a (near-)monophthong, [a(ː)] or [aɛ] (Wells, 1982d, pp. 537–538). In prestigious varieties, there is also an allophonic relationship that mirrors Canadian Raising, between [a(ː)] and [aɪ], with [aɪ] before a voiceless consonant within the same syllable and

[a(ː)] elsewhere. Given that the utterers in our corpus were mostly Harvard students, I will assume they spoke a more prestigious variant which has this allophonic rule. Furthermore, given the length of the monophthongal form is variable, for simplicity I will assume that the monophthongal form is long [aː]. The most usual realisation of the MOUTH vowel is [æʊ] (Wells, 1982d, p. 538), so this was the version adopted in my vowel set.

2.2.7.4.8 FLEECE and GOOSE The FLEECE vowel is commonly realised as monophthongal, and short or slightly lengthened, as in [i] or [iˑ], while the GOOSE vowel is generally diphthongal but central, [ʉu] or [ʉˑ] (Wells, 1982d, p. 539). The vowels [i] and [ʉu] respectively were chosen to represent these vowels in my vowel set. The lengthened variants were not chosen on the basis of them needing to use an additional diacritic for length and avoiding questions whether a lengthened segment constitutes, e.g. one or two alignment units (see Section 2.4.3 for discussion of the minimal alignment unit).

2.2.7.4.9 FACE and GOAT The FACE vowel has a mid starting point, as asserted by Wells (1982d, p. 539), as [e̝ɪ]. The GOAT vowel is realised as [oʊ] in most parts of the south. I have used [eɪ] for FACE, and [oʊ] for GOAT for my vowel set. The change in starting point for FACE was ignored in order to remove the need for an additional diacritic for vowel quality.

2.2.7.4.10 PALM The PALM vowel is realised as either LOT or non-rhotic START. Both realisations were kept in my vowel set, with an arbitrary preference for the non-rhotic START realisation. The LOT realisation will only be used if it is indicated in the transcriptions provided by the reporters.

2.2.7.4.11 THOUGHT and CLOTH The THOUGHT vowel is most stereotypically realised as [ɒʊ] a closing diphthong, but it can also be realised as monophthongal

[ɔ] or [ɒ]. I will keep all three variants for my vowel set, giving priority to [ɒʊ]; [ɔ] and [ɒ] will be used only if it is indicated in the transcriptions provided by the reporters. Given that the underlying form for CLOTH is the same as THOUGHT, I will assume there is a CLOTH-THOUGHT merger.

2.2.7.4.12 CHOICE The CHOICE vowel may have two allophones for those who have [ɪ] for HAPPY , with [ɔɛ̈] occurring word finally, and [ɔɪ] or [ɔe] elsewhere. Given it is not reported in our data whether a particular utterer has ɪ for HAPPY or not, it is not possible to know if this allophonic rule should apply or not. Limited by the level of detail provided by the reporters, I will assume [ɔɪ] as the sole realisation for CHOICE.

2.2.7.4.13 R-vowels The vowel set of the R-vowels is summarised in Table 2.9 and Table 2.10, namely NURSE to LETTER in the non-pre-vocalic group, and CURRENT to LETTERING in the pre-vocalic group.

Wells (1982d, p. 550) summarised the phonemic forms for the R-vowels for both rhotic and non-rhotic accents for both the pre-vocalic and non-pre-vocalic groups, with specific mention of the surface forms for NURSE and LETTER. The surface realisations of some R-vowels were available, such as FORCE, CURE and START. The analyses by Wells indicated considerable variations in both surface realisation and phonemic analyses. To simplify the possible realisations, the first realisation from the list was picked for each vowel.

For the non-pre-vocalic group, the vowel set for the non-rhotic accent is as follows, NURSE: [ɜː], NEAR: [ɪə], SQUARE: [æə], START: [ɑː], NORTH: [ɔː], FORCE: [oə], CURE: [oə], and LETTER: [ə]. Similarly, for the rhotic accent, they are as follows, NURSE: [ɜˠ], NEAR: [ɪɹ], SQUARE: [æɹ], START: [ɑɹ], NORTH: [ɔɹ], FORCE: [oɹ], CURE: [oɹ], and LETTER: [ə].

The pre-vocalic group is the same for both accents: CURRENT: [ɜɹ], PERIOD:

[iɹ], MARY: [eɪɹ], MARY: [eɪɹ], SAFARI: [ɑɹ], AURA: [ɔɹ], ORAL: [oɹ], CURIE: [oɹ] and LETTERING: [əɹ]. CURRENT was chosen to be [ɜɹ] instead of the underlying form /ʌɹ/ to maintain consistency with the STRUT vowel and its non-pre-vocalic form NURSE.

2.2.7.4.14 HAPPY and COMMA While the HAPPY vowel has [ɪ] as the common realisation, the rhotic accent has [i] as another possible realisation. For the rhotic accent, I accepted both forms for the HAPPY vowel, giving more priority to [ɪ] and only using [i] if the reported transcription indicated otherwise. The COMMA vowel was [ə] for both accents, and although it was reported that the rhotic accent has a possible (stigmatized) form [ə˞] in southern mountain speech, this was rejected for being too regionally-specific.

2.2.7.4.15 Others The treatments on offglides and length marks are the same as in Section 2.2.7.1.2 and Section 2.2.7.1.3. All *-ing* suffixes have two realisations [iŋ] and [ɪŋ] (Wells, 1982d, p. 550); [iŋ] is assumed unless [ɪŋ] is part of the reported transcriptions.

2.2.7.5 New York City

To determine an appropriate vowel set for the New York City accent, I examined Wells (1982d, pp. 501–518). I identified the modifications I made to the phonemic vowel set by Wells (1982d, p. 503). I utilised Wells's detailed discussion of the surface realisation of the accent and made similar modifications to other accents for consistency. Wells (1982d, pp. 505–508) examined the rhotic and non-rhotic varieties of the New York City accent. The main difference between them is the realisation of the vowels followed by a historical /r/, namely NURSE to LETTER in the non-pre-vocalic group. In the following paragraphs, identical vowels for both varieties will first be discussed, and then the difference between rhotic and non-rhotic will be highlighted. The reported transcription will be used to choose between the two

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 530–552)			
KIT	/ɪ/	[ɪ] ∼ [ɪə], [ɪː]	[ɪ] ∼ [ɪə], [ɪː]	[ɪ] ∼ [ɪə], [ɪɪ]
DRESS	/ɛ/	[ɛ] ∼ [ɛə], [ɛː]	[ɛ] ∼ [ɛə], [ɛː]	[ɛ] ∼ [ɛə], [ɛɛ]
TRAP	/æ/, /æɪ/	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æj], [æə], [ææ]
LOT	/ɑ/	[ɑ] ∼ [ɑ̟]	[ɑ] ∼ [ɑ̟]	[ɑ]
STRUT	/ʌ/	[ɜ], [ɣ̟]	[ɜ], [ɣ̟]	[ɜ]
FOOT	/ʊ/	[ɨ̟]	[ɨ̟]	[ʊ]
BATH	/æ/, /æɪ/	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æj], [æə], [ææ]
CLOTH	/ɔ/	–	[ɒʊ̯], [ɔ], [ɒ]	[ɒw], [ɔ], [ɒ]
FLEECE	/i/	[i], [iˑ]	[i], [iˑ]	[i]
FACE	/eɪ/	[e̝ɪ]	[e̝ɪ]	[ej]
PALM	/ɑ/	[ɑː], [ɑ]	[ɑː], [ɑ]	[ɑɑ], [ɑ]
THOUGHT	/ɔ/	[ɒʊ̯], [ɔ], [ɒ]	[ɒʊ̯], [ɔ], [ɒ]	[ɒw], [ɔ], [ɒ]
GOAT	/oʊ/	[oʊ̯]	[oʊ̯]	[ow]
GOOSE	/u/	[ʉu], [ʉˑ]	[ʉu], [ʉˑ]	[ʉw]
PRICE	/aɪ/	[aː], [aɛ] ∼ [aɪ]	[aː], [aɛ] ∼ [aɪ]	[aa] ∼ [aj]
CHOICE	/ɔɪ/	[ɔɪ], [ɔe] ∼ [ɔɛ̈]	[ɔɪ], [ɔe] ∼ [ɔɛ̈]	[ɔj]
MOUTH	/æʊ/	[æʊ̯]	[æʊ̯]	[æw]
HAPPY	/ɪ/	[ɪ]	[ɪ]	[ɪ]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜː/	[ɜː], [ɜɪ]	[ɜː], [ɜɪ]	[ɜɜ]
NEAR	/ɪə/, /ɛə/	–	[ɪə]	[ɪə]
SQUARE	/æə/, /ɛə/	–	[æə]	[æə]
START	/ɑː/, /ɑ/	[ɑː]	[ɑː]	[ɑɑ]
NORTH	/ɔː/, /ɔ/	[ɔː]	[ɔː]	[ɔɔ]
FORCE	/oə/, /oʊ/	[oə], [oʊ̯]	[oə], [oʊ̯]	[oə]
CURE	/oə/, /ʊə/, /oʊ/	[oə], [ʊə], [oʊ̯]	[oə], [ʊə], [oʊ̯]	[oə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	/ʌɹ/	–	[ɜɹ]	[ɜɹ]
NEARER (NEAR)	/iɹ/	–	[iɹ]	[iɹ]
MARY (SQUARE)	/eɪɹ/	–	[eɪɹ]	[ejɹ]
SAFARI (START)	/ɑɹ/, /ɔɹ/	–	[ɑɹ]	[ɑɹ]
AURA (NORTH)	/ɔɹ/	–	[ɔɹ]	[ɔɹ]
ORAL (FORCE)	/oʊɹ/	[oɹ]	[oɹ]	[oɹ]
CURIE (CURE)	/oʊɹ/, /uɹ/	[oɹ]	[oɹ]	[oɹ]
LETTERING (LETTER)	/əɹ/	–	[əɹ]	[əɹ]

Table 2.9: Vowel set: Southern – non-rhotic: "∼" denotes an allophonic relationship

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 530–552)			
KIT	/ɪ/	[ɪ] ∼ [ɪə], [ɪː]	[ɪ] ∼ [ɪə], [ɪː]	[ɪ] ∼ [ɪə], [ɪɪ]
DRESS	/ɛ/	[ɛ] ∼ [ɛə], [ɛː]	[ɛ] ∼ [ɛə], [ɛː]	[ɛ] ∼ [ɛə], [ɛɛ]
TRAP	/æ/, /æɪ/	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æj], [æə], [ææ]
LOT	/ɑ/	[ɑ] ∼ [ɑ̟]	[ɑ] ∼ [ɑ̟]	[ɑ]
STRUT	/ʌ/	[ɜ], [ɣ̟]	[ɜ], [ɣ̟]	[ɜ]
FOOT	/ʊ/	[ʉ̟]	[ʊ]	[ʊ]
BATH	/æ/, /æɪ/	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æɪ], [æə], [æː]	[æ] ∼ [æj], [æə], [ææ]
CLOTH	/ɔ/	–	[ɒʊ], [ɔ], [ɒ]	[ɒw], [ɔ], [ɒ]
FLEECE	/i/	[i], [iˑ]	[i]	[i]
FACE	/eɪ/	[e̝ɪ]	[eɪ]	[ej]
PALM	/ɑ/	[ɑː], [ɑ]	[ɑː], [ɑ]	[ɑɑ], [ɑ]
THOUGHT	/ɔ/	[ɒʊ̯], [ɔ], [ɒ]	[ɒʊ̯], [ɔ], [ɒ]	[ɒw], [ɔ], [ɒ]
GOAT	/oʊ/	[oʊ̯]	[oʊ̯]	[ow]
GOOSE	/u/	[ʉu], [ʉˑ]	[ʉu]	[ʉw]
PRICE	/aɪ/	[aː], [aɛ] ∼ [aɪ]	[aː], [aɛ] ∼ [aɪ]	[aa] ∼ [aj]
CHOICE	/ɔɪ/	[ɔɪ], [ɔe] ∼ [ɔë]	[ɔɪ], [ɔe] ∼ [ɔë]	[ɔj]
MOUTH	/æʊ/	[æʊ̯]	[æʊ̯]	[æw]
HAPPY	/ɪ/	[ɪ], [i]	[ɪ], [i]	[ɪ], [i]
COMMA	/ə/, /əɪ/	–	[ə]	[ə]
NURSE	/ʌɪ/	[ɜˠ]	[ɜˠ]	[ɜɪ]
NEAR	/ɪɪ/, /ɛɪ/, /æɪ/	–	[ɪɪ]	[ɪɪ]
SQUARE	/æɪ/, /æɪɪ/, /ɛɪ/	–	[æɪ]	[æɪ]
START	/ɑɪ/, /ɔɪ/	[ɑɪ]	[ɑɪ]	[ɑɪ]
NORTH	/ɔɪ/	–	[ɔɪ]	[ɔɪ]
FORCE	/oʊɪ/, /oʊ/	[oɪ]	[oɪ]	[oɪ]
CURE	/oʊɪ/, /ɯɪ/, /ʊɪ/, /oʊ/	[oɪ]	[oɪ]	[oɪ]
LETTER	/əɪ/	–	[əɪ]	[əɪ]
MIRROR (KIT)	–	–	[ɪɪ]	[ɪɪ]
MERRY (DRESS)	–	–	[ɛɪ]	[ɛɪ]
CURRENT (NURSE)	/ʌɪ/	–	[ɜɪ]	[ɜɪ]
NEARER (NEAR)	/iɪ/	–	[iɪ]	[iɪ]
MARY (SQUARE)	/eɪɪ/	–	[eɪɪ]	[ejɪ]
SAFARI (START)	/ɑɪ/, /ɔɪ/	–	[ɑɪ]	[ɑɪ]
AURA (NORTH)	/ɔɪ/	–	[ɔɪ]	[ɔɪ]
ORAL (FORCE)	/oʊɪ/	[oɪ]	[oɪ]	[oɪ]
CURIE (CURE)	/oʊɪ/, /ɯɪ/	[oɪ]	[oɪ]	[oɪ]
LETTERING (LETTER)	/əɪ/	–	[əɪ]	[əɪ]

Table 2.10: Vowel set: Southern – rhotic: "∼" denotes an allophonic relationship

varieties during re-transcription. If the reported transcription did not indicate that the speaker was rhotic, then the assumption will be that the speaker was rhotic. The final vowel set for the New York City accent for this thesis is listed in the final column of Table 2.11 (non-rhotic) and Table 2.12 (rhotic).

2.2.7.5.1 FLEECE and GOOSE FLEECE and GOOSE are commonly realised as diphthongs. The starting point of the vowel varies sociolinguistically, with [ɪi] and [ʊu] more likely by the middle-class, and [i̟i] and [ʊ̟u] by the working-class. Furthermore, GOOSE has an additional common realisation that is monophthongal [uː] but slightly rounded.

In the vowel set, I selected the diphthongal variants more associated with the middle-class, [ɪi] and [ʊu] for FLEECE and GOOSE. Since HAPPY is underlyingly the same as FLEECE, I assumed it has the same surface forms.

2.2.7.5.2 BATH-raising /æ/ is realised as a closer, longer, diphthongal variant (e.g. [æːə]) before a final voiced stop, a voiceless fricative, or /m, n/, if one of these consonants is followed by an inflectional boundary or an obstruent, but not if the consonant is followed by a vowel or liquid. Elsewhere this vowel is realised as [æ] (Wells, 1982d, p. 510). The exact realisation of the raised variant varies sociolinguistically (Wells, 1982d, pp. 511–512). For simplicity, I included only the two common realisations [æːə] and [ɛːə].

2.2.7.5.3 CLOTH and THOUGHT /ɔə/ in CLOTH and THOUGHT is raised for the starting-point of the vowel, with [ɔ̝] by the standard and upper-middle-class and with [o] in casual middle-class speech or careful reading style of the lower class (Wells, 1982d, p. 513). For simplicity, I chose only the standard/upper-middle-class realisations.

2.2.7.5.4 NURSE and CHOICE The traditional realisation of NURSE and preconsonantal CHOICE is [ɜɪ]. However, this realisation was reported to be sharply disfavoured, and [ɜˇ] and [ɔɪ] were growing in favour instead for NURSE and CHOICE respectively (Wells, 1982d, p. 508). Therefore, [ɜˇ] and [ɔɪ] were chosen for my vowel set.

2.2.7.5.5 Centring diphthongs The centring diphthongs are /ɪə/, /ɛə/, /ɑə/, /ɔə/, /ɔə/ and /ʊə/, and they are not restricted to historical /r/ vowels. They are realised monophthongally when followed by an intervocalic consonant within a word, such as *hearing*, *dairy*, *Chicago*, *sausage* and *curious*.

The non-rhotic variety has the following realisations: [ɪːə] ∼ [ɪː], [ɛːə], ∼ [ɛː], [ɑ̟ːə] ∼ [ɑ̟ː], [ɔ̟ːə] ∼ [ɔ̟ː] and [ʊːə] ∼ [ʊː]. The realisations in the rhotic variety are simply appended with /r/, with the exception of /ɑə/ which has an additional realisation as [ɑ̟ɹ].

2.2.7.5.6 LOT and START In the non-rhotic variety, LOT and START are distinguished by the difference in their backness and/or duration, such that START is more back and/or longer than LOT, with START as [ɑ̟ːə] or [ɑ̟ː], and LOT as [ɑ]. In the rhotic variety, these differences may be present or absent given that are redundant with the presence of /r/, with START as [ɑ̟ɹ] and LOT as [ɑ̟]. In my vowel set for the rhotic variety, I gave priority to the realisations in which these differences are absent. Furthermore, LOT has undergone a historical sound change called LOT-lengthening, such that /ɑ/ was changed to a centring diphthong /ɑə/ before word-final voiced stops /b, d, dʒ, g/. These LOT words with /ɑə/ underwent the monophthongisation process as with other centring diphthongs and are homophonous with START.

2.2.7.5.7 Others The treatment of offglides, length marks and extrapolations of missing vowels is the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section

2.2.7.1.4.

Keyword non-rhotic	Phonemic Wells (1982d, pp. 503–515)	Surface	Extrapolated	This thesis
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/, /æə/, /ɛə/	[æ] ∼ [æːə], [ɛːə]	[æ] ∼ [æːə], [ɛːə]	[æ] ∼ [ææə], [ɛɛə]
LOT	/ɑ/, /ɑə/	[ɑ]; [ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ]; [ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ]; [ɑɑə] ∼ [ɑɑ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/æə/, /ɛə/	[æːə], [ɛːə]	[æːə], [ɛːə]	[ææə], [ɛɛə]
CLOTH	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
FLEECE	/i/, /ɪi/	[ɪi], [i̟i], [i̟ː]	[ɪi], [i̟i], [i̟ː]	[ɪj]
FACE	/eɪ/	–	[eɪ]	[ej]
PALM	/ɑə/	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑɑə] ∼ [ɑɑ]
THOUGHT	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
GOAT	/oʊ/	–	[oʊ]	[ow]
GOOSE	/u/, /ʊu/, /ɪu/	[ʊu], [ʊ̟u], [uː]	[ʊu], [ʊ̟u], [uː]	[ʊw]
PRICE	/ɑɪ/	–	[ɑɪ]	[ɑj]
CHOICE	/ɔɪ/	[ɔɪ], [ɜɪ]	[ɔɪ], [ɜɪ]	[ɔj]
MOUTH	/aʊ/	–	[aʊ]	[aw]
HAPPY	/i/, /ɪi/	[ɪi], [i̟i], [i̟ː]	[ɪi], [i̟i], [i̟ː]	[ɪj]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜ/	[ɜˤ], [ɜɪ]	[ɜˤ], [ɜɪ]	[ɜɹ]
NEAR	/ɪə/	[ɪːə] ∼ [ɪː]	[ɪːə] ∼ [ɪː]	[ɪɪə] ∼ [ɪɪ]
SQUARE	/ɛə/	[ɛːə] ∼ [ɛː]	[ɛːə] ∼ [ɛː]	[ɛɛə] ∼ [ɛɛ]
START	/ɑə/	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑɑə] ∼ [ɑɑ]
NORTH	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
FORCE	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
CURE	/ʊə/	[ʊːə] ∼ [ʊː]	[ʊːə] ∼ [ʊː]	[ʊʊə] ∼ [ʊʊ]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜˤ]	[ɜɹ]
NEARER (NEAR)	–	–	[ɪːɹ]	[ɪɪɹ]
MARY (SQUARE)	–	–	[ɛːɹ]	[ɛɛɹ]
SAFARI (START)	–	–	[ɑ̟ːɹ]	[ɑɑɹ]
AURA (NORTH)	–	–	[ɔ̟ːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔ̟ːɹ]	[ɔɔɹ]
CURIE (CURE)	–	–	[ʊːɹ]	[ʊʊɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.11: Vowel set: New York City – non-rhotic: "∼" denotes an allophonic relationship

Keyword non-rhotic	Phonemic Wells (1982d, pp. 503–515)	Surface	Extrapolated	This thesis
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/, /æə/, /ɛə/	[æ] ∼ [æːə], [ɛːə]	[æ] ∼ [æːə], [ɛːə]	[æ] ∼ [ææə], [ɛɛə]
LOT	/ɑ/, /ɑə/	[ɑ̟]; [ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ̟]; [ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ]; [ɑɑə] ∼ [ɑɑ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/æə/, /ɛə/	[æːə], [ɛːə]	[æːə], [ɛːə]	[ææə], [ɛɛə]
CLOTH	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
FLEECE	/i/, /ii/	[ii], [i̟i̟], [iː]	[ii], [i̟i̟], [iː]	[ij]
FACE	/eɪ/	–	[eɪ]	[ej]
PALM	/ɑə/	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑ̟ːə] ∼ [ɑ̟ː]	[ɑɑə] ∼ [ɑɑ]
THOUGHT	/ɔə/	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔ̟ːə] ∼ [ɔ̟ː]	[ɔɔə] ∼ [ɔɔ]
GOAT	/oʊ/	–	[oʊ]	[ow]
GOOSE	/u/, /ʊu/, /ɯ/	[ʊu], [ə̟u], [uː]	[ʊu], [ə̟u], [uː]	[əw]
PRICE	/ɑɪ/	–	[ɑɪ]	[ɑj]
CHOICE	/ɔɪ/	[ɔɪ], [ɜɪ]	[ɔɪ], [ɜɪ]	[ɔj]
MOUTH	/aʊ/	–	[aʊ]	[aw]
HAPPY	/i/, /ii/	[ii], [i̟i̟], [iː]	[ii], [i̟i̟], [iː]	[ij]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜɹ/	[ɜˠ], [ɜɪ]	[ɜˠ], [ɜɪ]	[ɜɹ]
NEAR	/ɪəɹ/	[ɪːəɹ] ∼ [ɪːɹ]	[ɪːəɹ] ∼ [ɪːɹ]	[ɪəɹ] ∼ [ɪɹ]
SQUARE	/ɛəɹ/	[ɛːəɹ] ∼ [ɛːɹ]	[ɛːəɹ] ∼ [ɛːɹ]	[ɛɛəɹ] ∼ [ɛɛɹ]
START	/ɑəɹ/, /ɑɹ/	[ɑ̟ɹ]; [ɑ̟ːəɹ] ∼ [ɑ̟ːɹ]	[ɑ̟ɹ]; [ɑ̟ːəɹ] ∼ [ɑ̟ːɹ]	[ɑɹ]; [ɑɑəɹ] ∼ [ɑɑɹ]
NORTH	/ɔəɹ/	[ɔ̟ːəɹ] ∼ [ɔ̟ːɹ]	[ɔ̟ːəɹ] ∼ [ɔ̟ːɹ]	[ɔɔəɹ] ∼ [ɔɔɹ]
FORCE	/ɔəɹ/	[ɔ̟ːəɹ] ∼ [ɔ̟ːɹ]	[ɔ̟ːəɹ] ∼ [ɔ̟ːɹ]	[ɔɔəɹ] ∼ [ɔɔɹ]
CURE	/ʊəɹ/	[ʊːəɹ] ∼ [ʊːɹ]	[ʊːəɹ] ∼ [ʊːɹ]	[ʊʊəɹ] ∼ [ʊʊɹ]
LETTER	/əɹ/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜˠ]	[ɜɹ]
NEARER (NEAR)	–	–	[ɪːɹ]	[ɪɪɹ]
MARY (SQUARE)	–	–	[ɛːɹ]	[ɛɛɹ]
SAFARI (START)	–	–	[ɑ̟ɹ], [ɑ̟ːɹ]	[ɑɹ], [ɑɑɹ]
AURA (NORTH)	–	–	[ɔ̟ːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔ̟ːɹ]	[ɔɔɹ]
CURIE (CURE)	–	–	[ʊːɹ]	[ʊʊɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.12: Vowel set: New York City – rhotic: "∼" denotes an allophonic relationship

2.2.7.6 Philadelphia

To determine an appropriate vowel set for the Philadelphian accent, I examined Labov (2001), which describes the Philadelphia vowel system in the 1970s based on case studies. Since the Philadelphian accented data being used in this thesis are predominantly from Labov's natural misunderstandings corpus (Labov, 2010b)

which was collected mostly during the 1980s, the vowel system adopted for my vowel set is chosen to be the 1970s' system (as described by Labov (2001)) as opposed to a more contemporary system (Prichard and Tamminga, 2012). Labov (2001) primarily reported formant values without using an IPA representation; therefore, I consulted with a linguist who has worked on Philadelphian phonetics and phonology for further clarifications.[10] Table 2.13 summarises the vowel system. The second column contains a more narrow transcription based on the descriptions of Labov (2001). The third column contains a simplified version of the second column using a smaller set of IPA symbols. The final vowel set for the Philadelphian accent for this thesis is listed in the final column. Some vowels were not described by Labov (2001): they are the historical-/r/ vowels and their pre-vocalic equivalents, which I assume to be the same as those in the General American vowel set. A majority of the Philadelphian vowels are similar to those of General American, and the ones that are different are discussed below.

2.2.7.6.1 TRAP–BATH The short 'a' vowel in TRAP and BATH is reported to have a lexical split (Labov, 1989). The short 'a' vowel is frequently realised as [æə] before the following consonants: [m, n, f, θ] and [s], as well as in specific lexical items, such as *mad*, *glad* and *bad*; it is realised as [æ] elsewhere. In Labov's (1989) descriptions, the two realisations are tense and lax, which I indicated as being [æə] and [æ] respectively.

2.2.7.6.2 FACE FACE has two allophones. In the free environment (before no consonants), it is realised as [eɪ]. In the checked environment (before a consonant), it is realised like FLEECE but more retracted, and therefore a possible representation could be [i̟ɪ]. To avoid using diacritics for vowels, two alternative representations are [iɪ] which ignores the retraction, or [ɪ]/[ɪː] which assumes a lower vowel height. To

[10] I thank Dr. Kyle Gorman for his help with interpreting and translating Labov's (2001) descriptive summary of the vowel system using IPA, all errors remain mine.

decide between these two simplified representations, [iɪ] and [ɪɪ], we must consider also the realisation of the FLEECE vowel, which is [iː]. After applying the offglide treatment (see Section 2.2.7.1.2), FLEECE would be [ij], and similarly FACE could either be [ij] or [ɪj]. To avoid complete neutralisation with FLEECE, the checked vowel for FACE is best represented as [ɪɪ], and the offglide treatment as [ɪj].

2.2.7.6.3 GOAT The realisation of GOAT differs in three environments, *free* (not before a consonant), *checked* (before a consonant) and pre-lateral (before [l]). The starting height of the GOAT vowel is mid-high in the free and checked environments. In terms of frontness, the free vowel is as front as FACE in the free environment, making [ø] an appropriate representation of the starting point. The checked vowel is as front as FOOT, and therefore [ɵ] would be an appropriate representation. The ending-point of the free and checked vowels is assumed to be [ʊ]. In summary, the free vowel is realised as [øʊ], and the checked vowel is realised as [ɵʊ]. The pre-lateral vowel is realised as that of THOUGHT, therefore [ɔː]. Since neither [ø] nor [ɵ] have been used in other accents, they are both simplified to [ə], and GOAT is therefore realised as [əʊ] in both free and checked positions.

2.2.7.6.4 GOOSE The realisation of GOOSE differs in three environments, *free* (before no consonants), *checked* (before a consonant) and pre-lateral (before [l]). The starting-point of the GOOSE vowel is high in all the environments. In terms of frontness, the free vowel is as front as FACE in the free environment, making [y] an appropriate representation of the starting point; the checked vowel is as front as FOOT, and therefore [ʉ] would be an appropriate representation. The ending-point of the free and checked vowels is assumed to be [ʊ]. To recap, the free vowel is realised as [yʊ], and the checked vowel is realised as [ʉʊ]. The pre-lateral vowel is realised as [uː].

Since [y] has not been used in other accents, the free vowel will be merged with

the checked vowel so GOOSE is realised as [ʉʊ] in both free and checked positions.

2.2.7.6.5 PRICE Similar to Canadian raising, PRICE has two allophones. The vowel would be realised as [ɐɪ] before a voiceless consonant, and [aɪ] elsewhere. To avoid using a new symbol [ɐ], [ə] was used instead, and therefore [ɐɪ] was replaced by [əɪ].

2.2.7.6.6 Others The treatments of offglides, length marks and extrapolations of missing vowels are the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section 2.2.7.1.4.

2.2.7.7 Canada

I examined Wells (1982d, pp. 490–497) to determine an appropriate vowel set for the Canadian accent (not the Newfoundland variety). I identified the modifications I made to the phonemic vowel set by Wells (1982d, pp. 492–493). I utilised Wells's detailed discussion of the surface realisation of the accent and made modifications similar to those made to the other accents for consistency. The final vowel set for the Canadian accent for this thesis is listed in the final column of Table 2.14.

2.2.7.7.1 Canadian Raising The PRICE and MOUTH vowels are underlyingly /aɪ/ and /aʊ/, but they are realised as [əi] and [ʌʊ] respectively before a voiceless consonant. This is commonly known as Canadian Raising. PRICE has two realisations in the vowel set used for this thesis: [əi] before a voiceless consonant and [aɪ] elsewhere; similarly MOUTH has [ʌʊ] before a voiceless consonant and [aʊ] elsewhere.

2.2.7.7.2 THOUGHT–CLOTH–LOT–PALM–START Most of the Canadian accents have the same vowel for THOUGHT, CLOTH, LOT, PALM and START. Phonetically this vowel has the quality [ɑ]. It may be lightly rounded [ɒ], but not

Keyword	Narrow	Broad	This thesis
	Labov (2001) and Labov (1989)		
KIT	[ɪ]	[ɪ]	[ɪ]
DRESS	[ɛ]	[ɛ]	[ɛ]
TRAP	[æ] ∼ [æə]	[æ] ∼ [æə]	[æ] ∼ [æə]
LOT	[ɑː]	[ɑː]	[ɑɑ]
STRUT	[ʌ]	[ʌ]	[ʌ]
FOOT	[ʊ]	[ʊ]	[ʊ]
BATH	[æ] ∼ [æə]	[æ] ∼ [æə]	[æ] ∼ [æə]
CLOTH	[ɔː]	[ɔː]	[ɔɔ]
FLEECE	[iː]	[iː]	[ij]
FACE	[eɪ] ∼ [i̟ɪ]	[eɪ] ∼ [ɪː]	[ej] ∼ [ɪj]
PALM	[ɑː]	[ɑː]	[ɑɑ]
THOUGHT	[ɔː]	[ɔː]	[ɔɔ]
GOAT	[øʊ] ∼ [ɵʊ] ∼ [ɔː]	[əʊ] ∼ [ɔː]	[əw] ∼ [ɔɔ]
GOOSE	[yʊ] ∼ [ʉʊ] ∼ [uː]	[ʉʊ] ∼ [uː]	[ʉw] ∼ [uw]
PRICE	[aɪ] ∼ [ɐɪ]	[aɪ] ∼ [ɐɪ]	[aj] ∼ [əj]
CHOICE	[ɔɪ]	[ɔɪ]	[ɔj]
MOUTH	[aʊ]	[aʊ]	[aw]
HAPPY	[i]	[i]	[i]
COMMA	[ə]	[ə]	[ə]
NURSE	–	–	[ɜɜɹ]
NEAR	–	–	[ɪɹ]
SQUARE	–	–	[ɛɹ]
START	–	–	[ɑɑɹ]
NORTH	–	–	[ɔɔɹ]
FORCE	–	–	[ɔɔɹ]
CURE	–	–	[ʊɹ]
LETTER	–	–	[əɹ]
MIRROR (KIT)	–	–	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜɜɹ]
NEARER (NEAR)	–	–	[ɪɹ]
MARY (SQUARE)	–	–	[ɛɹ]
SAFARI (START)	–	–	[ɑɑɹ]
AURA (NORTH)	–	–	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔɔɹ]
CURIE (CURE)	–	–	[ʊɹ]
LETTERING (LETTER)	–	–	[əɹ]

Table 2.13: Vowel set: Philadelphia: "∼" denotes an allophonic relationship

in START, and long. I will ignore the lightly rounded variant in my vowel set, and instead adopt a long [ɑ], namely [ɑː] for the five reference vowels.

2.2.7.7.3 Others The treatments of offglides, length marks and extrapolations of missing vowels are the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section 2.2.7.1.4. It is worth noting that the treatment of offglides was applied to both of the realisations of the PRICE vowel, [əi] and [aɪ], despite their difference in the vowel quality of the ending vowel; concretely, they are treated as [əj] and [aj] respectively. For GOAT, the underlying monophthong /o/ is treated as being realised as a diphthong [oʊ], following the same argument in Section 2.2.7.2.1. To maintain consistency with GenAm, the following vowels are treated as being long: NURSE, NORTH, and FORCE, along with their pre-vocalic counterparts; similarly FLEECE and GOOSE are treated as being long following with offglides [j] and [w] respectively.

2.2.7.8 Australia

I examined Wells (1982d, pp. 592–605) to determine an appropriate vowel set for the Australian accent. I identified and discussed the modifications I made to the phonemic vowel set by Wells (1982d, p. 596). I utilised Wells's detailed discussion of the surface realisation of the accent and made modifications similar to those made to the other accents for consistency. The surface realisations are summarised in the third column in Table 2.15. The final vowel set for the Australian accent for this thesis is listed in the final column in Table 2.15.

The Australian accent can be divided into three main groups – *Cultivated*, *General* and *Broad* (Wells, 1982d, pp. 592–605). Phonetically they differ mainly in the quality of the closing diphthongs, namely, FLEECE, GOOSE, FACE, GOAT, PRICE and MOUTH. In addition to these three subgroups, Australian English has been found to have high variability in the realisations of the vowels, especially of the closing diphthongs (Wells, 1982d, p. 596) which poses a challenge for determining their underlying and surface

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 492–493)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/	–	[æ]	[æ]
LOT	/ɑ/	[ɑː]	[ɑː]	[ɑɑ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/æ/	–	[æ]	[æ]
CLOTH	/ɑ/	[ɑː]	[ɑː]	[ɑɑ]
FLEECE	/i/	–	[iː]	[ij]
FACE	/eɪ/	–	[eɪ]	[ej]
PALM	/ɑ/	[ɑː]	[ɑː]	[ɑɑ]
THOUGHT	/ɑ/	[ɑː]	[ɑː]	[ɑɑ]
GOAT	/o/	–	[oʊ]	[ow]
GOOSE	/u/	–	[uː]	[uw]
PRICE	/aɪ/	[aɪ] ∼ [əi]	[aɪ] ∼ [əi]	[aj] ∼ [əj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/ɑʊ/	[ɑʊ] ∼ [ʌʊ]	[ɑʊ] ∼ [ʌʊ]	[ɑw] ∼ [ʌw]
HAPPY	/i/	–	[i]	[i]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜɹ/	[ɜˤ]	[ɜˤ]	[ɜɜɹ]
NEAR	/ɪɹ/	–	[ɪɹ]	[ɪɹ]
SQUARE	/ɛɹ/	–	[ɛɹ]	[ɛɹ]
START	/ɑɹ/	[ɑ̟ːɹ], [a̠ːɹ]	[ɑːɹ]	[ɑɑɹ]
NORTH	/oɹ/	–	[oːɹ]	[ooɹ]
FORCE	/oɹ/	–	[oːɹ]	[ooɹ]
CURE	/ʊɹ/	–	[ʊɹ]	[ʊɹ]
LETTER	/əɹ/	[ɚˤ]	[ɚˤ]	[əɹ]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜːɹ]	[ɜɜɹ]
NEARER (NEAR)	–	–	[ɪɹ]	[ɪɹ]
MARY (SQUARE)	–	–	[ɛɹ]	[ɛɹ]
SAFARI (START)	–	–	[ɑːɹ]	[ɑɑɹ]
AURA (NORTH)	–	–	[oːɹ]	[ooɹ]
ORAL (FORCE)	–	–	[oːɹ]	[ooɹ]
CURIE (CURE)	–	–	[ʊɹ]	[ʊɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.14: Vowel set: Canada: "∼" denotes an allophonic relationship

forms. For simplicity, I will adopt only the General Australian accent group.

2.2.7.8.1 FLEECE and GOOSE FLEECE and GOOSE are phonemically /iː/ and /uː/ respectively. Phonetically, FLEECE and GOOSE share the same starting-point. FLEECE is realised as [i̟i] or in the notation of Wells (1982d, p. 534) [ɨi]; while GOOSE is realised as [i̟ɯ] or in the notation of Wells (1982d, p. 534) [ɨɯ]. GOOSE has another surface realisation [ʊ̟ʉ̟], but this is ignored for simplicity. To avoid the use of diacritics with vowels, [i̟] is substituted with a vowel of similar quality. Two possibilities for the substitution are [ɨ] and [ɪ]. [ɨ] differs from [i̟] in terms of its height – [i̟] is near-close, while [ɨ] is close. [ɪ] differs from [i̟] in terms of its backness – [i̟] is central while [ɪ] is near-front. [ɪ] was chosen over [ɨ] as a substitute starting vowel for FLEECE and GOOSE; this was motivated by the fact that [ɪ] was already used as the KIT vowel across many accents, while [ɨ] was not used at all, and therefore using the latter option would introduce sparsity issues in subsequent analyses. In summary, the vowels for FLEECE and GOOSE were simplified to [ɪi] and [ɪɯ] respectively. Since the phonemic representation of HAPPY was identical to that of FLEECE, the surface realisation of HAPPY was treated as that of FLEECE.

2.2.7.8.2 STRUT and START STRUT is realised phonetically as [a̲]. START has the same vowel quality as STRUT but it is phonetically long [a̲ː] (Wells, 1982d, p. 599). The quality of this vowel is retracted as indicated by the diacritic [x̲] (note that [x] is merely a placeholder for the diacritic). For simplicity, this diacritic is ignored for my vowel set, yielding [a] and [aː]. Since BATH and PALM have the same phonemic representation as START, the surface realisation is applied equally to all three vowels.

2.2.7.8.3 NEAR, SQUARE and CURE Phonemically, NEAR has the following representations: /ɪə/, /iːə/ and /iː/, and CURE has /ʊə/, /ɔː/, /uːə/ and /uː/.

NEARER has the realisation [ɪːɹ] and CURIE has two realisations, [ʊəɹ] and [ʊːɹ]. For simplicity, the first form is chosen to be the surface realisation, with [ɪə] for NEAR, [ɪːɹ] for NEARER, [ʊə] for CURE and [ʊəɹ] for CURIE. Similarly to NEARER, MARY (pre-vocalic SQUARE) has a longer monophthongal realisation [eːɹ].

2.2.7.8.4 Others The treatments of offglides, length marks and extrapolations of missing vowels are the same as in Section 2.2.7.1.2, Section 2.2.7.1.3 and Section 2.2.7.1.4. It is worth noting that the treatment of offglides was applied to the following vowels: FLEECE, FACE, GOAT, GOOSE, PRICE, CHOICE and MOUTH. This treatment of offglides effectively simplified the ending-point of these vowels; for instance, the ending points of FLEECE, FACE and PRICE are [i], [ɪ] and [ɪ̟] respectively, and the offglide treatment would convert them to [j]. Similarly, the ending-points of GOOSE, GOAT and MOUTH were [ɯ], [ʉ̞], and [o] respectively, and the offglide treatment would convert them to [w]. KIT and TRAP are both reported to be closer than those in RP, and are therefore [ɪ̟] and [æ̟] respectively, but these diacritics were ignored for simplicity.

2.2.7.9 Others

Six other accent groups were tabulated: New Zealand, South Africa, India, Caribbean, Ireland and Scotland. However, since together they cover only ≈ 60 data points which is $\approx 1\%$ of the corpus, I will only discuss them together briefly in this section.

2.2.7.9.1 New Zealand I examined Wells (1982d, pp. 605–610) in order to determine an appropriate vowel set for the New Zealand accent. Table 2.16 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 609) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 596–600)			
KIT	/ɪ/	[ɪ̞]	[ɪ̞]	[ɪ]
DRESS	/e/	[e]	[e]	[e]
TRAP	/æ/	[æ̞]	[æ̞]	[æ]
LOT	/ɒ/	–	[ɒ]	[ɒ]
STRUT	/ʌ/	[a̠]	[a̠]	[a]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/aː/	–	[a̠ː]	[aa]
CLOTH	/ɒ/	–	[ɒ]	[ɒ]
FLEECE	/iː/	[i̞i̞]	[i̞i̞]	[ɪj]
FACE	/ʌɪ/	[ʌɪ]	[ʌɪ]	[ʌj]
PALM	/aː/	[a̠ː]	[a̠ː]	[aa]
THOUGHT	/ɔː/	–	[ɔː]	[ɔɔ]
GOAT	/ʌʊ̈/	[ʌʉ̞]	[ʌʉ̞]	[ʌw]
GOOSE	/uː/	[ɨɯ], [ʊʉ̞]	[ɨɯ], [ʊʉ̞]	[ɪw]
PRICE	/ɑɪ/	[ɒɪ̞]	[ɒɪ̞]	[ɒj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/æʊ̈/	[æo]	[æo]	[æw]
HAPPY	/iː/	–	[i̞i̞]	[ɪj]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜː/	–	[ɜː]	[ɜɜ]
NEAR	/ɪə/, /iːə/, /iː/	–	[ɪə], [iːə], [iː]	[ɪə]
SQUARE	/eə/	–	[eə]	[eə]
START	/aː/	[a̠ː]	[a̠ː]	[aa]
NORTH	/ɔː/	–	[ɔː]	[ɔɔ]
FORCE	/ɔː/	–	[ɔː]	[ɔɔ]
CURE	/ʊə/, /ɔː/, /uːə/, /uː/	–	[ʊə], [ɔː], [uːə], [uː]	[ʊə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɪ]	[ɪɪ]
MERRY (DRESS)	–	–	[eɪ]	[eɪ]
CURRENT (NURSE)	–	–	[ɜːɪ]	[ɜɜɪ]
NEARER (NEAR)	–	[ɪːɪ]	[ɪːɪ]	[ɪɪɪ]
MARY (SQUARE)	–	[eːɪ]	[eːɪ]	[eeɪ]
SAFARI (START)	–	–	[aːɪ]	[aaɪ]
AURA (NORTH)	–	–	[ɔːɪ]	[ɔɔɪ]
ORAL (FORCE)	–	–	[ɔːɪ]	[ɔɔɪ]
CURIE (CURE)	–	[ʊəɪ], [ʊːɪ]	[ʊəɪ], [ʊːɪ]	[ʊəɪ]
LETTERING (LETTER)	–	–	[əɪ]	[əɪ]

Table 2.15: Vowel set: Australia

forms (cf. Section 2.2.7.1.4), as tabulated in the fourth column. Finally, I made similar modifications as those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

Wells (1982d, pp. 605–610) described the general New Zealand vowel system and a few variations in a broad New Zealand accent. Since these variations were limited to a few vowel types, namely FLEECE, GOOSE and NURSE, they were ignored and are therefore not reflected in the table.

The lateral /l/ has a substantial influence on any preceding vowels, in particular, GOAT and GOOSE. GOOSE is realised as [uː] before /l/ and as [ʉː] elsewhere. GOAT is realised as [ɒʊ] before /l/ and as [ʌʊ] elsewhere.

DRESS and TRAP are realised more closely than that of in RP, with [e] for DRESS and [ɛ] for TRAP. DRESS can be realised even closer than [e], yielding [ɪ]. KIT is realised more centrally and lowered [ɪ̞̈]. For simplicity, I assumed KIT has no such centralised realisation and remained [ɪ] so to maintain a contrast between DRESS and kit, I simplified the possible realisations of DRESS to only [e], and not [ɪ].

NURSE has the realisation [ɞ̞ː]. To avoid using diacritics, and the symbol [œ], I chose [ɜː] to be its substitute. In terms of openness and backness, [ɜ] is a good candidate, since the diacritics denote that [œ] should be centralised and lowered, although the rounded quality is not captured and is therefore ignored.

2.2.7.9.2 South Africa I examined Wells (1982d, pp. 610–622) to determine an appropriate vowel set for the South African accent. Table 2.17 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 616) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 605–610)			
KIT	/ə/	[i̞]	[i̞]	[ɪ]
DRESS	/e/	[e], [ɪ], [ɪə]	[e], [ɪ], [ɪə]	[e]
TRAP	/æ/	[ɛ]	[ɛ]	[ɛ]
LOT	/ɒ/	–	[ɒ]	[ɒ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/aː/	–	[aː]	[aa]
CLOTH	/ɒ/	–	[ɒ]	[ɒ]
FLEECE	/iː/		[iː]	[ij]
FACE	/ʌɪ/	–	[ʌɪ]	[ʌj]
PALM	/aː/	–	[aː]	[aa]
THOUGHT	/ɔː/	–	[ɔː]	[ɔɔ]
GOAT	/ʌʊ/	[ʌʊ̜] ∼ [ɒʊ̜]	[ʌʊ̜] ∼ [ɒʊ̜]	[ʌw] ∼ [ɒw]
GOOSE	/uː/	[ʉː], [iʉ] ∼ [uː]	[ʉː], [iʉ] ∼ [uː]	[ʉw] ∼ [uw]
PRICE	/ɑɪ/	–	[ɑɪ]	[ɑj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/æʊ/	–	[æʊ̜]	[æw]
HAPPY	/iː/	–	[iː]	[ij]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜː/	[ɵ̞ː]	[ɵ̞ː]	[ɜɜ]
NEAR	/iə/, /iːə/, /iː/, /eə/	[iə]	[iə]	[iə]
SQUARE	/eə/	[eə], [ɪə]	[eə], [ɪə]	[eə]
START	/aː/	–	[aː]	[aa]
NORTH	/ɔː/	–	[ɔː]	[ɔɔ]
FORCE	/ɔː/	–	[ɔː]	[ɔɔ]
CURE	/ʊə/, /uːə/, /uː/, /ɔː/	–	[ʊə], [uːə], [uː], [ɔː]	[ʊə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[əɹ]	[əɹ]
MERRY (DRESS)	–	–	[eɹ], [ɪɹ], [ɪəɹ]	[eɹ]
CURRENT (NURSE)	–	–	[ɵ̞ːɹ]	[ɜɜɹ]
NEARER (NEAR)	–	–	[iəɹ]	[iəɹ]
MARY (SQUARE)	–	–	[eəɹ], [ɪəɹ]	[eəɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[ɔːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔːɹ]	[ɔɔɹ]
CURIE (CURE)	–	–	[ʊəɹ], [uːəɹ], [uːɹ], [ɔːɹ]	[ʊəɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.16: Vowel set: New Zealand: "∼" denotes an allophonic relationship

modifications similar to those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this accent as adopted in this thesis is tabulated in the final column.

This accent has two sub-accents, broad and conservative, a distinction denoted by "|" in the table: the vowels on the left of the bar are broad and those on the right are conservative. For the transcriptions of the corpus, I will assume the conservative accent unless indicated otherwise by the reported transcriptions and demographics.

KIT has two sets of realisations. The first set is either [ɪ] or [i], when the vowel is in a stressed syllable and before or after a velar consonant (e.g. *kiss* and *lick*), or before /ʃ, tʃ, dʒ/ (e.g. *fish*, *ditch* and *bridge*) or after /h/ (e.g. *hit*). The second set is either [ɪ̈] or [ə] in the complementary set of environments in the first set. They are best treated as allophones of the same phoneme. [ɪ̈] maybe replaced by [ə] in stressed syllables, e.g. *dinner* has the realisation [ˈdənə], and therefore I simplified the second set to only [ə]. For simplicity, I chose only [ɪ] in the first set.

FLEECE and GOOSE are monophthongs, unlike accents which have developed diphthongization, such as Southern England, Australia and New Zealand, and I therefore did not apply the offglide treatment which converts the second element to [j] and [w] respectively.

2.2.7.9.3 Scotland I examined Wells (1982c, pp. 395–408) in order to determine an appropriate vowel set for the Scottish accent. Table 2.18 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982c, p. 399) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, as summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 612–616)			
KIT	/ɪ/, /ə/	[ɨ, ə] ∼ [ɪ, i]	[ɨ, ə] ∼ [ɪ, i]	[ə] ∼ [ɪ]
DRESS	/e/	[e] | [ɛ]	[e] | [ɛ]	[e] | [ɛ]
TRAP	/æ/	[ɛ] | [æ]	[ɛ] | [æ]	[ɛ] | [æ]
LOT	/ɒ/	–	[ɒ]	[ɒ]
STRUT	/ʌ/	–	[ʌ]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/ɑː/	[ɑ̟ː] | [ɑː]	[ɑ̟ː] | [ɑː]	[ɑɑ]
CLOTH	–	–	[ɒ]	[ɒ]
FLEECE	/iː/	–	[iː]	[ii]
FACE	/əɪ/	[ʌɪ], [ʌe] | [ëɪ], [əɪ]	[ʌɪ], [ʌe] | [ëɪ], [əɪ]	[ʌj] | [əj]
PALM	/ɑː/	[ɑ̟ː] | [ɑː]	[ɑ̟ː] | [ɑː]	[ɑɑ]
THOUGHT	/ɔː/	[oː]	[oː]	[oo]
GOAT	/əʊ/	[ʌʊ], [ʌː] | [əʊ]	[ʌʊ], [ʌː] | [əʊ]	[ʌw] | [əw]
GOOSE	/uː/	[ʉː]	[ʉː]	[ʉʉ]
PRICE	/aɪ/	[ɒɪ] | [a̲ɪ]	[ɒɪ] | [a̲ɪ]	[ɒj] | [aj]
CHOICE	/ɔɪ/	–	[oɪ]	[oj]
MOUTH	/aʊ/	[æʊ] | [a̲ʊ]	[æʊ] | [a̲ʊ]	[æw] | [aw]
HAPPY	–	[i̟], [i]	[i̟], [i]	[i]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ɜː/	[öː] | [ɜː]	[öː] | [ɜː]	[ɜɜ]
NEAR	/ɪə/	–	[ɪə]	[ɪə]
SQUARE	/eə/	[eː] | [ɛə]	[eː] | [ɛə]	[ee] | [ɛə]
START	/ɑː/	[ɑ̟ː] | [ɑː]	[ɑ̟ː] | [ɑː]	[ɑɑ]
NORTH	/ɔː/	[oː]	[oː]	[oo]
FORCE	/ɔː/	[oː]	[oː]	[oo]
CURE	/ʊə/	–	[ʊə]	[ʊə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɨɹ , əɹ] ∼ [ɪɹ , iɹ]	[əɹ] ∼ [ɪɹ]
MERRY (DRESS)	–	–	[eɹ] | [ɛɹ]	[eɹ] | [ɛɹ]
CURRENT (NURSE)	–	–	[öːɹ] | [ɜːɹ]	[ɜɜɹ]
NEARER (NEAR)	–	–	[ɪəɹ]	[ɪəɹ]
MARY (SQUARE)	–	–	[eːɹ] | [ɛəɹ]	[eeɹ] | [ɛəɹ]
SAFARI (START)	–	–	[ɑ̟ːɹ] | [ɑːɹ]	[ɑɑɹ]
AURA (NORTH)	–	–	[oːɹ]	[ooɹ]
ORAL (FORCE)	–	–	[oːɹ]	[ooɹ]
CURIE (CURE)	–	–	[ʊəɹ]	[ʊəɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.17: Vowel set: South African: "|" divides two sub-accents: the left is broad and the right is conservative; "∼" denotes an allophonic relationship.

Wells (1982c, pp. 395–408) described a typical Scottish accent while mentioning specific differences with accents in cities such as Glasgow, and with educated speech. The accent adopted for the vowel system is the typical Scottish accent.

In the adopted system, I assumed that there are two mergers LOT–THOUGHT ([ɔ]), and TRAP–PALM ([a]). The rule known as Aitken's Law was applied to all monophthongs except STRUT, KIT and [ə]; this means that all monophthongs are long if they are followed either by a word boundary, morpheme boundary, a voiced fricative, or /r/, and otherwise they are short. PRICE has two allophonic realisations, [ʌi] ∼ [ae]: it is realised as [ae] when it is in the same environment as Aitken's Law or if the vowel is in syllable-final positions within a word, and it is realised as [ʌi] elsewhere. The lengthened monophthongs [iː] and [ʉː] were assumed to have no diphthongisation, given that they were lengthened as per Aitken's Law, and therefore I did not apply the offglide treatment which would convert the second element to [j] and [w] respectively.

2.2.7.9.4 Ireland I examined Wells (1982c, pp. 417–428) in order to determine an appropriate vowel set for the Irish accent and Table 2.19 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982c, p. 419) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

Wells (1982c, pp. 417–428) described a typical Irish accent while also mentioning specific differences with the Dublin accent. The accent adopted for the vowel system is the typical Irish accent, and other Irish accents were ignored. STRUT and FOOT are reported to have free variation between opposition and neutralisation. For the

Keyword	Phonemic Wells (1982c, pp. 399–408)	Surface	Extrapolated	This thesis
KIT	/ɪ/	[ɪ]	[ɪ]	[ɪ]
DRESS	/ɛ/	[ɛ̞]	[ɛ̞]	[ɛ]
TRAP	/a/	[a]	[a]	[a]
LOT	/ɔ/	[ɔ]	[ɔ]	[ɔ]
STRUT	/ʌ/	[ʌ̟]	[ʌ̟]	[ʌ]
FOOT	/u/	[ʉ], [ʏ]	[ʉ], [ʏ]	[ʉ]
BATH	/a/	[a̠]	[a̠]	[a]
CLOTH	/ɔ/	[ɔ]	[ɔ]	[ɔ]
FLEECE	/i/	[i]	[i]	[i]
FACE	/e/	[e]	[e]	[e]
PALM	/a/	[a]	[a]	[a]
THOUGHT	/ɔ/	[ɔ]	[ɔ]	[ɔ]
GOAT	/o/	[o], [oʊ]	[o], [oʊ]	[o]
GOOSE	/u/	[ʉ], [ʏ]	[ʉ], [ʏ]	[ʉ]
PRICE	/ae/, /ʌi/	[ʌi] ∼ [ae]	[ʌi] ∼ [ae]	[ʌj] ∼ [aj]
CHOICE	/ɒɪ/	[ɒɪ], [ɔɪ]	[ɒɪ], [ɔɪ]	[ɒj]
MOUTH	/ʌu/	[ʌ̟u]	[ʌ̟u]	[u]
HAPPY	/e/, /ɪ/, /i/	–	[eː], [ɪ], [iː]	[ee]
COMMA	/ʌ/	–	[ʌ̟]	[ʌ]
NURSE	/ɜɹ/	–	[ɜːɹ]	[ɜɜɹ]
NEAR	/iɹ/	–	[iːɹ]	[iiɹ]
SQUARE	/eɹ/	–	[eːɹ]	[eeɹ]
START	/aɹ/	–	[aːɹ]	[aaɹ]
NORTH	/ɔɹ/	–	[ɔːɹ]	[ɔɔɹ]
FORCE	/oɹ/	–	[oːɹ], [oʊɹ]	[ooɹ]
CURE	/uɹ/	–	[ʉːɹ]	[ʉʉɹ]
LETTER	/əɹ/	–	[əɹ]	[əɹ]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛ̞ːɹ]	[ɛɛɹ]
CURRENT (NURSE)	–	–	[ɜːɹ]	[ɜɜɹ]
NEARER (NEAR)	–	–	[iːɹ]	[iiɹ]
MARY (SQUARE)	–	–	[eːɹ]	[eeɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[ɔːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[oːɹ], [oʊɹ]	[ooɹ]
CURIE (CURE)	–	–	[ʉːɹ]	[ʉʉɹ]
LETTERING (LETTER)	–	–	[əɹ]	[əɹ]

Table 2.18: Vowel set: Scotland: "∼" denotes an allophonic relationship; Aitken's law is applied to all monophthongs except STRUT, KIT and [ə].

vowel set, I assumed that this contrast is not neutralised. Of the many possible realisations of STRUT, [ɵ], [ɔ̈], [ɤ], [ə], [ə] was chosen because it was a symbol used by other vowel sets.

2.2.7.9.5 India I examined Wells (1982d, pp. 624–632) to determine an appropriate vowel set for the Indian accent. Table 2.20 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 626) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

The vowel system of Indian English is quite similar to that of RP but with the following differences. Firstly, the phonemic status is dubious for the two contrasts of /ʌ/ vs. /ə/ , and /ɒ/ vs. /ɔ/. Because these contrasts are dubious, for the variable realisation of STRUT, [ʌ] and [ə] are simplified to only [ʌ]. Similarly I assume /ɔ/ does not exist this system, which means both CLOTH and THOUGHT are phonemically /ɒ/, and NORTH is phonemically /ɒɹ/.

Secondly, the NURSE vowel is not phonemically distinct and it could pair with /əɹ/ or /ʌɹ/, this was simplified to only /əɹ/. Thirdly, length distinctions are not always consistent; for example, it is not clear if CLOTH and THOUGHT are different in length or not. I decided to apply the same length distinctions from RP to this accent, which allowed me to determine whether a vowel is long or not for cases when their surface realisations were not described in Wells (1982d, pp. 624–632); for example, GOOSE, START, NURSE, THOUGHT and NORTH are treated as long, while CLOTH is treated as short. Fourthly, FACE and GOAT are monophthongs. Finally, strong vowels are commonly used in weak syllables; for instance, *different* [ˈdɪfɹɛnt] with [ɛ]

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982c, pp. 417–428)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/	[a]	[a]	[a]
LOT	/ɒ/	[ɑ], [ɑ̟]	[ɑ], [ɑ̟]	[ɑ]
STRUT	/ʌ/	[ɵ], [ɔ̈], [ɤ], [ə]	[ɵ], [ɔ̈], [ɤ], [ə]	[ə]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/æ/,/aː/	[æ], [aː]	[æ], [aː]	[æ]
CLOTH	/ɒ/,/ɔː/	[ɑ], [ɑː]	[ɑ], [ɑː]	[ɑ]
FLEECE	/iː/	[iː], [ɪi]	[iː], [ɪi]	[ij]
FACE	/eː/	[eː]	[eː]	[ee]
PALM	/aː/	[aː]	[aː]	[aa]
THOUGHT	/ɔː/	[ɑː]	[ɑː]	[ɑɑ]
GOAT	/oː/	[oː]	[oː]	[oo]
GOOSE	/uː/	[uː], [ʊu]	[uː], [ʊu]	[uw]
PRICE	/aɪ/	[əɪ]	[əɪ]	[əj]
CHOICE	/ɔɪ/	[əɪ]	[əɪ]	[əj]
MOUTH	/aʊ/	–	[aʊ]	[aw]
HAPPY	/iː/	[i],[ɪ̟]	[i],[ɪ̟]	[i]
COMMA	/ə/	–	[ə]	[ə]
NURSE	/ʌɹ/,/ɛɹ/	–	[ʌɹ], [ɛɹ]	[ʌɹ]
NEAR	/iːɹ/	–	[iːɹ]	[ijɹ]
SQUARE	/eːɹ/	–	[eːɹ]	[eeɹ]
START	/aːɹ/	–	[aːɹ]	[aaɹ]
NORTH	/ɔːɹ/	–	[ɔːɹ]	[ɔɔɹ]
FORCE	/oːɹ/	–	[oːɹ]	[ooɹ]
CURE	/uːɹ/	–	[uːɹ]	[uwɹ]
LETTER	/əɹ/	–	[əɹ]	[əɹ]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɵːɹ], [ɔ̈ ːɹ], [ɤːɹ], [əːɹ]	[əəɹ]
NEARER (NEAR)	–	–	[iːɹ]	[ijɹ]
MARY (SQUARE)	–	–	[eːɹ]	[eeɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[ɔːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[oːɹ]	[ooɹ]
CURIE (CURE)	–	–	[uːɹ]	[uwɹ]
LETTERING (LETTER)	–	–	[əɹh]	[əɹ]

Table 2.19: Vowel set: Ireland

instead of [ə].

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 624-632)			
KIT	/ɪ/	[ɪ]	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/æ/	–	[æ]	[æ]
LOT	/ɒ/	[ɒ]	[ɒ]	[ɒ]
STRUT	–	[ʌ], [ə]	[ʌ], [ə]	[ʌ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/a/, /æ/	–	[a], [æ]	[a]
CLOTH	/ɒ/, /ɔ/	–	[ɒ], [ɔ]	[ɒ]
FLEECE	/i/	[iː]	[iː]	[ij]
FACE	/e/	[eː]	[eː]	[ee]
PALM	/a/	–	[a]	[a]
THOUGHT	/ɔ/, /ɒ/	–	[ɔː], [ɒː]	[ɒɒ]
GOAT	/o/	[oː]	[oː]	[oo]
GOOSE	/u/	–	[uː]	[uw]
PRICE	/aɪ/	[ʌɪ]	[ʌɪ]	[ʌj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/aʊ/	[ʌʊ]	[ʌʊ]	[ʌw]
HAPPY	/ɪ/, /i/	[iː]	[iː]	[ij]
COMMA	/a/, /ə/	–	[a], [ə]	[a], [ə]
NURSE	/əɹ/, /ʌɹ/	–	[əːɹ], [ʌːɹ]	[əəɹ]
NEAR	/ɪə/	–	[ɪə]	[ɪə]
SQUARE	/eə/	[ɛː]	[ɛː]	[ɛɛ]
START	/ɑ/	–	[ɑː]	[ɑɑ]
NORTH	/ɒ/, /ɔ/	–	[ɒː], [ɔː]	[ɒɒ]
FORCE	/o/	[oː]	[oː]	[oo]
CURE	/ʊə/	–	[ʊə]	[ʊə]
LETTER	/ə/	–	[ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[əːɹ], [ʌːɹ]	[əəɹ]
NEARER (NEAR)	/ɪəɹ/	[iːɹ]	[iːɹ]	[ijɹ]
MARY (SQUARE)	/eəɹ/	[eːɹ]	[eːɹ]	[eeɹ]
SAFARI (START)	/ɑɹ/	–	[ɑːɹ]	[ɑɑɹ]
AURA (NORTH)	/ɒɹ/, /ɔɹ/	–	[ɒːɹ], [ɔːɹ]	[ɒɒɹ]
ORAL (FORCE)	/oɹ/	–	[oːɹ]	[ooɹ]
CURIE (CURE)	/ʊəɹ/	–	[ʊəɹ]	[ʊəɹ]
LETTERING (LETTER)	/əɹ/	–	[əɹ]	[əɹ]

Table 2.20: Vowel set: India

2.2.7.9.6 Caribbean There are five regional accents in the corpus for the Caribbean group: Jamaica, Trinidad, Guyana, Barbados and the Leewards. Some of these accents have two sub-accents, acrolectal and basilectal, and this distinction is denoted by "|" in the tables; the vowels on the left of the bar are basilectal and those on the right are acrolectal. I will assume the acrolectal accent for the transcriptions of the corpus, unless indicated otherwise by the reported transcriptions and demographics.

2.2.7.9.6.1 Caribbean – Jamaica I examined Wells (1982d, pp. 574–577) in order to determine an appropriate vowel set for the Jamaican accent. Table 2.21 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 576) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

Rhoticity status within the accent is complex, and varies according to the sub-accents. Wells (1982d, pp. 574–577) discussed three sub-accents, acrolectal, mesolectal, and basilectal. For simplicity, I will ignore the mesolectal accent. In the basilectal accent, the /r/ in /r/-vowels is realised phonetically in NEAR, SQUARE, START, NORTH, FORCE and CURE, but not in LETTER, i.e. except in weak syllables. Furthermore, this is conditioned by its following environment. /r/ is not realised before a consonant in the same morpheme, e.g. it is not realised in *beard*, but it is realised in *near*. This allophonic relationship is denoted with "∼" in the table, such as [iɛɹ] ∼ [iɛ] for NEAR and SQUARE, [aaɹ] ∼ [aa] for START and NORTH, and [uoɹ] ∼ [uo] for FORCE and CURE. For simplicity in the final vowel set, any vowels containing [ɐ] were not chosen, and [ɔ̈] was simplified as [ɔ] with the diacritic removed.

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 574–577)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/a/	–	[a]	[a]
LOT	/a/ | /ɒ/	[a] | [ɒ]	[a] | [ɒ]	[a] | [ɒ]
STRUT	/ʌ/	[ɔ̈]	[ɔ̈]	[ɔ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/aː/	–	[aː]	[aa]
CLOTH	/aː/ | /ɔː/	–	[aː] | [ɔː]	[aa] | [ɔɔ]
FLEECE	/iː/	–	[iː]	[ij]
FACE	/eː/	[iɛ] | [eː]	[iɛ] | [eː]	[iɛ] | [ee]
PALM	/aː/	–	[aː]	[aa]
THOUGHT	/aː/ | /ɔː/	–	[aː] | [ɔː]	[aa] | [ɔɔ]
GOAT	/oː/	[uo] | [oː]	[uo] | [oː]	[uo] | [oo]
GOOSE	/uː/	–	[uː]	[uw]
PRICE	/aɪ/	[aɪ]	[aɪ]	[aj]
CHOICE	/aɪ/ | /ɔɪ/	–	[aɪ] | [ɔɪ]	[aj] | [ɔj]
MOUTH	/ɔʊ/	–	[ɔʊ̈]	[ɔw]
HAPPY	–	[ɪ], [i]	[ɪ], [i]	[ɪ], [i]
COMMA	/a/ | /ə/	[ɐ] | –	[ɐ] | [ə]	[ə]
NURSE	/ʌ/ | /ɜːɹ/	[ɔ̈] | [ɜˠː]	[ɔ̈] | [ɜˠː]	[ɔ] | [ɜɜɹ]
NEAR	/eːɹ/	[iɛɹ] ∼ [iɛ] | [eːɹ]	[iɛɹ] ∼ [iɛ] | [eːɹ]	[iɛɹ] ∼ [iɛ] | [eeɹ]
SQUARE	/eːɹ/	[iɛɹ] ∼ [iɛ] | [eːɹ]	[iɛɹ] ∼ [iɛ] | [eːɹ]	[iɛɹ] ∼ [iɛ] | [eeɹ]
START	/aːɹ/, /aː/	[aːɹ] ∼ [aː] | –	[aːɹ] ∼ [aː] | [aːɹ]	[aaɹ] ∼ [aa] | [aaɹ]
NORTH	/aːɹ/, /aː/ | /ɔːɹ/, /ɔː/	[aːɹ] ∼ [aː] | –	[aːɹ] ∼ [aː] | [ɔːɹ], [ɔː]	[aaɹ] ∼ [aa] | [ɔɔɹ]
FORCE	/oːɹ/	[uoɹ] ∼ [uo] | –	[uoɹ] ∼ [uo] | [uoɹ]	[uoɹ] ∼ [uo] | [uoɹ]
CURE	/oːɹ/	[uoɹ] ∼ [uo] | –	[uoɹ] ∼ [uo] | [uoɹ]	[uoɹ] ∼ [uo] | [uoɹ]
LETTER	/a/ | /ə/	[ɐ] | [ə]	[ɐ] | [ə]	[ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɔ̈ɹ] | [ɜˠː]	[ɔɹ] | [ɜɜɹ]
NEARER (NEAR)	–	–	[iɛɹ] | [eːɹ]	[iɛɹ] | [eeɹ]
MARY (SQUARE)	–	–	[iɛɹ] | [eːɹ]	[iɛɹ] | [eeɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[aːɹ] | [ɔːɹ]	[aaɹ] | [ɔɔɹ]
ORAL (FORCE)	–	–	[uoɹ]	[uoɹ]
CURIE (CURE)	–	–	[uoɹ]	[uoɹ]
LETTERING (LETTER)	–	–	[ɐɹ] | [əɹ]	[əɹ]

Table 2.21: Vowel set: Caribbean – Jamaica: "|" divides two sub-accents: the left is basilectal and the right is acrolectal; "∼" denotes an allophonic relationship.

2.2.7.9.6.2 Caribbean – Trinidad I examined Wells (1982d, pp. 577–580)in order to determine an appropriate vowel set for the Trinidadian accent. Table 2.22 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 580) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface

and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

The accent is non-rhotic. For simplicity, any vowels containing [ɐ] were not chosen for the final vowel set, and [ɔ̈] was simplified as [ɔ] with the diacritic removed and similarly [a̲] was simplified as [a].

2.2.7.9.6.3 Caribbean – Guyana I examined Wells (1982d, pp. 581–583) to determine an appropriate vowel set for the Guyanese accent. Table 2.23 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 582) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

2.2.7.9.6.4 Caribbean – Barbados I examined Wells (1982d, pp. 583–585) to determine an appropriate vowel set for the Barbadian accent. Table 2.24 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 585) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to the other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 577–580)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/a/	–	[a̱]	[a]
LOT	/ɒ/	[ɔ̈]	[ɔ̈]	[ɔ]
STRUT	/ɒ/ | /ʌ/	[ɔ̈] | [ə]	[ɔ̈] | [ə]	[ɔ] | [ə]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/a/ | /ɑ/	–	[a̱] | [ɑ]	[a] | [ɑ]
CLOTH	/ɔ/ | /ɒ/	[ɔː] | [ɔ̈ː]	[ɔː] | [ɔ̈ː]	[ɔɔ]
FLEECE	/i/	–	[i]	[i]
FACE	/e/	–	[e]	[e]
PALM	/a/ | /ɑ/	[a̱] | [ɑ]	[a̱] | [ɑ]	[a] | [ɑ]
THOUGHT	/ɒ/ | /ɔ/	[ɔ̈ː] | [ɔː]	[ɔ̈ː] | [ɔː]	[ɔɔ]
GOAT	/o/	–	[o]	[o]
GOOSE	/u/	–	[u]	[u]
PRICE	/aɪ/	–	[aɪ]	[aj]
CHOICE	/ɔɪ/	–	[ɔɪ]	[ɔj]
MOUTH	/ɔʊ/	–	[ɔʊ]	[ɔw]
HAPPY	/i/	–	[i]	[i]
COMMA	/a/ | /ə/	–	[a̱] | [ə]	[a] | [ə]
NURSE	/ɒ/ | /ɜ/	[ɔ̈] | [ɜ]	[ɔ̈] | [ɜ]	[ɔ] | [ɜ]
NEAR	/eə/	[ia]	[ia]	[ia]
SQUARE	/eə/	[ia]	[ia]	[ia]
START	/a/ | /ɑ/	[a̱] | [ɑ]	[a̱] | [ɑ]	[a] | [ɑ]
NORTH	/ɒ/ | /ɔ/	[ɔ̈ː] | [ɔː]	[ɔ̈ː] | [ɔː]	[ɔɔ]
FORCE	/ɒ/ | /ɔ/	[ɔ̈ː] | [ɔː]	[ɔ̈ː] | [ɔː]	[ɔɔ]
CURE	/ɒ/ | /ɔ/	–	[ɔ̈ː] | [ɔː]	[ɔɔ]
LETTER	/a/ | /ə/	–	[a̱] | [ə]	[a] | [ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɔ̈ɹ] | [ɜɹ]	[ɔɹ] | [ɜɹ]
NEARER (NEAR)	–	–	[iaɹ]	[iaɹ]
MARY (SQUARE)	–	–	[iaɹ]	[iaɹ]
SAFARI (START)	–	–	[a̱ɹ] | [ɑɹ]	[aɹ] | [ɑɹ]
AURA (NORTH)	–	–	[ɔ̈ːɹ] | [ɔːɹ]	[ɔɔɹ]
ORAL (FORCE)	–	–	[ɔ̈ːɹ] | [ɔːɹ]	[ɔɔɹ]
CURIE (CURE)	–	–	[ɔ̈ː] | [ɔːɹ]	[ɔɔɹ]
LETTERING (LETTER)	–	–	[a̱ɹ] | [əɹ]	[aɹ] | [əɹ]

Table 2.22: Vowel set: Caribbean – Trinidad: "|" divides two sub-accents: the left is basilectal and the right is acrolectal.

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 581–583)			
KIT	/ɪ/	–	[ɪ]	[ɪ]
DRESS	/ɛ/	–	[ɛ]	[ɛ]
TRAP	/a/	–	[a]	[a]
LOT	/a/ | /ɑ/	–	[a] | [ɑ]	[a] | [ɑ]
STRUT	/ʌ/	[ɔ]	[ɔ]	[ɔ]
FOOT	/ʊ/	–	[ʊ]	[ʊ]
BATH	/aː/	–	[aː]	[aa]
CLOTH	/aː/ | /ɔː/	–	[aː] | [ɔː]	[aa] | [ɔɔ]
FLEECE	/iː/	–	[iː]	[ij]
FACE	/eː/	–	[eː]	[ee]
PALM	/aː/	–	[aː]	[aa]
THOUGHT	/aː/ | /ɑː/	–	[aː] | [ɑː]	[aa] | [ɑɑ]
GOAT	/oː/	–	[oː]	[oo]
GOOSE	/uː/	–	[uː]	[uw]
PRICE	/aɪ/	–	[aɪ]	[aj]
CHOICE	/aɪ/ | /ɔɪ/	–	[aɪ] | [ɔɪ]	[aj] | [ɔj]
MOUTH	/ɔʊ/	–	[ɔʊ]	[ɔw]
HAPPY	–	[i]	[i]	[i]
COMMA	/a/ | /ə/	–	[a] | [ə]	[a] | [ə]
NURSE	/ʌ/ | /ʌɹ/	[ɔ] | [ɔɹ]	[ɔ] | [ɔɹ]	[ɔ] | [ɔɹ]
NEAR	/eə/	–	[eə]	[eə]
SQUARE	/eə/	–	[eə]	[eə]
START	/a/ | /ɑ/	–	[a] | [ɑ]	[a] | [ɑ]
NORTH	/ɒ/ | /ɔ/	–	[ɒ] | [ɔ]	[ɒ] | [ɔ]
FORCE	/ɒ/ | /ɔ/	–	[ɒ] | [ɔ]	[ɒ] | [ɔ]
CURE	/ɒ/ | /ɔ/	–	[ɒ] | [ɔ]	[ɒ] | [ɔ]
LETTER	/a/ | /ə/	–	[a] | [ə]	[a] | [ə]
MIRROR (KIT)	–	–	[ɪɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[ɛɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɔɹ]	[ɔɹ]
NEARER (NEAR)	–	–	[eəɹ]	[eəɹ]
MARY (SQUARE)	–	–	[eəɹ]	[eəɹ]
SAFARI (START)	–	–	[aɹ] | [ɑɹ]	[aɹ] | [ɑɹ]
AURA (NORTH)	–	–	[ɒɹ] | [ɔɹ]	[ɒɹ] | [ɔɹ]
ORAL (FORCE)	–	–	[ɒɹ] | [ɔɹ]	[ɒɹ] | [ɔɹ]
CURIE (CURE)	–	–	[ɒɹ] | [ɔɹ]	[ɒɹ] | [ɔɹ]
LETTERING (LETTER)	–	–	[aɹ] | [əɹ]	[aɹ] | [əɹ]

Table 2.23: Vowel set: Caribbean – Guyana: "|" divides two sub-accents: the left is basilectal and the right is acrolectal

final column.

Multiple modifications were made to simplify the final vowel set. STRUT and the starting-point of the PRICE and MOUTH vowels have three realisations [ʌ], [ɔ̈], [ə]; however, only the [ʌ] realisation is chosen. The diacritics for lowered and raised vowels were removed, such as [ɪ̞] in KIT and MIRROR, [e̞] in DRESS and MERRY, [e̞ː] and [e̞ə] in FACE, and [o̞ː] in GOAT, and [ʊ̞] in FOOT.

2.2.7.9.6.5 Caribbean – The Leewards In order to determine an appropriate vowel set for the Leewards accent, I examined Wells (1982d, pp. 585–588) which focussed on the speech of Montserrat. Table 2.25 contains a summary of the vowel set for this accent. I tabulated the phonemic vowel set by Wells (1982d, p. 588) in the second column. I utilised Wells's detailed discussion of the surface realisation of the accent, which is summarised in the third column of the table. I completed the vowel set by extrapolation from the surface and phonemic forms (cf. Section 2.2.7.1.4), which is tabulated in the fourth column. Finally, I made modifications similar to those made to other accents for consistency (cf. Section 2.2.7.1.2 and Section 2.2.7.1.3). The final vowel set for this thesis is tabulated in the final column.

In open syllables, the realisations of historical long vowels are shortened, e.g. *tea* is [ti] but BEAT is [biːt]. This allophonic relationship is denoted with "∼" in the table. Furthermore, [ɔ̈] and [ï] are simplified as [ɔ] and [i] respectively with the diacritics removed.

2.3 Written English corpus

This section describes a corpus that I constructed to serve as a "control" for subsequent analyses of the misperception corpus and various psycholinguistic norms can be derived from such a corpus.

While speech corpora are a valuable resource for linguists and speech engineers,

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 583–585)			
KIT	/ɪ/	[ɪ̟]	[ɪ̟]	[ɪ]
DRESS	/ɛ/	[e̞]	[e̞]	[ɛ]
TRAP	/a/	–	[a]	[a]
LOT	/ɒ/	[ɒ], [ɑ]	[ɒ], [ɑ]	[ɒ], [ɑ]
STRUT	/ʌ/	[ʌ], [ɜ̈], [ə]	[ʌ], [ɜ̈], [ə]	[ʌ]
FOOT	/ʊ/	[ʊ̟]	[ʊ̟]	[ʊ]
BATH	/aː/	–	[aː]	[aa]
CLOTH	/ɒː/	[ɒː], [ɑː]	[ɒː], [ɑː]	[ɒɒ], [ɑɑ]
FLEECE	/iː/	–	[iː]	[ij]
FACE	/eː/	[e̞ː], [e̞ə]	[e̞ː], [e̞ə]	[ee], [eə]
PALM	/aː/	–	[aː]	[aa]
THOUGHT	/ɒː/	–	[ɒː]	[ɒɒ]
GOAT	/oː/	[o̞ː], [oə]	[o̞ː], [oə]	[oo], [oə]
GOOSE	/uː/	[ʉ̟ː]	[ʉ̟ː]	[uw]
PRICE	/ʌi/	[ʌi], [ɜ̈i], [əi]	[ʌi], [ɜ̈i], [əi]	[ʌj]
CHOICE	/ʌi/ | /oi/	–	[ʌi] | [oi]	[ʌj] | [oj]
MOUTH	/ʌʊ/	[ʌʉ̟], [ɜ̈ʉ̟], [əʉ̟]	[ʌʉ̟], [ɜ̈ʉ̟], [əʉ̟]	[ʌw]
HAPPY	/iː/	–	[iː]	[ij]
COMMA	/ə/	–	[ə]	[ə]
NURSE	–	[ɜ˞ː]	[ɜ˞ː]	[ɜɜɹ]
NEAR	/eːɹ/	–	[eːɹ]	[eeɹ]
SQUARE	/eːɹ/	–	[eːɹ]	[eeɹ]
START	/aːɹ/	–	[aːɹ]	[aaɹ]
NORTH	/ɒːɹ/	–	[ɒːɹ]	[ɒɑɹ]
FORCE	/oːɹ/	–	[oːɹ]	[ooɹ]
CURE	/oːɹ/	–	[oːɹ]	[ooɹ]
LETTER	/əɹ/	[ɚ]	[ɚ]	[əɹ]
MIRROR (KIT)	–	–	[ɪ̟ɹ]	[ɪɹ]
MERRY (DRESS)	–	–	[e̞ɹ]	[ɛɹ]
CURRENT (NURSE)	–	–	[ɜ˞ː]	[ɜɜɹ]
NEARER (NEAR)	–	–	[eːɹ]	[eeɹ]
MARY (SQUARE)	–	–	[eːɹ]	[eeɹ]
SAFARI (START)	–	–	[aːɹ]	[aaɹ]
AURA (NORTH)	–	–	[ɒːɹ]	[ɒɑɹ]
ORAL (FORCE)	–	–	[oːɹ]	[ooɹ]
CURIE (CURE)	–	–	[oːɹ]	[ooɹ]
LETTERING (LETTER)	–	–	[ɚ]	[əɹ]

Table 2.24: Vowel set: Caribbean – Barbados: "|" divides two sub-accents: the left is basilectal and the right is acrolectal.

Keyword	Phonemic	Surface	Extrapolated	This thesis
	Wells (1982d, pp. 585–588)			
KIT	/i/	[i]	[i]	[i]
DRESS	/e/	[e]	[e]	[e]
TRAP	/a/	–	[a]	[a]
LOT	/ɒ/ | /a/	–	[ɒ] | [a]	[ɒ] | [a]
STRUT	/o/	[ɔ̈]	[ɔ̈]	[ɔ]
FOOT	/u/	[u]	[u]	[u]
BATH	/aː/	[aː] ∼ [a]	[aː] ∼ [a]	[aa] ∼ [a]
CLOTH	/aː/ | /ɒː/	[aː] ∼ [a] | [ɒː] ∼ [ɒ]	[aː] ∼ [a] | [ɒː] ∼ [ɒ]	[aa] ∼ [a] | [ɒɒ] ∼ [ɒ]
FLEECE	/iː/	[iː] ∼ [i]	[iː] ∼ [i]	[ij] ∼ [i]
FACE	/ie/ | /eː/	[ie] | [eː] ∼ [e]	[ie] | [eː] ∼ [e]	[ie] | [ee] ∼ [e]
PALM	–	–	[aː] ∼ [a]	[aa] ∼ [a]
THOUGHT	–	–	[aː] ∼ [a] | [ɒː] ∼ [ɒ]	[aa] ∼ [a] | [ɒɒ] ∼ [ɒ]
GOAT	/uo/ | /oː/	[uɔ̈] | [oː] ∼ [o]	[uɔ̈] | [oː] ∼ [o]	[uɔ] | [oo] ∼ [o]
GOOSE	/uː/	[uː] ∼ [u]	[uː] ∼ [u]	[uu] ∼ [u]
PRICE	/ai/ | /ɒi/	–	[ai] | [ɒi]	[aj] | [ɒj]
CHOICE	–	–	[ai] | [ɒi]	[aj] | [ɒj]
MOUTH	/ou/	–	[ou]	[ow]
HAPPY	–	[i]	[i]	[i]
COMMA	/u/	[u] | [ə]	[u] | [ə]	[u] | [ə]
NURSE	/oɪ / | /ɜː/	[ɔ̈ː], [ɔ̈˞ː] | [ɔ̈]	[ɔ̈ː], [ɔ̈ ˞ː] | [ɔ̈]	[ɔɔ] | [ɔ]
NEAR	–	[ïa] | [iɐ]	[ïa] | [iɐ]	[ia]
SQUARE	–	[ïa] | [iɐ]	[ïa] | [iɐ]	[ia]
START	/aː/	–	[aː]	[aa]
NORTH	–	[aː] ∼ [a] | [ɒː] ∼ [ɒ]	[aː] ∼ [a] | [ɒː] ∼ [ɒ]	[aa] ∼ [a] | [ɒɒ] ∼ [ɒ]
FORCE	–	[uo], [ʊʌ] | [oɔ̈]	[uo], [ʊʌ] | [oɔ̈]	[uo], [ʊʌ] | [ɔɔ]
CURE	–	[uo], [ʊʌ]	[uo], [ʊʌ]	[uo], [ʊʌ]
LETTER	–	–	[u] | [ə]	[u] | [ə]
MIRROR (KIT)	–	–	[iɪ]	[iɪ]
MERRY (DRESS)	–	–	[eɪ]	[eɪ]
CURRENT (NURSE)	–	–	[ïɔ̈ɪ] | [ɪɔ̈]	[ɔɔɪ] | [ɔɪ]
NEARER (NEAR)	–	–	[ïaɪ] | [iɐɪ]	[iaɪ]
MARY (SQUARE)	–	–	[ïaɪ] | [iɐɪ]	[iaɪ]
SAFARI (START)	–	–	[aːɪ]	[aaɪ]
AURA (NORTH)	–	–	[aːɪ] | [ɒːɪ]	[aaɪ] | [ɒɒɪ]
ORAL (FORCE)	–	–	[uoɪ], [ʊʌɪ] | [oɔ̈ɪ]	[uoɪ], [ʊʌɪ] | [ɔɔɪ]
CURIE (CURE)	–	–	[uoɪ], [ʊʌɪ]	[uoɪ], [ʊʌɪ]
LETTERING (LETTER)	–	–	[uɪ] | [əɪ]	[uɪ] | [əɪ]

Table 2.25: Vowel set: Caribbean – The Leewards: "|" divides two sub-accents: the left is basilectal and the right is acrolectal; "∼" denotes an allophonic relationship.

they are often small (< five million words) and frequently concentrate on a narrow speech type such as telephone exchanges in the SWITCHBOARD corpus (Calhoun et al., 2010) and prompted speech as in the TIMIT corpus (Garofolo et al., 1993), and are therefore unlikely to be representative of everyday speech. In order to obtain reliable word frequency norms, a corpus needs to be of 16 – 30 million words (Brysbaert and New, 2009). Together, the small corpus size and narrow speech types make the existing spoken corpora unsuitable for extracting a wide range of psycholinguistic norms.

Given that the existing speech corpora are not suitable, I chose to use corpora that have been compiled from TV and film subtitle texts. This method of using film subtitles to construct language corpora, SUBTLEX, was developed by New et al., 2007 for French, and subsequently used for English (Brysbaert and New, 2009)[11], Dutch (Keuleers, Brysbaert, and New, 2010), Polish (Mandera et al., 2014), Greek (Dimitropoulou et al., 2009), Brazilian Portuguese (Tang, 2012) and many other languages. Crucially lexical frequencies derived from these SUBTLEX corpora have been proven to be excellent predictors of behavioural task measures such as reaction times in written lexical decision tasks and consistently outperformed other larger written corpora or smaller spoken corpora, primarily due to their size (typically at least 20 million words, to half a billion words) and spoken register, as they are essentially transcribed spoken speech. Recent work by Ernestus and Cutler (2014) has shown that SUBTLEX is also an excellent predictor of reaction time in auditory lexical decision tasks. This finding directly supports the use of SUBTLEX for speech perception data. Together, the use of SUBTLEX for extracting psycholinguistic norms (beyond token frequencies) for this thesis is justified by the aforementioned validation studies.

[11]The reasons for not using Brysbaert and New (2009) as the control corpus are that Brysbaert and New (2009) is a frequency list and not a raw corpus which is not useful for creating a language model and furthermore the new corpus is approximately ten times bigger and thus more representative.

In the following sections, I will outline the compilation and processing of the corpus, from orthographic transcriptions to phonetic transcriptions.

2.3.1 Source

The corpus was compiled using subtitles from `http://opensubtitles.org` that are tagged as being English. The files were obtained by Paweł Mandera from the Center for Reading Research, Ghent University, Belgium. Time-stamps and other non-linguistic information were removed. Crucially, these files have been deduplicated, meaning near/exact duplicates of the same TV episode or film have been removed (Tang, 2012)[12] leaving 69,382 files.

2.3.2 Processing

Although the downloaded subtitle files were self-identified as English, errors were often made by the uploaders who sometimes included the original subtitles of the films with the translated versions in the zip package. To remove any non-English subtitle files, Shuyo's (2010) language detection library for Java, *langdetect*, was applied. The model calculates language probabilities from features of spelling using a naïve Bayesian model with character n-gram, using language profiles generated from Wikipedia abstracts. It has an above-99% precision for 53 languages. Only the files with a probability of 99% of being English were kept. This filtered out 60 files, leaving 69,322 files.

All 69,322 files underwent multiple steps of text normalisation as described in Section 2.1.2.1, regarding the treatments of upper/lowercases, punctuation marks, full stops, apostrophes, hyphens, and abbreviations. Two files were removed after the normalisation process due to corrupted formatting, leaving 69,320 files.

Since subtitle documents can contain large portions of text in languages other

[12]I profoundly thank Paweł Mandera for sharing these deduplicated files with me.

than English, (for instance, sung portions of musicals are often not translated), additional cleaning was required. These mixed-language files can be undetected by the language detection model above because *langdetect* only tests random samples from each piece of text. Following the footsteps of SUBTLEX–PL (Mandera et al., 2014), preliminary word frequencies were first calculated for all of the documents, and a file was removed if the 30 most frequent word types (from the preliminary word frequencies) did not cover at least 30% of the total tokens in the file. The 30% threshold was determined by manual inspections. I found that files with a threshold below 30 tended to contain more foreign proper names (e.g. names and places), a large portion of foreign text (e.g. sung speech in a foreign language not translated into English), or simply many typos. This threshold filter removed 3181 files, approximately 4.6% of the total, leaving 66,139 files. The final corpus has 353.4 million word tokens, and 755,559 word types.

2.3.2.1 Transcription

eSpeak (`http://espeak.sourceforge.net/`), a text-to-speech software, was used to transcribe the text-normalised corpus into IPA. The transcription system uses a combination of pronunciation rules and dictionaries. In order to determine the reliability of this transcription system, Marian et al. (2012) compared the transcription of eSpeak to the Carnegie Mellon Pronouncing Dictionary (Weide, 2014), and found that English eSpeak transcriptions correlates strongly with the CMU database with $R = 0.97$ ($N = 26{,}474$, $p < 0.001$). This strong and significant correlation suggests that the eSpeak system is reliable. I used the American English transcription setting, *en-us*, in this software.

First, all word types were extracted from the 66,139 files. Second, the word types were then converted into pronunciation using eSpeak. Third, although the system's output is IPA transcriptions, the sets of IPA symbols it provides were then

normalised to the General American system described in Section 2.2.7.2. Fourth, given that eSpeak can transcribe intervocalic /t/ (in the environment of tapping) as tap but not /d/, I therefore applied the tapping rules as described in Section 2.2.2 to transcribe the intervocalic /d/s as taps as well as intervocalic /t/s and /d/s as taps across word boundaries. Finally, aspiration of the voiceless stops was transcribed as described in Section 2.2.2.

2.4 Phonetic alignment

The key aspect of analysing mishearing data is to identify the differences between the intended speech and perceived speech. However, two important difficulties are the size of the misheard sequence and the identification of the differences between the intended and the perceived. Such alignments fall under the umbrella term, *Pairwise String Alignment*.

Slips containing only one 'error' - an insertion, deletion or substitution - can be analysed rather straightforwardly, e.g. 'thug' → 'hug', [θʌg] → [hʌg]; the change is [θ] → [h]. However, it is not always simple to align slips with multiple errors, e.g. 'sleeping bag' → 'single man', [sliːpɪŋ bæg] → [sŋgəl mæn]. The number of possible alignments between two sequences (say both have a length of L ≈ 10) are $\frac{2^{2L}}{\sqrt{\pi L}}$ (Durbin et al., 1998), which are ≈ 200,000. It is therefore clear that the complexity of the analysis increases with the number of errors, and thus a manual alignment by visually identifying changes is unfeasible. Furthermore, manually aligning mishearing data would be a subjective process and the quality would vary depending on the analyst and his/her judgement. So there is an evident need for a computational method that is objective and automatic.

2.4.1 A review of alignment algorithms

There are broadly two approaches to aligning phonetic sequences – phonetically based and phonetically blind. On the one hand, considering that the one of the research aims is to find segmental changes and discover how phonetic similarity motivates these changes, a phonetically blind alignment algorithm may be more suitable for the current analyses as it is more likely to avoid any potential issues of circularity. On the other hand, if the algorithm is phonetically blind the quality of alignments might be poor. For example a consonant-consonant substitution would be penalized as much as a consonant-vowel one. In the following sections, I will review the two existing approaches of alignments, starting with the phonetically based approach.

2.4.1.1 Phonetically based algorithms

In phonetically based alignments, it is necessary to establish the phonetic similarity between phones, for example, [p] is phonetically more similar to [t] than to [l]. This relies on the assumption that similar sounds are more likely to correspond to each other, and therefore more likely to be aligned with each other. In linguistics, phonetically based alignments have been developed for aligning pairs of cognates, bilingual texts, speech misproduction data and more. One example is an algorithm called *ALINE* (Kondrak, 2003), which uses phonetic similarity to improve the alignment accuracy of cognates by defining multi-valued articulatory phonetic features such as voice, lateral, place and nasal. However, there is no widely accepted procedure to determine phonetic similarity (Laver, 1994, p.391), and taking the case of ALINE (Kondrak, 2003), the weight of the phonetic features are free parameters so it is not clear what would be a principled way of determining the values of these weights. Much of the time, linguists rely on intuitions and some system of phonetic/phonological features. The following is an overview of various phonetically based alignment methods used by previous speech misperception work.

2.4.1.1.1 Manual alignment Bird's (1998) procedure was to first align syllables between the intended and perceived utterances based on stress, and then do the segmental alignments. The criteria for the segmental alignments were based on the author's intuition. In one particular example, the utterance "It's the milkman" was misperceived as "It's a nightmare.". The author pointed out that one could argue that [l] aligns with [t] with [k] being deleted. However, in the author's own words, "this goes against observation and logic", because [l] is more often deleted and [k] and [t] differ only in their place of articulation. The author's alignment procedure therefore relies on a combination of linguistic knowledge and arbitrary judgement, e.g. [l] is more often deleted (in a coda) and the number of phonological feature differences between segments. Three drawbacks with Bird's (1998) approach are apparent. Firstly, it is a manual process, which is not feasible for large data sets; secondly, since the stress pattern was aligned first, this indirectly biases the rate of stress errors over segmental errors, thus resulting in fewer stress errors; thirdly, the reliance on intuition is open to a wide range of problems, such as reliability and inter-coder agreement.

2.4.1.1.2 Semi-automated alignment Browman's (1980) alignment procedure was in the same vein as that of Bird (1998) and involved an initial manual alignment of syllables which aimed to maximize similarity using the judgement of the author. After syllable alignment, a specialised algorithm was devised to automate the segmental alignments. The algorithm follows three principles which were applied in the order of their importance (most important to the least important) – identity, common feature maximization and segmental order information; in addition, it is able to detect segmental metatheses. For detailed descriptions of the algorithm, see Browman (1980).

While Browman's (1980) approach has better automaticity than Bird (1998), it does have one major drawback in that it explicitly uses phonological features in the

alignment process which arguably results in more alignment biases (although these biases would be highly systematic), favouring the phonological features chosen.

Inspired by Bird's (1998) and Browman's (1980) methods, Tang and Nevins (2014) devised a method that employs an automated algorithm like that of Browman (1980), but does not rely on a particular phonological feature set, instead relying on linguists' intuitions like that of Bird (1998). Tang's method adapted the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), which is commonly used to align DNA sequences in molecular biology.

The Needleman-Wunsch algorithm employs a dynamic programming technique, which solves problems through the use of previously computed solutions on sub-problems. It has multiple free parameters: the cost of matches, mismatches (substitutions) and gaps (insertions and deletions). See Figure 2.14 for details. More

$$cell(i,j) = max \left\{ \begin{array}{c} cell(i-1,j-1) + S(X_i, Y_j) \\ cell(i-1,j) + g \\ cell(i,j-1) + g \end{array} \right\}$$

Figure 2.14: Needleman-Wunsch algorithm: the $S(X_i, Y_j)$ is the score obtained from the substitution matrix for two segments corresponding to the column and row of cell(i,j), g is the gap penalty.

specifically, the Needleman-Wunsch algorithm involves a gap penalty which can be constant, linear and affine. Different gap penalty schemes require different modifications to the algorithm. The affine gap penalty scheme uses two parameters, the cost of a gap opening and the cost for a gap extension which is a function of the gap length l. These are combined using the equation $GapPenalty = g_{open} + lg_{extend}$ or $GapPenalty = g_{open} + (l-1)g_{extend}$ (Eidhammer, Jonassen, and Taylor, 2004); this scheme can favour big gaps over many smaller gaps of the equivalent size and vice-versa. This equation was chosen for this study because it could be beneficial for capturing errors that involve whole-word deletions rather than just simple isolated segmental deletions.

Motivated by how Browman (1980) manually aligned the syllables between the intended and perceived utterances and the assumption that syllables are most likely to be preserved in misperception, the compromise arrived at was to use an algorithm which is phonetically blind in the sense of distinctive features but is sensitive to syllables, i.e. the alignment is biased towards aligning by syllables. A simple implementation of this with the Needleman-Wunsch affine algorithm was devised. Firstly, phonetic blindness was achieved by inputting an identity matrix for the substitution matrix that the algorithm requires (see Durbin et al. (1998) for details of substitution matrices). Secondly, the alignment by syllable was done by simply replacing all the vowels with the same segment "V" to represent the nucleus of a syllable, as this would therefore bias the alignment algorithm to align by syllables and act as a soft constraint. Stress was ignored in the alignment process, unlike the method of Bird (1998).

Of the four parameters (Match, Mismatch, Gap opening, and Gap extension), the match cost was fixed with the value 1 to minimize the complexity of the optimization; this is also a default value for the match cost in most substitution matrices, and therefore only three parameters remain. Manual parameter optimization would be challenging so the computational Monte Carlo method (Metropolis and Ulam, 1949) was employed for finding a suitable set of parameters. The training data was 10% of the corpus which was manually aligned by the authors. Half of the training data was for calibration and the other half for validation. X (the number of generated sets of parameters) and the upper and lower limits of each parameter were systematically increased until a 100% match rate was achieved.

A remaining and recurring criticism[13] is that this procedure is only *partially* phonetically blind since manual alignment can implicitly introduce phonetic biases, just like that of Bird (1998), but this was not a major drawback in the analyses

[13]I thank Dr. Katrin Skoruppa for pointing this weakness out.

done by Tang and Nevins (2014) as they only considered unambiguously aligned errors. That is, the immediately adjacent segments of a mismatch pair of segments are matched. For instance, in [kɪt] and [kæt] [k] and [t] are identical in both of the intended and perceived utterances. In any case, restricting analyses to only unambiguously aligned errors would mean losing data that involved "ambiguous" alignments and the restriction itself is an assumption and any findings can only be generalised to these restricted cases[14]. Furthermore, any semi-automated procedures requiring manual alignments are not only time-consuming and but are also prone to errors and biases.

2.4.1.2 Phonetically blind algorithm

A phonetically blind alignment algorithm should not have been exposed to prior knowledge of linguistic information. For instance, it should not know anything about phonological features and it should not be able to distinguish between a consonant and a vowel. One way of achieving this would be to use a dynamic alignment algorithm, such as the Needleman-Wunsch algorithm, with arbitrarily chosen fixed weights for match, mismatch, and gaps. However, this may not be satisfactory as it could result in unwanted alignments, such that it would freely align a vowel with a consonant. This leads us to add another criterion for our ideal alignment method: the algorithm needs to be able to learn linguistic information from the data itself – for example a vowel is more similar to another vowel than to a consonant. Adding this to our previously listed criteria, an ideal alignment method needs to be a) objective, thus it would be independent of analysts' performance, b) automatic, and c) to be able to learn linguistic information without supervision. Looking at all of these criteria, it is clear that we need an *unsupervised* alignment algorithm that is able to *learn* linguistic information from the alignments themselves. Two recent studies,

[14]I thank Dr. Mark Huckvale for pointing out this issue.

Hirjee and Brown (2010) and Wieling, Prokić, and Nerbonne (2009), employed such a method for linguistic data. Their alignment procedures are summarised in the section below.

2.4.1.2.1 PMI-based Needleman-Wunsch algorithm (Hirjee and Brown, 2010) Hirjee and Brown (2010) applied an iterative alignment method to aligning misheard lyrics. The data in their study are closely related to the data in this thesis, in that they are using misheard data from a different domain, making their alignment method particularly suitable for our analyses of mishearing in spoken speech.

Hirjee and Brown (2010) adapted an optimal global alignment algorithm from Durbin et al. (1998) without specifying the name of the algorithm. It is reasonable to assume that it is the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) as it is a global alignment algorithm documented in Durbin et al. (1998). The algorithm requires a substitution matrix. This matrix encodes the likelihood of a particular pair of segments co-occurring (i.e. aligned), and it is calculated using log-odds scores (Henikoff and Henikoff, 1992). This log-odds metric is in fact identical to *pointwise mutual information* (PMI) (Church and Hanks, 1990).

$$PMI(x, y) = log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Figure 2.15: Pointwise Mutual Information

Where:

- $p(x, y)$ is the probability of aligning segment x with segment y. It is estimated by calculating the number of observations that segment x aligns with segment y, divided by the total number of aligned segments.
- $p(x)$ and $p(y)$ are estimated by calculating the number of observations of segment x (and segment y) divided by the total number of segments.

Informally, PMI compares the probability of observing segment x with segment y with the probability of observing these two segments independently (i.e. chance). A positive PMI indicates the pair of segments is more likely to co-occur, while a negative PMI indicates the pair of segments is more likely to co-occur by chance. The PMI values of all of the possible segment pairs are used as the substitution matrix for the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970).

The overall procedure by Hirjee and Brown (2010) is outlined below:

1. Calculate initial confusion matrix: For a given pair of utterances, firstly the segmental length of the two utterances are matched by adding as many segments "-" (which represent insertions or deletions) as needed; secondly the segments are aligned sequentially starting from the left for all the pairs of utterances. Finally, a confusion matrix is obtained by counting the number of observations for given pairs of segments.

2. Create the PMI substitution matrix: the confusion matrix from the previous step is used to calculate the PMI substitution matrix using Equation 2.15.

3. Create new alignments: New alignments are obtained using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) with the PMI substitution matrix calculated in the previous step.

4. Recalculate the confusion matrix: With the new alignments obtained in the previous step, a new confusion matrix can then be calculated.

5. Repeat Steps 2, 3 and 4, until there is no change in the alignments (i.e. convergence is reached).

Validation: Hirjee and Brown's (2010) focus was on improving the retrieval of lyrics with misheard lyrics queries in search engines, e.g. a user entered the search terms "kiss this guy" (instead of the correct lyrics, "kiss the sky") to search for

the Jimi Hendrix's song, "Purple Haze". The validation was done by assessing the performance of the authors' method at finding the best matches for a query set of misheard lyrics in a collection of full song lyrics containing the correct version of each query. They compared four other methods based on simple edit distance (sensitive to match and mismatch segments), phoneme edit distance (sensitive to phonological features), a syllable alignment pattern searching method (sensitive to syllables), and a probabilistic syllable alignment method but trained on the misheard lyrics just like the previously described method. The authors' method outperformed all of the other four methods and had the highest retrieval accuracy.

2.4.1.2.2 PMI-based Levenshtein distance (Wieling, Prokić, and Nerbonne, 2009) Wieling, Prokić, and Nerbonne (2009) developed a similar PMI-based method which was applied to phonetic dialect data. Their method has three major differences to Hirjee and Brown (2010). Firstly, it employs a different alignment algorithm. Secondly, it differs in how the initial confusion matrix is obtained. Thirdly, it differs in how the PMI values are calculated and transformed. These three differences are described below.

The first difference from Hirjee and Brown (2010) is the choice of the alignment algorithm. The alignment algorithm chosen by Wieling, Prokić, and Nerbonne (2009) was the Levenshtein distance (Levenshtein, 1966). The use of this algorithm for measuring linguistic distances is well-motivated and has been used to successfully measure linguistic distances for multiple languages, such as Irish (Kessler, 1995), Dutch and Norwegian (Heeringa, 2004).The regular Levenshtein distance has a simple cost scheme: mismatch (substitution, insertion or deletion) has a cost of 1, and match has a cost of 0. This cost scheme, which has no linguistic knowledge, can lead to alignments of vowels with consonants, so this could lead to unwanted alignments. To tackle this, Wieling, Prokić, and Nerbonne (2009) adapted the regular Levenshtein distance which disallowed alignments between vowels and consonants –

henceforth, the VC-sensitive Levenshtein algorithm.

The second difference is how the initial confusion matrix is obtained. Wieling, Prokić, and Nerbonne (2009) obtained the initial confusion matrix by applying the VC-sensitive Levenshtein algorithm with the default cost of 1 for a mismatch and 0 for a match.

The third difference is how the PMI values are calculated and transformed. Since the Levenshtein distance takes a distance matrix (as opposed to a substitution matrix), the PMI values were scaled in the range of 0 to 1 by subtracting each PMI value from 0 and adding the maximum PMI value. Later work by Wieling and Nerbonne (2011) used an improved version of this method to measure the linguistic distances of five languages (Bantu, Bulgarian, German, Dutch and Norwegian); the improvement was achieved by ignoring identical sound segment pairs in calculating the PMI values. The intuition behind this decision is that identical pairs should have a distance of 0 and the interest lies in the distance of non-identical sound segment pairs relative to each other. This also means that when converting PMI values to distances, the normalization should be the range of a small value to 1; this is to ensure that only identical segments have a distance of 0. This study also mentioned additional details on how to tackle sparse matrices; that is, when a segment pair does not occur ($p(x, y) = 0$). The solution was to add a small value to $p(x, y)$, $p(x)$ and $p(y)$ in Equation 2.15.

The overall procedure by Wieling and Nerbonne (2011) is outlined below:

1. Calculate initial confusion matrix: the VC-sensitive Levenshtein algorithm with a regular cost matrix is used to align pairs of utterances. A confusion matrix is then obtained by counting the number of observations for a given pairs of segments.

2. Create the PMI values: the confusion matrix from the previous step is used to calculate the PMI values using Equation 2.15 with the modification of ignoring

identical sound segment pairs.

3. Convert PMI values to distances: the distance for identical segments is set to 0. For non-identical segments, the PMI values are converted to distances by subtracting each PMI value from 0, and normalizing the PMI scores to range between a small value and 1.

4. Create new alignments: New alignments are obtained using the VC-sensitive Levenshtein algorithm with the distance matrix calculated in the previous step.

5. Recalculate the confusion matrix: With the new alignments obtained in the previous step, a new confusion matrix can then be calculated.

6. Repeat Steps 2, 3, 4 and 5, until there is no change in the alignments (i.e. convergence is reached).

Validation: Even without any further modification to how the PMI values are calculated, Wieling, Prokić, and Nerbonne (2009) already found that the method can achieve superior alignments. They evaluated its performance against other methods (such as the Hamming algorithm, the VC-sensitive Levenshtein algorithm without PMI and Pair Hidden Markov Model) using manually aligned Bulgarian dialect data set which are considered to be the gold standard alignments (because it is manually aligned by linguists) and found this PMI-based VC-sensitive Levenshtein distance method was able to achieve the lowest alignment (word pairs) error rate (4.50%) as well as being computationally efficient compared to other methods. Besides validating the method with manually-aligned dialect data, Wieling, Margaretha, and Nerbonne (2012) found that the PMI-based VC-sensitive Levenshtein distance method was able to induce phonetic distances for Dutch and German vowels and these alignment induced distances correlated highly with acoustic distances. Finally, in a large accent (nativeness) rating task with over 1,000 native American English listeners,

Wieling et al. (2014) found that the phonetic distance induced by the same method strongly correlates with subjective accent judgements. Overall, this method has been used successfully to capture a) linguistic knowledge of dialectologists who manually aligned dialect data sets (Wieling, Prokić, and Nerbonne, 2009); b) acoustic similarity of vowels (Wieling, Margaretha, and Nerbonne, 2012), and c) perceptual similarity of sentences (Wieling et al., 2014).

2.4.2 Algorithm selection and adaptation

Having reviewed the multiple alignment methods used in linguistics, I will now discuss which alignment method should be selected for analysing the mishearing data in this thesis and what (if any) adaptations should be made.

The alignment methods reviewed in the previous section are summarised in Table 2.26. It is immediately clear the manual and semi automatic methods would not be ideal, given the large amount of data used in this thesis, their bias towards subjective linguistic knowledge, and that they have not been validated. The ideal candidates are the methods used by Hirjee and Brown (2010) and Wieling and Nerbonne (2011), as they are fully automatic, with no or minimal biases (in the case of Wieling and Nerbonne (2011), consonants and vowels are not allowed to be aligned with each other), and the linguistic knowledge was induced through iterative learning, that was free from subjective biases by the authors or phonological features. Most importantly, they have both been validated. To break the tie between these two PMI-based methods, we need to look into the quality of the validations. As summarised in the previous section, Wieling and Nerbonne's (2011) method was extensively validated in multiple studies (Wieling et al., 2014; Wieling, Margaretha, and Nerbonne, 2012; Wieling and Nerbonne, 2011; Wieling, Prokić, and Nerbonne, 2009), and was shown to be appropriate for capturing linguistic knowledge when performing manual alignments and also acoustic and perceptual similarities, while the scope of Hirjee and

Brown's (2010) validation was comparably narrow and showed an improvement in lyrics retrieval in search engines. Together, this led to the conclusion that Wieling and Nerbonne's (2011) method was the best candidate.

One modification was made to Wieling and Nerbonne's (2011) method for this thesis. The hard constraint that disallows consonants and vowels to be aligned with each other, while it is a reasonably sensible constraint, can be overly harsh. This is evidenced by the fact that in the original study that used the method in Wieling, Prokić, and Nerbonne (2009), the authors found that the errors made by the alignment method were caused by an inability to align vowels with consonants (and presumably vice-versa). The modification would therefore be to change the hard constraint to a soft constraint. Concretely, when obtaining the initial confusion matrix, the cost matrix for the Levenshtein algorithm was set to have a high cost (for instance, 100) for aligning consonants with vowels, and in subsequent iterations the cost matrix is free to be updated and this constraint can therefore be overridden. Furthermore, since the glides [j] and [w] are used as offglides (see Section 2.2.7.1.2), which means some of glides are consonantal and some are vocalic (in the sense that they form part of a vowel), the initial soft constraint did not apply to them, meaning they were allowed to align with both consonants and vowels in the first iteration.

This modified version of Wieling and Nerbonne's (2011) method, the PMI-based VC-sensitive Levenshtein distance method, is therefore chosen for use throughout this thesis for any analyses that require segmental alignments.

2.4.3 Minimal alignment unit

In pairwise alignments, one must define the alignment unit. But this decision is not a trival one; as discussed below, it can affect the scope of analyses and allow/restrict certain kinds of misperception.

Previous work on misperception (experimental and naturalistic) mostly aligned

Source	Automatic	Algorithm	Bias	Linguistic Knowledge	Validated
Bird (1998)	No	–	Stress-pattern	Authors	No
Browman (1980)	Semi	Tailor-made	Syllable	Authors, Phonological features	No
Tang and Nevins (2014)	Semi	Needleman-Wunch	Syllable	Authors	No
Hirjee and Brown (2010)	Yes	Needleman-Wunch	–	Inductive	Yes
Wieling and Nerbonne (2011)	Yes	Levenshtein	Consonants and Vowels (Hard)	Inductive	Yes

Table 2.26: An overview of alignment methods

their data on a phoneme level (Miller and Nicely, 1955; Wang and Bilger, 1973; Cutler and Butterfield, 1992; Browman, 1980; Bond, 1999; Labov, 2010b). One should question whether phonemes are the correct units of alignment. The implication of using phonemes as the units for alignment is that the phonemes that are "complex" (occupy more than one timing slot, e.g. long vowels, diphthongs and affricates) cannot be misperceived into multiple "simple" phonemes (occupy only one timing slot). Phonotactics can often tell us whether these complex phonemes are divisble or not (e.g.Wells (1990) treated affricates as indivisible using arguments from phonotactics), but this is not the case in perception. For example, it is possible that the affricate [tʃ] is misperceived as two simple phones, [t] and [s], where the first segment of the affricate was perceived correctly as [t] while the second segment was misperceived as [s]. Similarly with rhotic vowels (see Section 2.2.7.2.2), the rhoticity can be misperceived as another sonorant. In sum, I regard phonemes to be too large a unit for alignment purposes; this would impose too big a restriction on the possible kinds of misperception.

This thesis will use the IPA segments in Section 2.2.2.2 as the minimal alignment units. With regard to the long vowels and rhotic vowels, the length and the rhoticity are encoded as a separate segment. Crucially, the length marks are encoded as a copy of the previous segment (see Section 2.2.7.1.3). The diphthongs are divisible,

meaning that the offglides can freely align. Similarly, affricates are divisible. In order to analyse these complex phones as a unit after the alignment process, the aligned segments can be re-parsed to regroup the sub-segments (from complex phones) back to being a single unit, e.g. the segments [t] and [ʃ] (from [tʃ]) were aligned with [t] and [s] respectively, and the regrouping will convert [t] [ʃ] back to [tʃ]. This would mean being able to detect the kind of misperception mentioned above ([tʃ] into [t] and [s]) without losing the possibility of analysing the interactions between complex and simple phones.

Finally, syllable structure (onsets, nuclei and codas), syllable breaks and stress marks are not part of the alignment units. Instead they will be re-associated with the corresponding segments after the alignment process. Alternatively, this information could in fact be tagged as part of the segments. For instance, stress marks could be encoded as part of the syllable by attaching stress to all of the segments (assuming syllables are stress-bearing units) or just the nuclei or the rhyme within the syllables prior to the alignment process. Similarly, syllabic structures can also be encoded in this manner, such that every segment is tagged with a syllabic label. Syllabic breaks can be encoded in multiple ways, for example using a three label system – pre-break, post-break, and non-adjacent. This tagging method was not used, as the chosen alignment method in Section 2.4.2 relies on the alignment units to be non-sparse due to its use of alignments as a way of calculating perceptual distances which in turn is used to improve the quality of the alignments, and this tagging method will undoubtedly introduce a lot of alignment units that are of very low frequency and therefore causing data-sparsity problems.

2.5 Contributions

This section will summarise the contribution of this chapter to the thesis as a whole and the potential contributions to the field of linguistics.

Section 2.1 documented the naturalistic English misperception corpora and the steps that were taken to compile them into a single mega corpus. This resulted in a mega corpus of naturalistic English misperception, taking into account the variability between each subcorpus (e.g. missing demographics, storage formats and others). This is on par with the Fromkin's Speech Error Database (Fromkin, 2000) that contains about 9,000 instances of speech errors for selected languages, and which was developed by combining multiple speech error corpora from independent researchers. The need to compile such a mega corpus for speech errors and the issues involved were discussed by Schütze and Ferreira (2007), and many of the arguments are directly applicable to our speech misperception data. In sum, this chapter documented a similar undertaking for naturalistic misperception of English, containing about 5,000 instances, thus making our present mega corpus one of a kind. In the future, the aim would be to make this corpus accessible to other researchers via a web platform as in the case with the Fromkin's Speech Error Database.

Section 2.2 documented the phonetic transcriptions of the mega corpus. In Section 2.2.7, I systematically examined the vowel sets of 14 English dialects in order to perform dialectal transcriptions. For each dialect, I examined the phonemic vowel set and the surface vowel set as reported by the existing literature. Missing surface vowels were extrapolated, and then finally the complete surface vowel set was simplified and normalised to avoid introducing IPA symbols that are less widely used across the dialects. The contribution to the thesis of mediating the data-sparsity issue with the simplification of the surface vowels is that one can then analyse the mega corpus as a whole without the need to subset the corpus by dialect, therefore allowing the use of analytical techniques that require large samples. As a more gen-

eral contribution, the thorough examinations of the dialects would serve as a starting point for any cross-dialectal comparison of vowel sets, and furthermore the vowel sets themselves could be used to aid the development of text-to-speech synthesis of the dialects examined; for instance, a text-to-speech synthesis system of New York City accented English.

In Section 2.2.4, I devised a novel data-driven method for selecting "legal" onsets for a syllabification method using the Maximal Onset Principle. For practical purposes, this method allowed me to proceed to conduct analyses on the misperception corpora, involving questions concerning sub-syllabic units (onsets, nucleus and codas), without the need to make arbitrary decisions on whether certain marginal onsets (e.g. [vl]) are legal onsets or not. Another advantage of this method is that it is data-driven, which means it is relatively theory-independent[15]. Admittedly, the proposed method is rudimentary with room for further development. For example, in its present form the method does not take into account of any prosodic information, such as stress, which is argued to be a determining factor for the syllabification process (McCarthy, 1979). The suggested method needs to be further examined for compatibility with the existing phonological/speech processing theories, as well as its applicability to other languages.

Section 2.4 reviewed some of the alignment algorithms that were used by previous studies of speech misperception corpora, and more generally on phonetic transcription. I examined the pros and cons of each of the algorithms. With a focus on their appropriateness for the mishearing data in this thesis, I selected a data-driven alignment algorithm with a few minor modifications. This contributes to the thesis by providing a means of aligning the misperception data with minimal human/linguistic biases, while achieving sensible alignments. More generally, this section serves as a

[15]Although a more theory-independent method would be to employ a kind of entropy-based syllabification, which is not appropriate for our multi-dialectal corpus, as discussed in Section 2.2.4.

detailed review of alignment algorithms for speech perception data and argues for the merits of the chosen data-driven approach for aligning phonetic transcriptions.

Chapter 3

Bottom-up phonetic and phonological factors

3.1 Introduction

The focus of this chapter is to examine the bottom-up phonetic and phonological factors that play a role in naturalistic misperception. Previous analyses of naturalistic misperception using the sub-corpora of the our combined mega corpus (Browman, 1978; Bird, 1998; Bond, 1999; Labov, 2010b; Tang and Nevins, 2014) have identified numerous phonetic and phonological factors, as described in Chapter 1, Section 1.2. However, none of these studies have attempted to quantify the amount of phonetic biases. To best quantify the bias, the most direct method is to analyse the misperception data at the lowest levels, namely segmental confusions; this is essentially a confusion matrix of segments. The question is whether any phonetic/phonological factors can be found at such a low level. Three approaches were devised to tackle this question.

3.1.1 Phonetic biases

The first approach is to examine the rate of confusions and separate them into the most basic phonetic dimensions, features. The confusion rates of consonants will be computed separately for place, manner and voicing – namely, place confusions, manner confusions and voicing confusions. The rates of vowels will be computed separately for height and backness – namely, height confusions and backness confusions. This will allow us to examine if there are any phonetic/phonological trends on a featural level, and whether these trends could be explained using phonetic/phonological theories.

We will then focus the level of analysis on a segmental level. The overall approach is to compare the segmental relationships in the naturalistic matrices with those that are phonetically/phonologically based. This will allow us to quantify how much of the segmental relationships in the naturalistic matrices can be explained by pure phonetic/phonological factors.

Concretely, we will convert the segmental confusion matrices into distance matrices which contain the pairwise distance between any two segments (e.g. [p] and [b], [t] and [b], etc.), separately for consonants and for vowels. The confusion matrices will first be converted into similarity matrices, which will then be converted into distance matrices. It is important to note that the distance conversion is perceptually grounded by employing the distance metric by Shepard (1972) and Shepard (1987). This metric captures the fact that perceptual distance has an exponential relationship with similarity. Other distance conversions, such as taking the inverse of similarity, are inadequate for perception data.

Similarly, the phonetic/phonological distance matrices will be computed using acoustic measurements of vowels and feature values of consonants. The naturalistic distance matrices and the phonetic/phonological distance matrices will then be compared in two ways. First, correlation tests between two sets of distance matrices

will be employed, which reflect the global similarity. Second, the distance matrices will be projected into a hierarchical structure for consonants and a two-dimensional space for the vowels. These projected structures will be then compared visually and quantitatively using correlation tests.

3.1.2 Ecological validity

The second approach is to compare the segmental relationships in the naturalistic matrices with those in experimentally induced misperception matrices. The experimental matrices will be selected from studies that focused on perception of the lowest level, without much top-down effects from the lexicon, by using stimuli that are either a CV or VC syllable, most of which are non-words. These experimental confusion matrices should therefore be more phonetically based than the naturalistic matrices. By comparing the naturalistic matrices with the experimental matrices, we could further examine the amount of phonetic factors involved in naturalistic misperception, as in the first approach. In addition, we could examine the ecological validity of the experimental matrices. Given that experimental matrices are generated under different experimental manipulations, some manipulations might yield matrices that are more similar to the naturalistic matrices than others. We will focus on two common manipulations: signal-to-noise ratio and bandpass filtering.

Previous attempts have suggested that naturalistic misperception data are congruent with experimental data in Bond, Moore, and Gable (1996), and other studies summarised in Chapter 1, Section 1.3. However, as pointed out by Bond (1999, pp. 135–158), there are difficulties when comparing naturalistic and experimental data. Firstly, the naturalistic data are reported data which rely on the memory of the reporters and the interlocutors involved in a given misperception instance; however, this is not the case for experimental data in which memory/reporting errors can be eliminated.

Secondly, in the experiments, listeners are asked to give their response by writing or by repeating what they heard orally. This creates three additional sources of errors, which are largely independent from perceptual errors, and are spelling errors, speech errors, and listener compliance errors. The latter is when listeners are forced to produce a response even when they have no response to give, and such a response is a reflection of guessing as a response strategy. These three additional sources of errors can potentially add "noise" to the data. While in naturalistic settings, listeners are not forced to write or repeat anything they heard, it is possible the perceptual errors are speech errors: that is, the intended utterance was erroneously spoken.

The third and the most important difficulty is regarding the listening conditions. Experimental data have clearly shown that the more degraded the signal due to manipulation of the listening condition, the more errors are generated. While the listening condition is indeed an important factor in perception, there are currently no benchmark listening conditions that are most representative of our everyday life; therefore, there is no benchmark of confusability of speech sounds (Bailey and Hahn, 2005). Although the listening conditions are not known for the naturalistic data, they are nonetheless collected under a wide range of listening conditions, and the conditions would tend to be more benign than those created in experiments with extreme manipulations of the signal. While these difficulties are indeed valid, most cannot be resolved as they are part of the nature of the data. I argue that the naturalistic data should, in fact, be regarded as the benchmark misperception data because, firstly, they are as ecologically valid as you can get and, secondly, they are not collected under specific listening conditions but a wide range of conditions, thus making them more representative. With the naturalistic data serving as a benchmark, experimenters can better understand which manipulations are required to induce a specific misperception pattern by comparing the experimental data with the naturalistic data. Furthermore, the benchmark corpus can also serve as a starting

point for identifying the existence or absence of a given pattern before conducting a controlled experiment in the laboratory.

In sum, this approach of comparing experimental and naturalistic data cannot only quantify the amount of phonetic biases, but can also serve as a method for evaluating the ecological validity of experimental manipulations. Given that our focus is misperception at the lowest level, the procedure is to first convert the confusion matrices of segments into distance matrices, and then apply correlation tests between the naturalistic distance matrices and the experimental distance matrices in both global and structural levels. Finally, the level of ecological validity is indicated by the strength of the correlation.

3.1.3 Asymmetrical patterns

The method of examining the confusion matrices as distance matrices has a major disadvantage: it loses any asymmetrical information. By examining the asymmetries in both naturalistic and experimental matrices, we can again evaluate to what extent phonetics/phonology play a role in naturalistic misperception, as well as the ecological validity of some of the experimental conditions.

Since not all asymmetrical patterns in perception have an immediate theoretical explanation, we will select three well-known asymmetrical patterns in sound change. If these asymmetrical patterns are also found in misperception, then they could serve as evidence for Ohala's (1989) account of sound change, in which the listeners are a source. Crucially, if a certain sound change is said to be perceptually motivated, then it is vital for the asymmetrical patterns to be found in naturalistic conditions, as well as in a wide range of experimental conditions, and to not simply be experimental by-products. The method of analysis is to extract the confusion matrices containing the relevant segments, convert them into proportions, and then we finally apply the c bias measure (stands for criterion) used in choice theories, which can quantify the

direction and strength of the asymmetries. If the c biases are found to be consistent across the naturalistic and experimental data, then the asymmetrical patterns are robust and reinforces the role of phonetics/phonology in low-level misperception and the complementarity of the naturalistic and experimental data.

3.1.4 Summary

This chapter is broken down into the following sections. First, Section 3.2 describes the extraction of the naturalistic data from the corpus. Second, Section 3.3 explains most of the methods that are used in this chapter in depth. Third, Section 3.4 conducts a descriptive analysis of the data, with the aim of identifying phonetic biases on a featural level, focusing on the confusion rates of consonants and vowels separately by place, manner and voicing for the consonants, and by height and backness for the vowels. Fourth, Section 3.5 and Section 3.6 evaluate the amount of phonetic biases in the naturalistic confusions for vowels (Section 3.5) and consonants (Section 3.6) separately. Fifth, Section 3.7 examines the ecological validity of specific experimental manipulations that are used in experimentally induced misperception studies by comparing experimental data of previous studies to the naturalistic corpus. Sixth, Section 3.8 selects three asymmetrical patterns in perception which mirror certain sound change in progress in English, and examines their robustness across both the naturalistic and experimental data. Finally, Section 3.9 concludes the findings and contribution made in this chapter.

3.2 Data extraction

The naturalistic data is the output of Chapter 2 – a phonetically transcribed corpus which is segmentally aligned. The full corpus contains 5,183 instances of misperceptions. For this chapter, the data contains all sub-corpora and all accent groups,

with two filters. Firstly, all Mondegreen (misperception of music lyrics) instances of misperceptions (253 instances) are excluded. Secondly, any non-English misperceptions (69 instances) are excluded. The filtered corpus contains 4,861 instances of misperceptions.

One might argue that additional filters on the demographics are needed for controlling the effects of dialectal interactions and various factors that could potentially contribute to misperception. One such filter is to exclude any instances that contain speakers and listeners whose accents are not North-American English. While this is possible, the biggest drawback is that it would remove 712 instances, which is a 15% reduction. Furthermore, practically speaking, there are numerous ways of filtering the demographics: e.g. 1) including only instances where the accent of the speaker matches that of the listener, 2) including only the General American accent, and many others. Testing the effect of each of the possible filters on all of the analyses in this chapter is computationally demanding, and it would make this chapter impossible to complete. Therefore, for practical reasons, no filters on the demographics are applied.

The following 28 consonants are considered: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ, j, w]. The following 16 vowels are considered: [i, ɪ, e, ɛ, æ, a, ɑ, ɒ, ɔ, o, u, ʉ, ɜ, ʌ, ʊ, ə]. Finally a gap segment for insertion and deletion [-] is included.

From the alignment, I exclude those aligned pairs that have a many-to-one or one-to-many relationship between the intended segment and the perceived segment. By many-to-one or one-to-many, I mean a complex segment such as [dʒ] being aligned with multiple segments such as [d] and [z]. This leaves us with 90,271 one-to-one aligned pairs.

Finally, given that the focus of this chapter is to analyse bottom-up phonetic/phonological factors in the naturalistic misperception, the aligned pairs are

considered without controlling for anything, i.e. context-free, e.g. their phonological contexts, which words they are from, and many others.

3.3 Method

In this section, we will review a range of analytical techniques that are customary with confusion matrices of perception data.

3.3.1 From counts to distance

Perceptual confusion matrices are count data. Each row of the matrix is the frequency distribution of responses of a particular stimulus and each column represents a possible response. The diagonals are the stimuli being perceived correctly (identical to themselves); the rest of the matrix are the errors. A toy confusion matrix, containing three consonants, is shown in Table 3.1.

Stim.\Resp.	t	p	k
t	50	10	0
p	4	60	6
k	10	20	40

Table 3.1: A toy confusion matrix in counts for the consonants [t, p, k]: the labels on the left represent the intended segments, and those on the top represent the perceived segments; the numbers are the counts of a given intended segment being perceived as a given perceived segment.

A prevalent method of analysing confusion matrices is to convert them into distance matrices. A distance matrix contains the pairwise distances between a set of points (in our case, phones), and it is symmetrical; therefore, the distance between segment A and segment B is the same as that between segment B and segment A. This conversion requires a number of steps. Firstly, the count matrix is converted into a proportion matrix. Secondly, the proportion matrix is converted into a simi-

larity matrix. Finally, the similarity matrix is converted into a distance matrix. An overview of the conversion process is given below.

3.3.1.1 Counts to proportions

The first step of this conversion is to convert the counts into proportions (same as probabilities). This is done for each row of the matrix by dividing each value of the row with the sum of the row (which is the number of times a particular stimulus is presented). To illustrate this, the toy confusion matrix in Table 3.1 is now converted into proportions and is shown in Table 3.2 as a proportion matrix.

Stim.\Resp.	t	p	k
t	0.833	0.166	0
p	0.057	0.857	0.086
k	0.142	0.286	0.571

Table 3.2: A toy confusion matrix in proportions: the labels on the left represent the intended segments, and those on the top represent the perceived segments; the numbers are the proportions of a given intended segment being perceived as a given perceived segment.

3.3.1.2 Proportions to similarity

A proportion matrix can be converted into a similarity matrix. There are a number of established similarity metrics that are used to perform this conversion. Two prevalent metrics are suggested in Shepard (1958) and Shepard (1972) respectively, as shown below.

$$S_{xy} = \sqrt{\frac{p_{x,y}p_{y,x}}{p_{x,x}p_{y,y}}}$$

Figure 3.1: Shepard's (1958) similarity

Where:

- S_{xy} is the similarity value between x and y.

$$S_{xy} = \frac{p_{x,y}+p_{y,x}}{p_{x,x}+p_{y,y}}$$

Figure 3.2: Shepard's (1972) similarity

- $p(x, y)$ is the proportion of times that the segment x was perceived as the segment y. It is the frequency of the segment x perceived as the segment y divided by the frequency of the segment x spoken.

- $p(y, x)$ is the proportion of times that the segment y was perceived as the segment x. It is the frequency of the segment y perceived as the segment x divided by the frequency of the segment y spoken.

- $p(x, x)$ is the proportion of times that the segment x was correctly perceived. It is the frequency of the segment x perceived as the segment x divided by the frequency of the segment x spoken.

- $p(y, y)$ is the proportion of times that the segment y was correctly perceived. It is the frequency of the segment y perceived as the segment y divided by the frequency of the segment y spoken.

Both of these metrics are closely related to Luce's choice rule (Luce, 1963, p. 113), which models human choice behaviours and, importantly, it can account for potential response biases – e.g. if one stimulus category has confusions distributed widely over the possible responses, while another stimulus category has confusions concentrated between two particular responses (Johnson, 2012, Ch. 5). This bias could be removed by weighing the confused proportions (the non-diagonal cells in the matrix) with the correct proportions (the diagonal cells).

Again, for illustration purposes, the previous toy proportion matrix (Table 3.2) was converted into two similarity matrices, each with one of two metrics, as shown in Table 3.3 and Table 3.4.

Stim.\Resp.	t	p	k
t	1	0.115	0
p	0.115	1	0.224
k	0	0.224	1

Table 3.3: A toy similarity matrix with Shepard's (1958) metric: the labels on the left represent the intended segments, and those on the top represent the perceived segments; the numbers are the similarity values of two given segments.

Stim.\Resp.	t	p	k
t	1	0.132	0.102
p	0.132	1	0.260
k	0.102	0.260	1

Table 3.4: A toy similarity matrix with Shepard's (1972) metric: the labels on the left represent the intended segments, and those on the top represent the perceived segments; the numbers are the similarity values of two given segments.

3.3.1.3 Similarity to distance

Finally, a distance matrix can be derived with a similarity matrix. A well-established metric for estimating a psychological distance is based on Shepard's law (Shepard, 1972; Shepard, 1987) which states that the perceptual distance between x and y has a exponential relationship with their similarity, as shown in Figure 3.3. This law is also found to play a role in other non-perceptual contexts – for instance, in information theory (Johnson, 2012, Ch. 5).

$$D_{xy} = -ln\left(S_{xy}\right)$$

Figure 3.3: Shepard's distance (Shepard, 1972; Shepard, 1987)

Where:

- D_{xy} is the distance value between x and y.
- S_{xy} is the similarity value between x and y.

3.3.1.4 Sparse matrix issues

In the previous sections, the overall procedure was outlined for converting confusion matrices from their count forms to distance. However, the procedure will break down due to matrix sparsity – that is, the occurrence of zero counts in a confusion matrix.

The procedure can break down, either at the proportion-to-similarity stage or the similarity-to-distance stage. Let us first take a look at the two similarity metrics (Figure 3.1 and Figure 3.2).

At the proportion-to-similarity stage, the metrics will break down if the diagonal cells of a matrix contain zeros. Take Shepard's (1958) similarity metric, $S_{xy} = \sqrt{\frac{p_{x,y}p_{y,x}}{p_{x,x}p_{y,y}}}$, if *either* $p_{x,x}$ or $p_{y,y}$ were zero – that is, for a given pair of phones, both were misperceived by 100% – then S_{xy} cannot be computed, since one cannot divide a value by zero. Similarly, Shepard's (1972) similarity metric, $S_{xy} = \frac{p_{x,y}+p_{y,x}}{p_{x,x}+p_{y,y}}$, also suffers from the same problem but less severely, since it will only break down if *both* $p_{x,x}$ and $p_{y,y}$ are zero.

At the similarity-to-distance stage, the distance metric (Figure 3.3) will break down if a similarity value is zero, since $ln(0)$ is not defined. Zero similarity values can arise if $p_{x,y}$ and $p_{y,x}$ are zeros. Again, Shepard's (1958) similarity metric will generate a zero similarity value if *either* $p_{x,y}$ or $p_{y,x}$ were zero. Similarly, Shepard's (1972) similarity metric will generate zero if both $p_{x,y}$ and $p_{y,x}$ were zeros.

It is clear that Shepard's (1972) similarity metric is less susceptible to breaking down due to sparse matrices, than Shepard's (1958) metric. This is because Shepard's (1958) metric requires all four parameters to be non-zeros, while Shepard's (1972) metric minimally requires either $p_{x,y}$ or $p_{y,x}$ to be non-zero (for the numerator), and either $p_{x,x}$ or $p_{y,y}$ to be non-zero (for the denominator). For this reason, Shepard's (1972) similarity metric is chosen to be the sole similarity metric used in this thesis.

In any case, even with the more robust Shepard's (1972) metric, we are still faced with possible matrix sparsity issues. This calls for a common solution called

smoothing. The next section will introduce several smoothing techniques, and will establish an adapted smoothing technique that will be used to obtain any distance matrix in this thesis.

3.3.1.5 Sparse matrix smoothing

The sparse matrix issue is, in fact, the same zero frequency problem commonly faced in language models (Jurafsky and Martin, 2008, Ch. 6) and word frequency estimation (Brysbaert and Diependaele, 2013); therefore, it would be beneficial to look at the common smoothing techniques that are used: additive smoothing, Witten-Bell smoothing, and Good-Turing smoothing. Smoothing is performed at the count-to-proportion stage of the procedure.

All smoothing techniques have two components: *discounting* and *backoff*. For each row of a matrix, any zero frequency cells are modelled as events that have not happened yet. Discounting is to take the probability mass away from events that have happened, i.e. the non-zero cells. Backoff is the process of redistributing this probability mass from discounting to the unseen events, i.e. the zero cells. Different smoothing techniques are different in terms of which of the non-zero events the probability mass is taken from, and how we redistribute the probability mass.

3.3.1.5.1 Additive smoothing The simplest kind of smoothing is called *Additive Smoothing*. In additive smoothing, a fixed numeric constant is added to all the cells, both zeros and non-zeros. The two common constants are 1 and a very small number (e.g. 0.00000001). The former is known as Laplace smoothing and the latter is known as Lidstone smoothing. The formula for this smoothing is shown in Figure 3.4, and it performs both discounting and backoff.

Depending on the total number of stimuli presented (the sum of each row), additive smoothing will take too much or too little probability mass from the seen events (the non-zero cells). For example, if a stimulus category has a small number of pre-

sentation (e.g. ten), adding 1 to all the responses (e.g. ten possible responses) will double the overall number of presentation, and the discounted probability mass will therefore be 50%. While one could adjust the additive constant, the precise value is unclear; should it be 1, 0.5 or 0.000000001 ? In general, additive smoothing is a poor smoothing method and the disadvantages are summarised in Gale and Church (1994).

$$Pr_i = \frac{c_i + x}{N + NZ + Z}$$

Figure 3.4: The new estimated probability with additive smoothing

Where:

- Pr_i is the estimated probability for each response category i.
- c_i is the number of counts for each response category i.
- x is the additive constant (1 for Laplace and a small number for Lidstone).
- Z is the number of zero response categories (the zero cells).
- NZ is the number of non-zero response categories (the non-zero cells).
- N is the total number of responses.

3.3.1.5.2 Good-Turing smoothing The *Good-Turing smoothing* (Good, 1953; Gale and Sampson, 1995) relies on the insight that probability mass assigned to categories with zero or low counts can be re-estimated using the probability mass assigned to categories with higher counts. Concretely in the context of a confusion matrix, for each row, all the cells are grouped into frequency bins (frequency of frequencies); for example, we count the number of cells with zeros, the number of cells with 1, the number of cells with 2, etc.

While it is a good smoothing technique for language models, it is inappropriate for confusion matrices. This is because it assumes that frequency bins are relatively

smooth, which is not the case for confusion matrices due to the small number of categories. For instance, a 20 by 20 matrix which has 20 response categories will have many zero/low frequency of frequencies because there will not be enough response categories for a given count value to occur more than once (Nagata, 1998).

3.3.1.5.3 Witten-Bell smoothing A better smoothing method is the *Witten-Bell smoothing* (Witten and Bell, 1991). It is based on the idea that we could estimate the probability of unseen response categories (the zero response categories) with the probability of encountering a response category for the first time (the non-zero response categories). The discounted probability mass of all the zero response categories is the number of non-zero response categories divided by the number of total responses plus the number of non-zero response categories, as shown in Figure 3.5.

$$Pr_{Zero} = \frac{NZ}{NZ+N}$$

Figure 3.5: The total discounted probability mass from Witten-Bell smoothing

Where:

- NZ is the number of non-zero response categories (the non-zero cells).
- N is the total number of responses.

The usual backoff is to divide the discounted probability mass by the number of zero response categories: that is, to redistribute this mass evenly for all the unseen events. This is shown in Figure 3.6.

$$Pr_{Zero_i} = \frac{NZ}{Z \times (NZ+N)}$$

Figure 3.6: The probability of each zero response category from Witten-Bell smoothing with an evenly distributed backoff.

Where:

- Pr_{Zero_i} is the estimated probability for each zero response category i.
- NZ is the number of non-zero response categories (the non-zero cells).
- Z is the number of zero response categories (the zero cells).
- N is the total number of responses.

Witten-Bell smoothing is especially appropriate for confusion matrices (Nagata, 1998). Nagata (1998) applied Witten-Bell smoothing for confusion matrices of Japanese orthographic character confusions in optical character recognition processes. While the size of their matrices (typically 3,000 by 3,000) are much bigger than ours (typically 20 by 20) in misperception, it is still smaller than the usual kind of data (e.g. the size of a vocabulary, typically in the region 100,000) that requires smoothing techniques such as Good-Turing. As mentioned above, both additive smoothing and Good-Turing are inappropriate for confusion matrices. Furthermore, Witten-Bell smoothing has been tested with much smaller number of distinct categories in the domain of confusion matrices; I therefore chose Witten-Bell as my default smoothing technique. In the next section, I will describe a novel weighted backoff method proposed by Nagata (1998), and a novel adaptation of it that I developed.

3.3.1.5.4 Iterative Witten-Bell smoothing In Witten-Bell smoothing, the usual backoff of redistributing the unseen probability mass evenly might not be the best method. Nagata (1998) developed a novel method of applying Witten-Bell smoothing. The novelty lies in their weighted backoff method. Rather than distributing the probability mass evenly, their method would weigh the mass by the similarity between the response category with a zero count and the stimulus category. Their weighted backoff is argued to be more appropriate because it is unreasonable to assume all unseen events are equally probable due to confusions being more likely to occur between similar categories (in their case, orthographic characters).

In their study, they obtained character similarity independently, computed from the feature vectors of the characters. Considering our own misperception data, we could obtain similarity between the categories in each matrix from the confusion matrix itself. This strongly resembles the iterative alignment method by Wieling and Nerbonne (2011) as described in Section 2.4.2, which derives a distance matrix from an initial alignment, which is then fed back to the alignment algorithm as weights; the process repeats iteratively until there is no further change in the resultant alignments. This is the basis of my adaptation of Nagata's (1998) weighted backoff – an iterative backoff – which is described below.

1. Calculate an initial proportion matrix: the unseen probability mass is first divided evenly (Figure 3.6).

2. Create the similarity matrix: the smoothed proportion matrix is then used to calculate a similarity matrix using Shepard's (1972) similarity metric (Figure 3.2).

3. Recalculate the proportion matrix: the resultant similarity matrix is then used to weigh the unseen probability mass in Step 1, such that the probability mass is distributed more/less to zero response categories that have higher/lower similarity with the target stimulus category.

4. Recalculate the similarity matrix: the smoothed (weighted) proportion matrix in Step 3 is then again used to calculate another similarity matrix as in Step 2.

5. Step 3 and 4 are repeated until the proportion matrix stabilises.

In order to evaluate the stability of the proportion matrix, for each iteration, a corresponding distance matrix is calculated using Figure 3.2 and Figure 3.3. We

employed the Mantel test – a correlation test that is especially appropriate for correlating distance matrices (see Section 3.3.2.1 for more information on this test). Furthermore, we select a non-parametric version of this test (Kendall's Tau), which requires no parametric assumptions, thus allowing us to better compare distance matrices of various sizes; the smaller the matrix is, the more likely it would violate parametric assumptions. Using the Mantel test (Kendall's Tau), the distance matrix derived from each iteration can be compared to the distance matrix from the previous iteration, with a correlation coefficient R. While testing this iterative backoff, I found that the R value usually stabilises at around 100 iterations with a value fluctuating just below 1 (e.g. 0.98, with 1 being a perfect correlation), which is a clear sign of stability being reached.

3.3.1.6 Summary

Section 3.3.1 outlined the procedure for converting confusion matrices in counts to distance matrices. I highlighted the sparse matrix issues and three common smoothing methods that tackle them. Shepard's (1972) similarity metric was selected to be the sole metric used in the thesis, since it is less susceptible to sparse matrices. Finally, I adapted the backoff component of one of the smoothing methods, Witten-Bell – which has been shown to be one of the more appropriate methods for confusion matrices – to be able to redistribute the probability mass to the zero response categories weighted by their similarity with the stimulus category, as learnt inductively from the matrix itself.

To recap, a given confusion matrix will be converted into proportions, with any zero response categories smoothed with iterative Witten-Bell smoothing (Figure 3.3.1.5.4). The resultant proportion matrix will be converted into a similarity matrix using Shepard's (1972) metric. Finally, the similarity matrix will be converted into a distance matrix using Shepard's law (Figure 3.3).

3.3.2 Global comparison

3.3.2.1 Mantel correlation: vowels and consonants

The Mantel test (Mantel, 1967) is used to access the global similarity of matrices. It can be applied to distance matrices of vowels and consonants. The Mantel test is a correlation test for assessing the similarity between two distance matrices. It is particularly appropriate for comparing matrices because the values in a distance matrix are not independent of each other. The null hypothesis is that there is no relationship between the two matrices. Its correlation coefficient value can be used to measure the strength of the relationship.

The coefficient of correlation can be computed using Pearson's r, Spearman's rho or Kendall's tau. The Pearson coefficient is parametric, while Spearman's rho and Kendall's tau are both non-parametric. Because our matrices can sometimes be quite small, depending on the number of phones the two matrices have in common (the size of a matrix is restricted to the number of phones we have), I selected one parametric coefficient (Pearson) and one non-parametric coefficient (Kendall's tau). I chose Kendall's tau over Spearman's rho, because Spearman's rho is more prone to error and discrepancies in the data. If a given pair of matrices were to be similar, then the different coefficients (and their significance levels) should be relatively similar; therefore it is important to test both parametric and non-parametric to rule out the chance of a spurious correlation due to the choice of coefficient.

To obtain a significance level for the Mantel test, a permutation test is required. The permutation test is done by:

1. Shuffle the rows and columns of one (or both) of the two matrices.

2. Recompute the correlation coefficient using the shuffled matrices.

3. Repeat the last two steps N times.

4. The p-value (one-tailed) is the number of times the correlation coefficient generated from shuffled matrices has a value greater than the original correlation coefficient from the unshuffled matrices.

To conduct this test, I use the implementation of the Mantel test from the *vegan* library (Oksanen et al., 2013) in R. It requires the user to specify the choice of the correlation coefficient, as well as the number of permutations. The minimum number of permutations to perform is around 1,000; because with 1,000 permutations, the smallest possible p-value is 0.001. For $\alpha = 0.05$, the uncertainty is $\pm 1\%$, which is acceptable. The more permutations, the lower the uncertainty (i.e. the more the better); therefore, if time permits (if the computation does not take too long to run), then a larger number of permutation should be chosen. Having tested the implementation in *vegan*, I found that 10,000 permutation is acceptable in terms of computation time; I therefore chose 10,000 to be the number of permutations for all the Mantel tests performed in this thesis.

3.3.3 Structural comparison

3.3.3.1 Hierarchical clustering: consonants

Agglomerative Hierarchical Clustering is used to compare and examine the internal structures of the distance matrices. Unlike the Mantel test (Section 3.3.2.1), hierarchical clustering seeks to form meaningful clusters which are nested using the distance between any two members (in our case, two phones). It is particularly appropriate for consonants rather than vowels, because consonants are inherently more complex with more phonetic dimensions than vowels. These phonetic dimensions are well-presented hierarchically (cf. Goldsmith, 1976; Dresher, 2008). Therefore, I will only apply hierarchical clustering to consonants and not vowels.

For an extensive review of this method, see Rokach and Maimon (2005). Ag-

glomerative Hierarchical Clustering builds a hierarchy of clusters using a measure of dissimilarity (distances), merged from bottom to top. The clustering output is a *dendrogram* which has a tree-like structure. A dendrogram reflects how the members are clustered in a nested structure, and the similarity levels between any two clusters that were merged.

There are three common strategies of clustering; these strategies are called "linkages". A *Complete* linkage considers the furthest neighbour, where the distance between two groups is defined as the distance between their two farthest-apart members. An *Average* linkage considers the average neighbour, where the distance between two groups is defined as the average distance between each of their members. A *Single* linkage considers the nearest neighbour, where the distance between two groups is defined as the distance between their two closest members.

These linkages have different disadvantages. The single linkage is known for its chaining effect: that is, a few members form a chain/bridge between two clusters to be merged as a single cluster. The average linkage can generate inappropriate clusters if the clusters are elongated, because the average linkage takes the average distance between each of the members of two clusters; then if two elongated clusters are parallel to each other, it could split each of them in half, and the split portions (one half from each cluster) would then form an incorrect cluster. The complete linkage tends to be affected by outliers that do not fit into the overall structure of the cluster.

There is not a standardised method of choosing linkage type. Some linkage types are more appropriate than others for a given set of data. As we have mentioned above, the linkages have different disadvantages, and their severity is dependent on the structure of the data: for instance, if it contains elongated clusters/outliers or not. Therefore, it is important to project the underlying structure of the data with all three common linkage strategies.

Having established the method for projecting the underlying structure of the data, we need methods of comparing two dendrograms quantitatively. I selected two such methods: the cophenetic correlation (Sokal and Rohlf, 1962) and the Baker's Gamma index (Baker, 1974).

The cophenetic[1] correlation is the correlation between two cophenetic distance matrices of two dendrograms. The cophenetic distance is the distance at which clusters are combined in a dendrogram. With a given dendrogram, a cophenetic distance matrix can be obtained. To compare two dendrograms, we first extract their corresponding distance matrix, and then we apply a correlation test on the two cophenetic distance matrices. Two variants of correlation coefficients were chosen: one parametric coefficient (Pearson) and one non-parametric coefficient (Kendall's tau), for the same reason described in Section 3.3.2.1.

The Baker's Gamma index is a measure of similarity between two dendrograms. Take two members of a dendrogram: the height of the dendrogram where the two members are first grouped under the same cluster is identified. The dendrogram is then horizontally "cut" at this height to give k clusters; therefore, this k value is associated with the two members. This process is then repeated for all members pairwise, to obtain their corresponding k values. By doing this for two dendrograms, we could get two sets of k values. These two sets of k values were then correlated using Spearman's rho correlation.

Both the cophenetic correlation and the Baker's Gamma index are correlation tests; therefore, their correlation values range from -1 to +1. To obtain a p-value for these tests, we apply a permutation test. This is done by shuffling the positions of the members in one (or both) of the dendrograms without changing the structure of the dendrogram itself; the shuffled dendrograms are then correlated to obtain a correlation coefficient. This process is then repeated N times. The p-value (one-

[1]Please note that "cophenetic" is *not* a typo of "cophonetic".

tailed) is the number of times the correlation coefficient generated from shuffled dendrograms has a value greater than the original correlation coefficient from the unshuffled dendrograms. The number of permutations was set to 1,000 times, which, for $\alpha = 0.05$, the uncertainty is $\pm 1\%$, which is acceptable.

The R package *dendextendRcpp* (Galili, 2014) is used to construct the hierachical clustering trees with the three common linkages, as well as the correlation tests (the cophenetic correlation and the Baker's Gamma index).

3.3.3.2 Multidimensional scaling: vowels

Just as hierarchical clustering can be used to project the structure of the distance matrices of consonants, classical Multi-Dimensional Scaling (MDS) can be used to project the structure of the distance matrices of vowels. Classical MDS can be used to visualise the level of dissimilarity (distance) of individual members (in our case, vowels) by decomposing them into a k-dimensional space (Gower, 1966). It takes a set of distances and returns a set of points, and the distances between the points approximately reflect the original distances. With $k = 2$, a given distance matrix can be projected into a two-dimensional space. MDS with $k = 2$ is especially appropriate for projecting a distance matrix of vowels, because vowels have primarily two dimensions (frontness and height).

The quality of the projected space can be evaluated visually by comparing it with an acoustic space, and observing the relative positions of the vowels correspond to those in an acoustic space. Quantitatively, the percentage of explained total variance in a two-dimensional solution can be calculated, which is an indicator for the goodness of fit; 100% would mean that the distance matrix can be fully explained in a two-dimensional space.

3.3.4 Interpretation of correlation

The strength level of the correlation coefficients (R/τ values) was interpreted with the following categorisation (*Correlations: Direction and Strength* n.d., Table 3). A perfect correlation has values ±1. A very strong correlation has values ±0.8 to ±0.9. A strong correlation has values ±0.5 to ±0.8. A moderate correlation has values ±0.3 to ±0.5. A modest correlation has values ±0.1 to ±0.3. A weak correlation has values below than ±0.1. A zero correlation has the value 0.

3.4 Descriptive statistics of phonetic bias

This section will provide descriptive statistics of the naturalistic confusion data. Focusing on the rates of substitutions, I aim to identify potential phonetic biases on a featural level. For consonant substitutions, the rates are divided by place, manner and voicing. For vowel substitutions, the rates are divided by height and backness. If phonetic biases were to present on a featural level, then any trend of the substitution rates could be explained with phonetic accounts of perception.

3.4.1 Overall error rates

Errors can be categorised into three broad types: substitutions (perceiving one segment for another segment), insertions (perceiving a segment when there isn't one) and deletions (failing to perceive a segment). We consider the following 28 consonants: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ, j, w], and the following 16 vowels: [i, ɪ, e, ɛ, æ, a, ɑ, ɒ, ɔ, o, u, ʉ, ɜ, ʌ, ʊ, ə].

The confusion matrix of the consonants in percentages (proportions × 100) is shown in Figure 3.7. Similarly, the confusion matrix of the vowels is shown in Figure 3.8. These figures are provided as references, and therefore will not be discussed.

Figure 3.9 summarises the error rates of all segments. The rates are computed

Intended \ Perceived	p	t	k	pʰ	tʰ	kʰ	b	d	g	θ	ð	f	v	s	z	ʃ	ʒ	tʃ	dʒ	h	l	ɹ	ɾ	m	n	ŋ	j	w	-
p	74.62	4.94	4.39	0.96	0.00	0.27	3.29	0.69	0.55	0.14	0.00	2.88	0.69	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.55	0.96	0.27	0.00	0.00	0.14	3.57
t	1.13	79.35	1.50	0.03	0.75	0.12	0.06	1.85	0.03	0.29	0.00	0.17	0.26	0.46	0.23	0.03	0.00	0.09	0.14	0.00	0.29	0.20	0.81	0.14	0.87	0.12	0.61	0.12	10.36
k	1.21	3.81	79.92	0.12	0.18	1.21	0.48	0.73	2.60	0.42	0.06	0.67	0.18	0.18	0.12	0.00	0.00	0.00	0.00	0.00	0.06	0.42	0.67	0.18	0.79	0.12	0.18	0.12	5.57
pʰ	1.54	0.00	0.00	80.15	1.54	3.97	3.33	0.90	0.00	0.00	0.26	2.30	0.26	0.38	0.13	0.00	0.00	0.00	0.13	1.02	0.38	0.00	0.13	0.38	0.13	0.00	0.00	0.26	2.82
tʰ	0.15	2.59	0.00	0.89	84.00	1.63	0.15	1.33	0.52	0.15	0.30	0.30	0.07	0.30	0.07	0.22	0.00	1.19	0.30	0.44	0.00	0.30	0.81	0.22	0.22	0.00	0.30	0.22	3.33
kʰ	0.34	0.43	2.22	3.41	2.30	81.48	0.26	0.85	1.71	0.00	0.00	0.51	0.00	0.34	0.00	0.00	0.00	0.17	0.17	1.54	0.09	0.17	0.17	0.09	0.09	0.00	0.09	0.34	3.24
b	1.68	0.00	0.37	1.17	0.22	0.73	80.18	1.61	0.88	0.07	0.22	1.76	1.90	0.29	0.15	0.07	0.07	0.00	0.07	0.51	0.73	0.15	0.29	1.90	0.66	0.00	0.00	0.51	3.80
d	0.20	2.90	0.63	0.16	0.36	0.12	0.79	77.22	1.31	0.28	0.60	0.28	0.36	0.36	0.36	0.04	0.04	0.04	0.48	0.04	0.56	0.52	1.51	0.24	1.79	0.16	0.44	0.16	8.10
g	0.17	0.50	2.69	0.34	0.17	1.68	1.34	1.85	80.60	0.34	0.34	0.50	0.17	0.25	0.00	0.00	0.00	0.00	0.59	0.76	0.25	0.08	0.67	0.34	0.42	0.42	0.92	0.34	4.28
θ	0.52	2.58	2.58	0.26	0.77	0.52	1.29	1.80	0.77	74.48	0.26	2.58	0.52	1.29	1.03	0.00	0.00	0.26	0.00	0.00	0.77	0.00	0.26	0.00	0.00	0.26	0.00	0.26	6.96
ð	0.06	0.13	0.06	0.13	0.51	0.06	0.57	1.21	0.06	0.19	87.62	0.19	0.13	0.06	0.06	0.06	0.00	0.00	0.00	0.89	0.70	0.19	0.44	0.51	0.19	0.00	0.57	0.57	4.83
f	1.86	0.54	0.39	0.93	0.15	0.93	0.93	0.62	0.31	1.47	0.15	81.11	1.32	2.24	0.00	0.23	0.00	0.00	0.08	0.31	0.39	0.31	0.08	0.77	0.70	0.00	0.23	0.54	3.41
v	0.66	0.41	0.75	0.17	0.17	0.08	3.40	0.75	0.17	0.33	0.25	1.41	79.85	0.17	0.41	0.00	0.00	0.00	0.00	0.17	0.75	0.50	1.08	1.49	1.74	0.00	0.08	1.24	3.98
s	0.05	0.32	0.11	0.05	0.21	0.05	0.05	0.21	0.19	0.11	0.00	0.58	0.03	87.34	3.48	0.61	0.00	0.11	0.05	0.27	0.21	0.27	0.08	0.29	0.13	0.03	0.08	0.03	5.05
z	0.05	0.44	0.25	0.00	0.05	0.00	0.05	0.54	0.05	0.05	0.20	0.00	0.25	6.02	85.50	0.00	0.00	0.05	0.05	0.15	0.20	0.49	0.15	0.00	0.15	0.15	0.15	0.00	5.03
ʃ	0.00	0.54	0.36	0.18	0.36	0.18	0.91	0.18	0.00	0.00	0.00	0.18	0.00	3.99	1.09	86.59	0.54	2.17	0.72	0.18	0.00	0.18	0.00	0.72	0.36	0.00	0.00	0.00	0.54
ʒ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.52	59.26	0.00	14.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.41
tʃ	0.00	0.27	1.08	0.54	2.44	0.54	0.81	0.00	0.00	0.27	0.27	0.27	0.00	2.17	0.00	5.96	0.00	78.86	4.34	0.27	0.00	0.54	0.00	0.00	0.27	0.00	0.00	0.00	1.08
dʒ	0.00	0.43	0.21	0.21	0.85	0.64	0.43	4.26	0.85	0.00	0.21	0.43	0.21	0.43	0.43	1.06	1.06	3.83	77.87	0.43	0.43	0.21	0.85	0.21	0.43	0.00	0.21	0.00	3.83
h	0.00	0.00	0.00	1.00	0.84	1.15	0.92	0.38	0.61	0.00	0.31	0.54	0.00	0.46	0.15	0.38	0.00	0.08	0.23	82.50	0.46	0.46	0.00	0.54	0.38	0.08	0.61	1.00	6.91
l	0.10	0.53	0.10	0.13	0.07	0.07	0.40	0.33	0.10	0.07	0.50	0.46	0.43	0.17	0.00	0.07	0.00	0.03	0.00	0.27	81.33	1.86	0.27	1.16	1.19	0.13	0.27	2.49	7.50
ɹ	0.02	0.25	0.18	0.02	0.04	0.04	0.20	0.13	0.04	0.04	0.02	0.07	0.13	0.29	0.07	0.02	0.00	0.00	0.02	0.27	1.19	86.92	0.18	0.38	0.67	0.09	0.36	1.03	7.29
ɾ	0.73	2.82	1.21	0.08	0.00	0.00	0.32	2.10	0.40	0.16	0.40	0.16	0.81	0.24	0.24	0.00	0.00	0.24	0.24	0.08	0.89	0.56	78.39	0.48	2.18	0.08	0.65	0.16	6.37
m	0.34	0.50	0.17	0.38	0.04	0.04	1.84	0.17	0.13	0.04	0.38	0.42	0.29	0.21	0.08	0.00	0.00	0.00	0.00	0.34	1.01	1.13	0.17	81.73	5.70	0.46	0.42	0.84	3.19
n	0.04	0.67	0.21	0.04	0.13	0.06	0.13	0.82	0.02	0.00	0.04	0.19	0.23	0.15	0.15	0.02	0.00	0.00	0.00	0.08	0.59	0.69	0.73	2.96	84.19	1.15	0.57	0.61	5.52
ŋ	0.00	0.11	0.21	0.00	0.00	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.21	0.43	0.43	0.11	0.00	0.00	0.00	0.00	0.43	0.32	0.00	1.92	6.39	79.77	0.64	0.64	7.99
j	0.00	0.29	0.07	0.02	0.04	0.00	0.02	0.27	0.11	0.00	0.16	0.04	0.04	0.04	0.05	0.02	0.00	0.02	0.02	0.04	0.20	0.27	0.04	0.13	0.38	0.20	90.16	0.56	6.86
w	0.02	0.10	0.07	0.07	0.00	0.00	0.41	0.07	0.02	0.05	0.17	0.12	0.41	0.05	0.00	0.05	0.00	0.00	0.02	0.27	2.28	0.97	0.07	0.48	0.56	0.07	1.36	86.30	6.00
-	0.94	10.28	2.61	0.64	2.01	1.03	2.19	7.16	1.37	0.69	1.97	1.33	1.71	5.83	2.61	0.39	0.09	0.00	0.34	3.08	6.60	12.51	1.76	2.83	8.91	1.46	13.50	6.17	0.00

Figure 3.7: Confusion matrix of consonants with substitution, insertion and deletion in percentages: the labels on the left are the intended segments, and those on the top are the perceived segments; the label "-" is an empty segment used to denote insertion (the last row) and deletion (the last column) errors; the number in the cells represents the response rate in percentage of a given intended segment as a given perceived segment; the numbers sum up to 100% in each row.

on consonants + vowels, consonants, and vowels, and on substitution, insertion and deletion, and a combined rate of substitution, insertion and deletion. The rate of substitutions of all segments (consonants and vowels) is 10.56%. The rate of insertions is 3.66%, and the rate of deletions is 5.17%. The differences between all possible combinations of substitution, insertion and deletion are significant under a

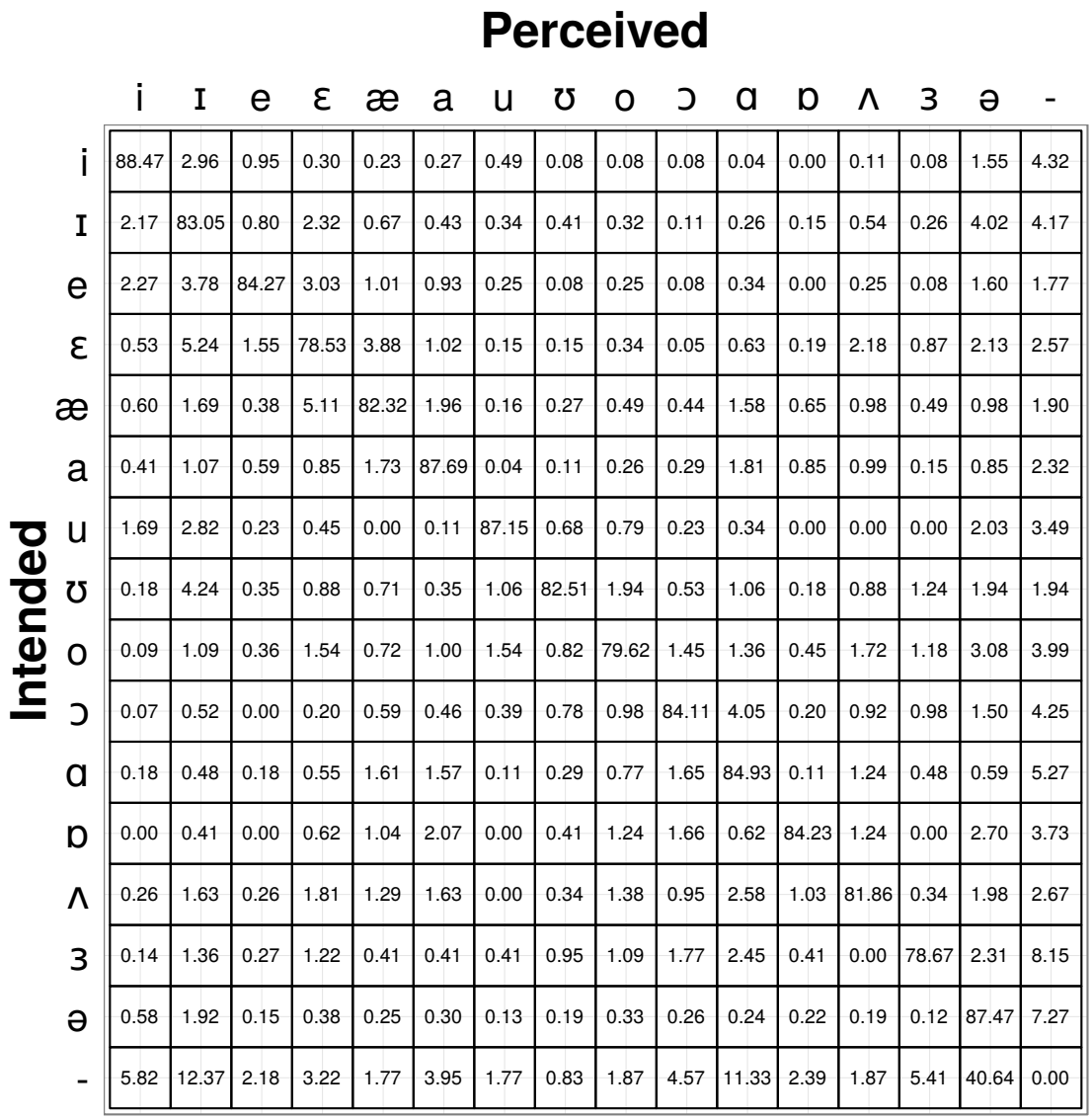


Figure 3.8: Confusion matrix of vowels with substitution, insertion and deletion in percentages: the labels on the left are the intended segments, and those on the top are the perceived segments; the label "-" is an empty segment used to denote insertion (the last row) and deletion (the last column) errors; the number in the cells represents the response rate in percentage of a given intended segment as a given perceived segment; the numbers sum up to 100% in each row.

chi-squared test: a) substitution, insertion and deletion ($\chi^2 = 3949.002$, df $= 2$, $p = 2.2 \times 10^{-16***}$); b) insertion and deletion ($\chi^2 = 244.527$, df $= 1$, $p = 2.2 \times 10^{-16***}$); c) substitution and insertion ($\chi^2 = 3268.22$, df $= 1$, $p = 2.2 \times 10^{-16***}$); and d) substitution and deletion ($\chi^2 = 1818.113$, df $= 1$, $p = 2.2 \times 10^{-16***}$).

There are more substitution errors than deletion errors and insertion errors com-

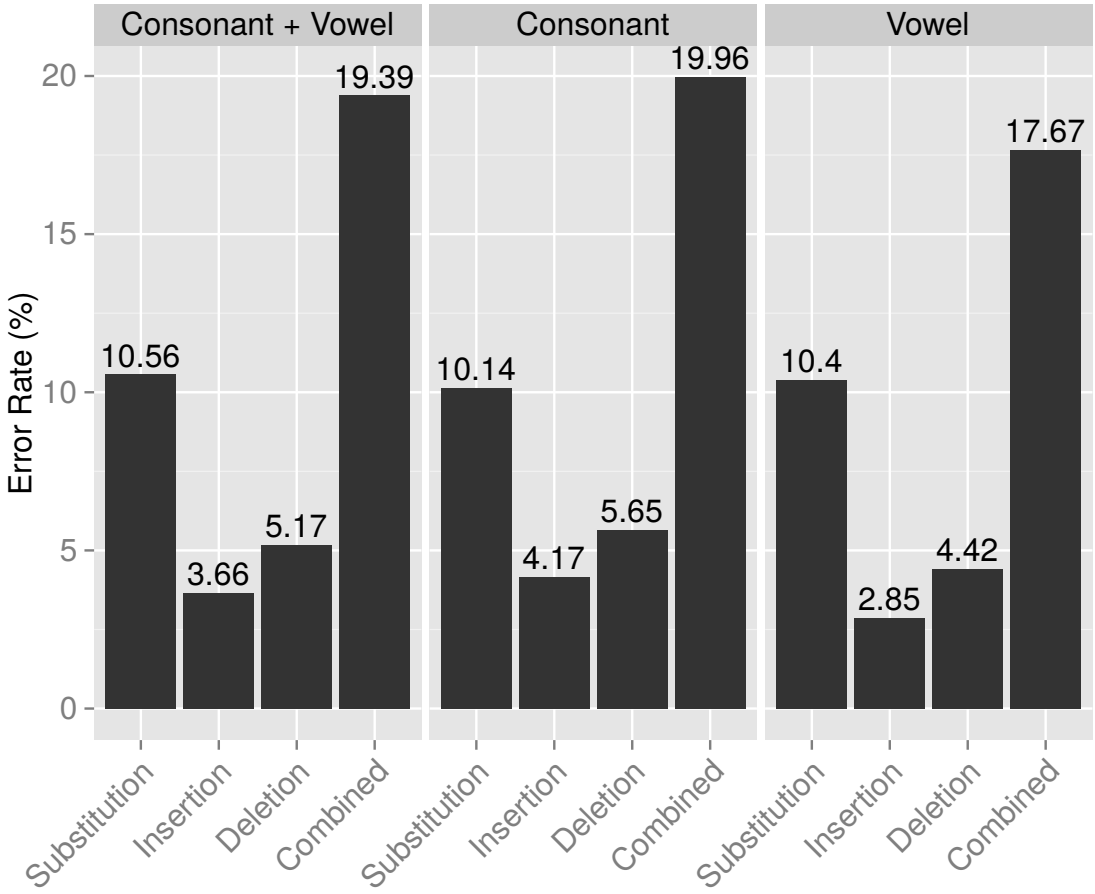


Figure 3.9: Error rate of segments: the three subplots are the rates for consonants + vowels, consonants, and vowels; within each subplot, the rates are shown for substitution, insertion and deletion, and a combined rate of substitution, insertion and deletion; the error rates are printed on top of each bar for clarity.

bined. Deletion errors are, in turn, more common than insertion errors. These relative error rates make intuitive sense. First of all, listeners can deduce the presence of a segment from multiple sources – durational information in the acoustics, phonotactics (e.g. if the listener heard a lax vowel in an English utterance, then the listener can predict a consonant should follow it), and many others. Therefore, one would expect more substitution errors than insertion/deletion errors. Second of all, it is unlikely for listeners to hallucinate a segment when there isn't one (cf. perceptual restoration Warren, 1970), than to fail to detect a segment when there is

one; therefore, one would expect more deletion errors than insertion errors.

By separating the segments into consonants and vowels, we could then find out whether there are more consonant errors than vowel errors, and if the relative error rates between substitutions, insertions and deletions hold for consonants and vowels separately. In terms of consonant errors, the overall rate is 19.96%, of which 10.14% are substitutions, 4.17% are insertions, and 5.65% are deletions. The differences between all possible combinations of substitution, insertion and deletion are significant under a chi-squared test: a) substitution, insertion and deletion ($\chi^2 = 1746.642$, df = 2, $p = 2.2 \times 10^{-16***}$); b) insertion and deletion ($\chi^2 = 132.0375$, df = 1, $p = 2.2 \times 10^{-16***}$); c) substitution and insertion ($\chi^2 = 1505.184$, df = 1, $p = 2.2 \times 10^{-16***}$); and d) substitution and deletion ($\chi^2 = 776.1648$, df = 1, $p = 2.2 \times 10^{-16***}$).

In terms of vowel errors, the overall rate is 17.67%, of which 10.39% are substitutions, 2.85% are insertions, and 4.42% are deletions. To summarise, consonants and vowels together have the following rates: Substitutions (10.56%) > Deletions (5.17%) > Insertions (3.66%); consonants have Substitutions (10.14%) > Deletions (5.65%) > Insertions (4.17%). Finally, vowels have Substitutions (10.39%) > Deletions (4.42%) > Insertions (2.85%). The differences between all possible combinations of substitution, insertion and deletion are significant under a chi-squared test: a) substitution, insertion and deletion ($\chi^2 = 1952.875$, df = 2, $p = 2.2 \times 10^{-16***}$); b) insertion and deletion ($\chi^2 = 119.1773$, df = 1, $p = 2.2 \times 10^{-16***}$); c) substitution and insertion ($\chi^2 = 1569.659$, df = 1, $p = 2.2 \times 10^{-16***}$); and d) substitution and deletion ($\chi^2 = 888.4415$, df = 1, $p = 2.2 \times 10^{-16***}$).

Firstly, we found that there are more consonant errors (19.96%) than vowel errors (17.67%). This difference is significant under a chi-squared test ($\chi^2 = 71.6599$, df = 1, $p = 2.2 \times 10^{-16***}$). This is a recurring finding in naturalistic misperception. Even in the earliest (and smallest, N = 47) naturalistic misperception corpus by Meringer

(1908), the author found that consonants are more erroneous than vowels. Furthermore, this advantage of vowels is also found in experimentally induced misperception data. For instance, in Cutler et al. (2004), the error rate for vowels is $\approx 20\%$ and the error rate for consonants is $\approx 30\%$. It is worth noting that the vowel errors in the naturalistic data could have been inflated by how the vowels were transcribed; for instance, long vowels are treated as two segments. Therefore, the difference between the rate of the consonant errors and that of the vowel errors is likely to be larger.

Secondly, from the breakdown, we see that the substitution rate is basically identical for consonants and vowels separately (both around 10%). The difference in error rates between consonants and vowels comes from insertion errors and deletion errors. This suggests that, given there is a vowel error, a deletion or an insertion is less likely than a substitution than in comparison to the context of consonant errors: i.e. vowels are less likely to be inserted or deleted than consonants. A simple explanation for why consonants are more erroneous than vowels (especially in terms of insertions and deletions) is that vowels are acoustically more robust than consonants, and consonants rely a great deal on vocalic cues in order to be correctly identified in perception (Wright, 2004). Another explanation is that there are more consonants than vowels, so the probability of perceiving the correct consonants based on chance alone is smaller than that of perceiving the correct vowels because of the number of possible choices is greater for consonants than for vowels. One might even speculate that the pattern is explicable using the relative importance of consonants and vowels for making lexical contrasts. In English, consonants carry more functional load than vowels: that is, they are used more to create lexical contrasts (Nespor, Peña, and Mehler, 2003; Surendran and Levow, 2004). However, this functional load account, in fact, predicts the opposite pattern; if consonants are more important lexically, then they should be perceptually more salient. Therefore, functional load is an unlikely explanation.

Finally, the relative error rates between substitutions, insertions and deletions remain relatively similar when we consider consonants and vowels together, and separately. This reinforces the explanation of their relative rates given earlier.

Having explored the overall error rates (substitutions, insertions and deletions) of consonants and vowels, we will now look more closely at substitutions in terms of place, manner and voicing for consonants and in terms of height and backness for vowels.

3.4.2 Consonant confusion of PVM

This section will look at consonant confusion separately of place, manner and voicing. In other words, place confusions, manner confusions, and voicing confusions are examined. Additionally, the confusion rates with two different sets of aligned pairs of segments are explored.

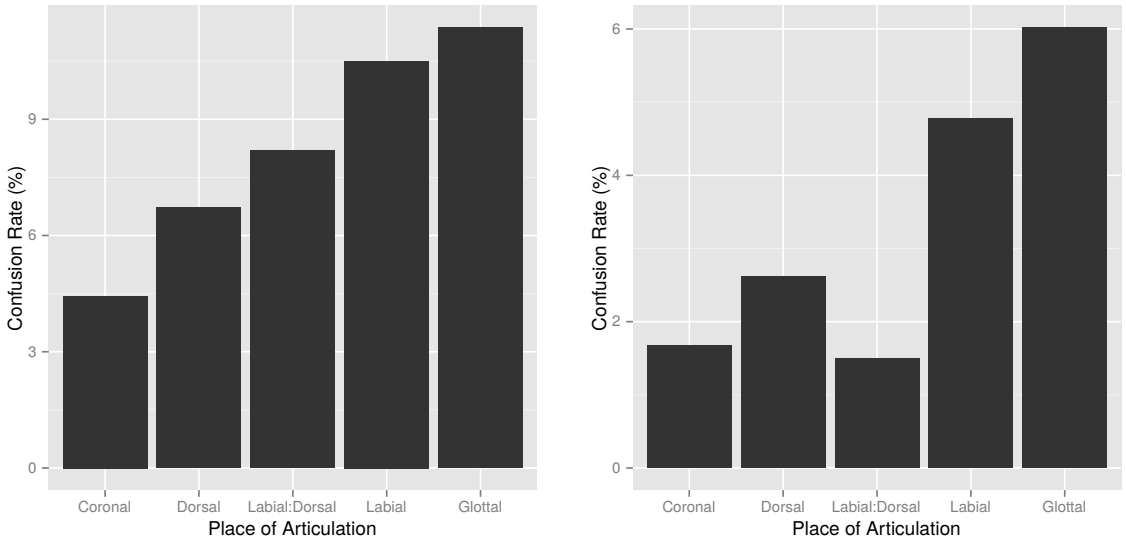
The first set is to simply use all the aligned pairs of segments. The second set is a subset of the first set, which is to only include aligned pairs that have their immediately adjacent segments correctly perceived, e.g. [kʌt] → [kɑt] where [k] and [t] are the same on both intended and perceived.

The second set has the advantage of having more control over the adjacent segments. By controlling the adjacent segments of a given pair of segments being correctly perceived, we can eliminate any crossover effects of multiple instances of misperception; for instance, say an aligned pair of segments is a mismatch (a misperception), and its adjacent segments could also be a mismatch, it is possible that these mismatches are dependent on each other.

3.4.2.1 Place

The place classification is based on Hayes's feature set which was extracted from the software *Pheatures Spreadsheet* (van Vugt, Hayes, and Zuraw, 2012). Four types of

place of articulation are used to cover all of the English consonants: *Coronal*, *Labial*, *Dorsal*, *Labial + Dorsal*, and *Glottal*. The place types, Labial + Dorsal and Glottal, have a narrow coverage – Labial + Dorsal covers only the glide [w], and Glottal covers only [h].



(a) All aligned pairs **(b)** Simple aligned pairs

Figure 3.10: Confusion rate of place of articulation for consonants: (a) All aligned pairs and (b) Simple aligned pairs; the confusion rates are shown as bar charts in percentages, with one bar per place of articulation.

Figure 3.10a and Figure 3.10b show the confusion rates of place with all aligned pairs and simple aligned pairs respectively. Among all the aligned pairs, the Glottal place has the highest rate, followed by Labial, then Labial + Dorsal, then Dorsal, and, finally, Coronal. The rates across both aligned sets (all aligned pairs and simple aligned pairs) are relatively consistent, with the exception of Labial + Dorsal which has the lowest rate in the simple aligned pair set (Figure 3.10b).

The trend of place confusion (from the least confusable to the most confusable) can be summarised as Coronal > Dorsal > Labial + Dorsal > Labial > Glottal. First, the statistical significance of this trend is examined. The difference between each level is compared with the mean of the subsequent levels. Concretely, the following

four contrasts are tested under a chi-squared test: Coronal vs. [Dorsal, Labial + Dorsal, Labial, Glottal], Dorsal vs. [Labial + Dorsal, Labial, Glottal], Labial + Dorsal vs. [Labial, Glottal] and Labial vs. Glottal.

The chi-squared test results on the all aligned pairs are summarised below:

- Coronal vs. [Dorsal, Labial + Dorsal, Labial, Glottal]: $\chi^2 = 348.5409$, df $= 1$, $p = 2.2 \times 10^{-16***}$
- Dorsal vs. [Labial + Dorsal, Labial, Glottal]: $\chi^2 = 70.7124$, df $= 1$, $p = 2.2 \times 10^{-16***}$
- Labial + Dorsal vs. [Labial, Glottal]: $\chi^2 = 17.6241$, df $= 1$, $p = 2.2 \times 10^{-16***}$
- Labial vs. Glottal: $\chi^2 = 0.7496$, df $= 1$, $p = 0.3866^{n.s.}$

Three out of the four contrasts are highly significant. The insignificant contrast is Labial vs. Glottal. This suggests that the trend is significant up to Labial, which can be summarised as: Coronal $>$ (***) Dorsal $>$ (***) Labial + Dorsal $>$ (***) Labial $>$ ($^{n.s.}$) Glottal.

This analysis is repeated for the simple aligned pairs. However, the place, Labial + Dorsal, is excluded because it clearly diverged from the trend. Three contrasts are tested: Coronal vs. [Dorsal, Labial, Glottal], Dorsal vs. [Labial, Glottal] and Labial vs. Glottal.

The chi-squared test results on the simple aligned pairs are summarised below:

- Coronal vs. [Dorsal, Labial, Glottal]: $\chi^2 = 143.0341$, df $= 1$, $p = 2.2 \times 10^{-16***}$
- Dorsal vs. [Labial, Glottal]: $\chi^2 = 54.4878$, df $= 1$, $p = 1.564 \times 10^{-16***}$
- Labial vs. Glottal: $\chi^2 = 2.6194$, df $= 1$, $p = 0.1056^{n.s.}$

Two out of the three contrasts are highly significant. The insignificant contrast is, again, Labial vs. Glottal. This suggests that the trend is significant up to Labial, which can be summarised as: Coronal $>$ (***) Dorsal $>$ (***) Labial $>$ ($^{n.s.}$) Glottal.

Overall the significance of the trends with all aligned pairs and simple aligned pairs are confirmed, with the exception of Labial > Glottal. However, the p-value of this contrast with the simple aligned pairs nearly reached near-significance ($p = 0.1056$), so there is nonetheless a tendency for Glottal to be more erroneous than Labial. While the statistical significance of the trend has been confirmed, the trend still needs to be explained. To do so, the perceptibility of place is first considered.

In terms of perceptibility of place, Jun (2004) proposed a scale of perceptibility (most perceptible to least perceptible) for unreleased stops – Dorsal > Labial > Coronal, which is based on the availability of acoustic cues and evidence from place assimilation. It predicts that Coronal is the most susceptible place to place assimilation, compared to Dorsal and Labial. Comparing this scale to our trend of substitution errors of place, they do not match at all, as we would otherwise expect Coronal to have the highest confusion rate. However, this mismatch is not too surprising, considering this scale of perceptibility is not appropriate for our data because it is limited to unreleased stops and cannot be generalised to all the consonants.

Alternative to the scale of perceptibility, we could consider predictions using markedness of place. In terms of markedness of place, Lombardi (2002) added the Pharyngeal place to Smolensky's (1993) place markedness scale (marked to unmarked) which is *Labial, *Dorsal > *Coronal > *Pharyngeal. The Pharyngeal place is assumed to include the Glottal place (McCarthy, 1994). Lombardi (2002) proposed this scale to explain consonant epenthesis patterns, with the least-marked place being the most frequent target of epenthesis. In terms of our substitution data, this scale predicts that Glottal should have the highest confusion rate, followed by Coronal, then Labial or Dorsal. This scale correctly predicted the high confusion rate for Glottal, but failed to predict the rate of Coronal which is actually the lowest, not the second highest.

Finally, the underspecification account of Coronal makes a different set of predic-

tions. The Featurally Underspecified Lexicon (FUL) model (Lahiri and Reetz, 2002) assumes that not all structured features are specified in the phonological representations of morphemes. Under this model of speech perception, listeners compare an incoming speech signal with the set of features in the phonological representation, with either match, mismatch or no-mismatch as outputs. For there to be a *match*, the signal and the lexicon must share the same feature. A *mismatch* requires the signal and the lexicon not sharing the same feature. Finally, a *no-mismatch* can happen in several conditions, but the one that is of current interest is when the extracted feature from the signal is underspecified in the lexicon. Focusing on the three major places of articulation, Labial, Coronal and Dorsal, under this model, the coronal feature is underspecified, and therefore a hypothesis could be made such that: Pr(Dorsal/Labial $\rightarrow$ Coronal) $>$ Pr(Coronal $\rightarrow$ Dorsal/Labial). This means that the probability of a Dorsal or Labial segment misperceived as a Coronal segment should be higher than the probability of a Coronal segment misperceived as a Dorsal or Labial segment. This is motivated by the model as there is a *no-mismatch* between Dorsal/Labial (acoustic signal) and Coronal (lexicon) which is underspecified; while there is a *mismatch* between Coronal (acoustic signal) and Dorsal/Labial (lexicon). The *no-mismatch* condition could contribute to more misperceptions into Coronal.

Given that non-Coronals are more likely to be perceived as Coronal than the reverse under this model, then we would expect that the confusion rate of Dorsal and Labial to be higher than that of Coronal, which is indeed the case. The underspecification hypothesis is indeed confirmed by a previous analysis on a subset of the naturalistic corpus in Tang and Nevins (2014). This is again confirmed in the current full naturalistic corpus (using all the aligned pairs), showing that the following asymmetrical patterns: Coronal perceived as Dorsal at 1.55% and as Labial at 2.05%, Labial perceived as Coronal at 7.26%, and Dorsal perceived as Coronal at 4.24%.

In sum, the underspecification of Coronal (Lahiri and Reetz, 2002) can explain the fact that Labial and Dorsal have higher confusion rates than Coronal. However, the underspecification account offers no explanations for Glottal, because it is unclear whether it has its own place feature or not [+*Glottal*], as it could be equally capture with [−*Labial*, −*Coronal*, −*Dorsal*], and it is likely to be language-dependent: e.g. in Arabic, the glottal stop patterns with guttural consonants as a natural class, therefore suggesting that they have place features (Lombardi, 2002). While it does not offer an explanation of the Glottal place, Lombardi's (2002) place markedness scale can explain the fact that Glottal has a higher error rate than Labial, Dorsal and Coronal. The place errors found in the naturalistic data can therefore be explained with a combination of the underspecification of Coronal (Lahiri and Reetz, 2002) and Lombardi's (2002) place markedness scale, with the following prediction (least confusable to most confusable): Coronal > [*Labial*, *Dorsal*] > Glottal.

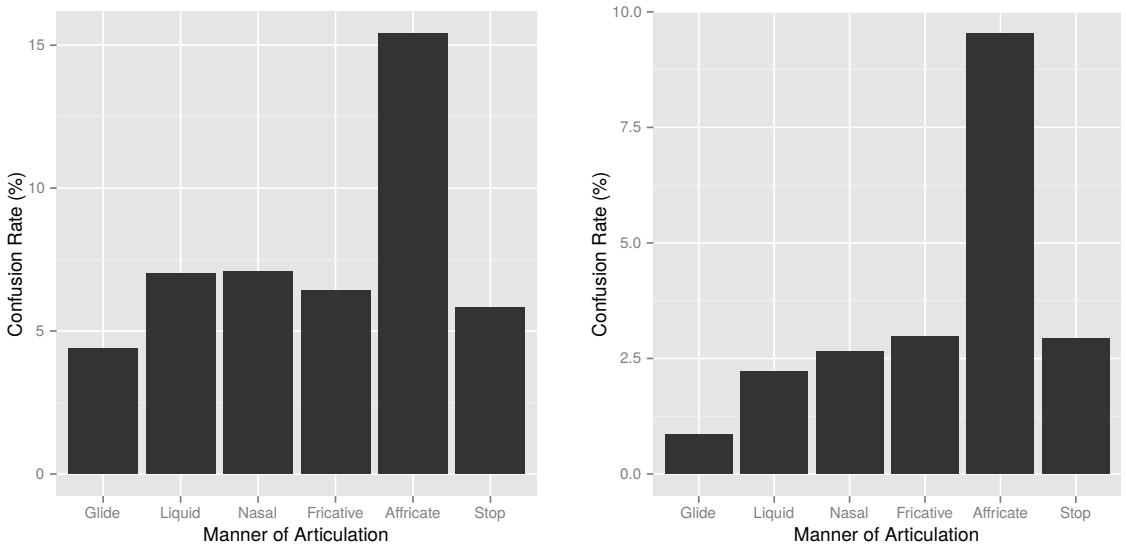
3.4.2.2 Manner

Five types of manner of articulation are considered – Glide, Liquid, Nasal, Fricative, Affricate and Stop. The precise classifications are as follows:

- Glide [j, w]
- Liquid [l, ɹ]
- Nasal [m, n, ŋ]
- Fricative [ʃ, ʒ, θ, ð, s, z, f, v, h]
- Affricate [tʃ, dʒ]
- Stop [p, t, k, b, d, g, p^h, t^h, k^h, ɾ]

[ɾ] is classified as a stop, and not a liquid. This is based on its phonetic (not phonological) properties. Furthermore, for English, [ɾ] is underlyingly /t, d/ which

are also stops.



(a) All aligned pairs **(b)** Simple aligned pairs

Figure 3.11: Confusion rates of manner of articulation for consonants: (a) All aligned pairs and (b) Simple aligned pairs; the confusion rates are shown as bar charts in percentages, with one bar per manner of articulation.

Figure 3.11a and Figure 3.11b show the confusion rates of manner with all aligned pairs and simple aligned pairs respectively. In both sets of aligned pairs, affricate has the highest confusion rate, and it is a lot higher (at least two times higher) than all the other manners, and glide has the lowest confusion rate. Besides affricate and glide, the other manners have different relative rates in both plots. This indicates that there is a crossover effect of multiple instances of misperception on manner (more so than place, since only Labial + Dorsal differed). With all aligned pairs (Figure 3.11a), nasal and liquid have the second highest rate (after affricate), followed by fricative, then stop, and, finally, glide. With simple aligned pairs (Figure 3.11b), stop and fricative have the second highest rate (after affricate), followed by nasal, then liquid and finally glide. The pattern with the simple aligned pairs has a striking resemblance with the sonority scale.

A typical sonority scale of manner (most sonorous to the least sonorous) is Glide > Liquid > Nasal > Fricative > Affricate > Stop (Parker, 2002). With the exception of affricate, the sonority scale of manner matches the error rate with the simple aligned pairs. In terms of perception, sonority offers an explanation that the more sonorous a sound is, the more acoustic energy it contains; it is therefore less susceptible to confusions.

Before we can accept this sonority explanation, the statistical significance of the sonority trend needs to be tested. Just as the analyses in the place section, the difference between each level is compared with the mean of the subsequent levels. However, the manner affricate is excluded from this analysis because it is a clear outlier. Therefore, the following four contrasts are tested under a chi-squared test: Glide vs. [Liquid, Nasal, Fricative, Stop], Liquid vs. [Nasal, Fricative, Stop], Nasal vs. [Fricative, Stop] and Fricative vs. Stop.

The chi-squared test results on the simple aligned pairs are summarised below:

- Glide vs. [Liquid, Nasal, Fricative, Stop]: $\chi^2 = 91.0267$, df $= 1$, $p = 2.2 \times 10^{-16***}$
- Liquid vs. [Nasal, Fricative, Stop]: $\chi^2 = 6.8949$, df $= 1$, $p = 0.008644 \times 10^{-16**}$
- Nasal vs. [Fricative, Stop]: $\chi^2 = 1.2758$, df $= 1$, $p = 0.2587^{n.s.}$
- Fricative vs. Stop: $\chi^2 = 0.0234$, df $= 1$, $p = 0.8784^{n.s.}$

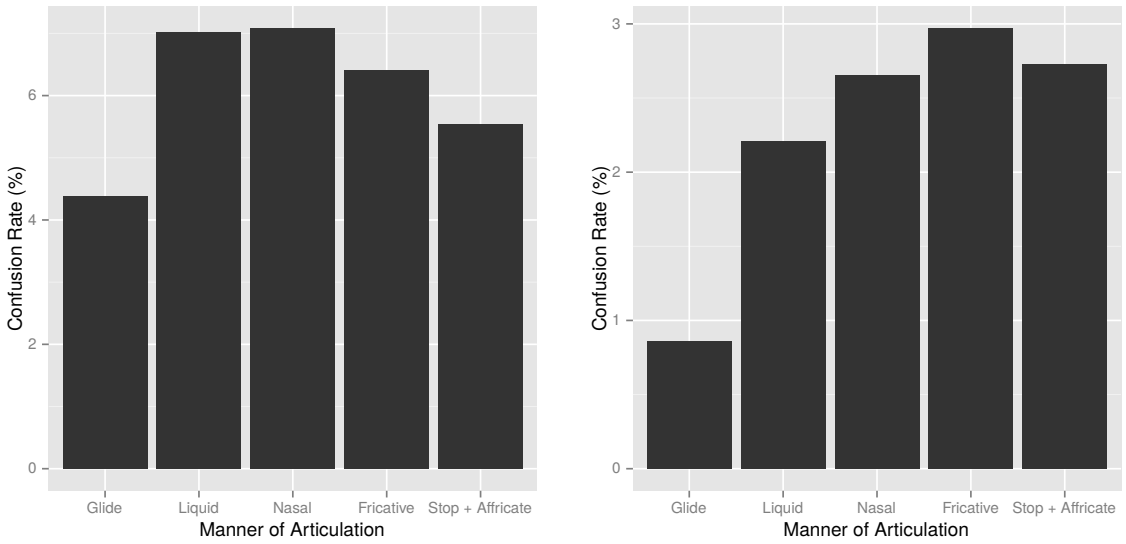
Two out of the four contrasts are highly significant. The insignificant contrasts are Nasal vs. [Fricative, Stop] and Fricative vs. Stop. This suggests that the trend is significant up to Nasal, which can be summarised as: Glide > (***) Liquid > (***) Nasal > ($^{n.s.}$) Fricative > ($^{n.s.}$) Stop. By inspecting the exact p-values, there is a decreasing trend from Glide to Fricative. While the contrast, Nasal vs. [Fricative, Stop], is not significant, its p-value (0.2587) is smaller than that of Fricative vs. Stop

(0.8784). Therefore, this suggests the trend weakens from the most sonorous manner, Glide, to the least sonorous manner, Stop.

While we could partially explain the confusion patterns in terms of sonority (at least with the simple aligned pairs), the exceptionally high affricate rate still requires an explanation. Looking more closely at the segments that the affricates have high confusions with, we found that [tʃ] is most frequently confused with [ʃ] at a rate of 6.03%, followed by [dʒ] at 4.38% and [t^h] at 2.46%; similarly [dʒ] is most frequently confused with [d] at a rate of 4.42%, followed by [tʃ] at 3.98% and [ʒ] at 1.11%. Of the three most frequently confused segments, one differs in voicing ([dʒ] as [tʃ], and [tʃ] as [dʒ]) (which is irrelevant to the high confusion rate of the affricate manner); crucially, the other two are smaller elements of the affricates. One explanation is that the high confusion rates with the affricates is due to the fact that it is composed of the stop portion and the fricative portion; it could therefore be confused as a stop (e.g. [dʒ] > [d]) or a fricative (e.g. [dʒ] > [ʒ]). This consequently doubles the chance of a confusion; coincidently, the confusion rate of affricate is two to three times as high as that of either the stop or the fricative.

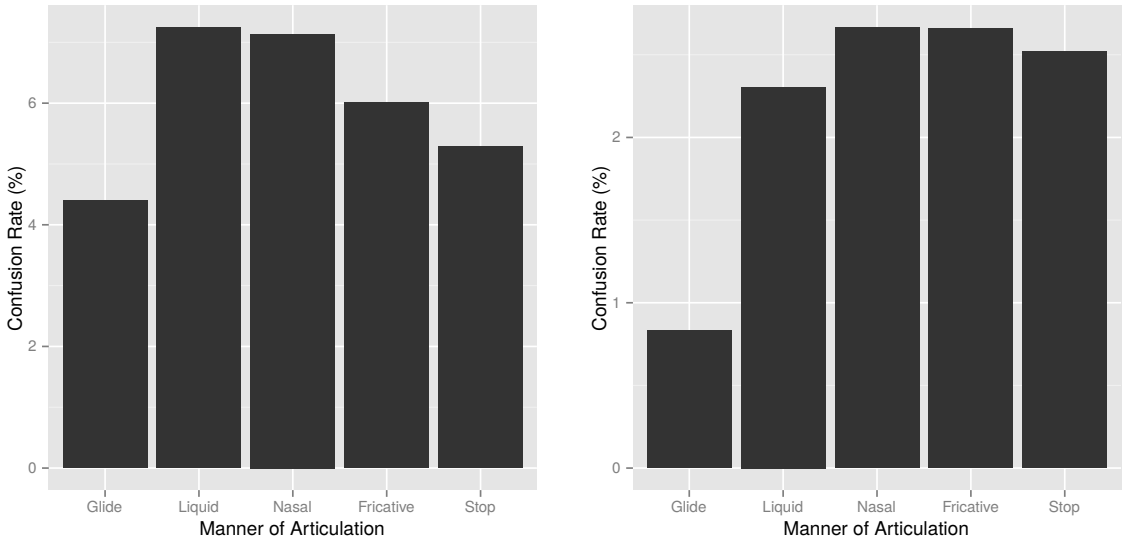
Perhaps this divergence of affricates with the sonority scale can be mediated by categorising affricates as strident stops, which would group affricates and stops together. This is in line with Jakobson, Fant, and Halle's (1952) treatment of affricates as [strident, -continuant]. Therefore, any stop and affricate confusions would no longer be counted as being confused. Figure 3.12 shows the outcome of this treatment. The confusion rate of stops is now lower than before, and indeed lower than that of fricatives. In fact, this treatment created a divergence of stops.

An alternative treatment for the affricates is to consider them as separate segments: a stop and a fricative. The outcome of this treatment is shown in Figure 3.13. This treatment reduced the confusion rates of both fricatives and stops. Fricative is at a similar rate as nasal, and stop is lower than fricative. Both diverged



(a) All aligned pairs **(b)** Simple aligned pairs

Figure 3.12: Confusion rates of manner of articulation for consonants, with stops and affricates as a single manner category: (a) All aligned pairs and (b) Simple aligned pairs; the confusion rates are shown as bar charts in percentages, with one bar per manner of articulation.



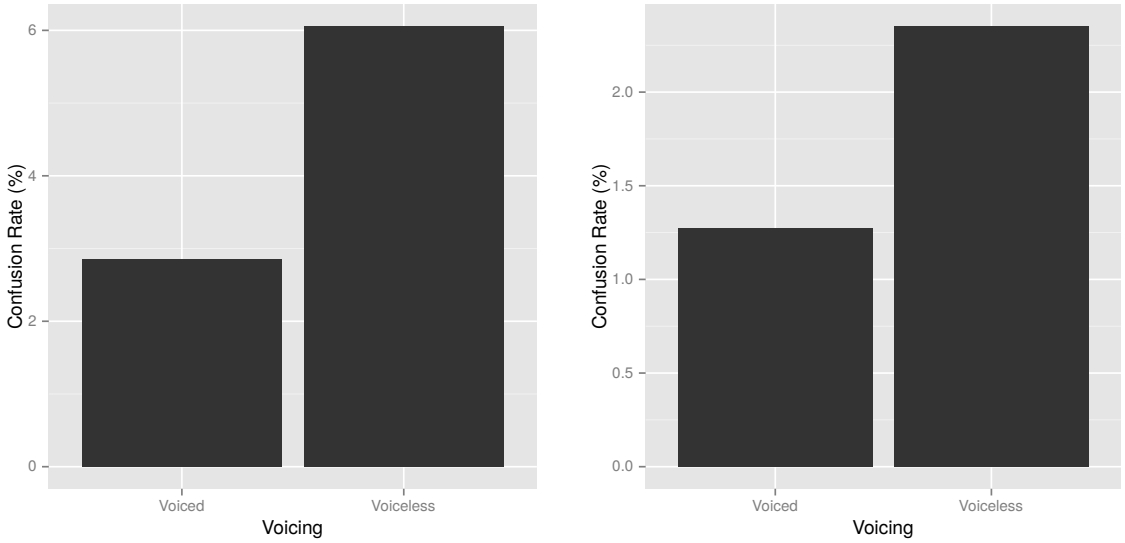
(a) All aligned pairs **(b)** Simple aligned pairs

Figure 3.13: Confusion rates of manner of articulation for consonants, with affricates being split as stops and fricatives: (a) All aligned pairs and (b) Simple aligned pairs; the confusion rates are shown as bar charts in percentages, with one bar per manner of articulation.

from the sonority scale. In sum, these two treatments of affricates cannot improve the overall fit with the sonority scale. In fact, the simplest treatment is to exclude affricates. The fit with the sonority scale is strongest with the simple aligned pairs and the exclusion of affricates. Interestingly, this treatment mirrors the arguments by Kehrein (2002, Ch. 2) who argues that, from the perspective of contrasts and natural classes cross-linguistically, the manner affricate should be eliminated as a phonological concept.

3.4.2.3 Voicing

Finally, the confusion rates of voicing, which is either voiced or voiceless, are examined. Figure 3.14a and Figure 3.14b show the confusion rate of voicing with all aligned pairs and simple aligned pairs respectively.



(a) All aligned pairs **(b)** Simple aligned pairs

Figure 3.14: Confusion rate of voicing for consonants: (a) All aligned pairs and (b) Simple aligned pairs; the confusion rates are shown as bar charts in percentages, with one bar per voicing category.

Across both plots, the trend is identical with voiceless consonants being more confusable than voiced consonants. The fact that the trend is identical indicates that

voicing is not susceptible to crossover effects from multiple instances of misperception: that is, voicing errors are relatively independent from the phonological environments, unlike manner errors. This difference between voiced and voiceless is significant under a chi-squared test for the all aligned pairs ($\chi^2 = 41866.8$, df $= 1$, $p = 2.2 \times 10^{-16***}$) and for the simple aligned pairs ($\chi^2 = 60.6747$, df $= 1$, $p = 6.733 \times 10^{-15***}$).

The error trend (high to low) voiceless > voiced can be explained by the fact that voiced consonants have more acoustic energy and therefore perceptually more salient; this also fits with a sonority account because voiced consonants are more sonorous than voiceless consonants.

3.4.3 Vowel confusion of height and backness

Similar to the consonant section, the confusion rates for vowels are analysed separately in terms of *Height* and *Backness* (with each dimension having three levels). The three levels of height are Close, Mid and Open. The three levels of backness are Front, Central and Back. The precise classifications are as follows:
Height:

- Close [i, ɪ, ʉ, ʊ, u]
- Mid [e, ɛ, ɜ, ə, ɔ, o, ʌ]
- Open [æ, a, ɑ, ɒ]

Backness:

- Front [i, ɪ, e, ɛ, æ, a]
- Central [ʉ, ɜ, ə]
- Back [ɔ, o, ʌ, ʊ, u, ɑ, ɒ]

3.4.3.1 Height

Figure 3.15 shows the confusion rate of vowels by height. Open and close vowels have similar error rates, and they are not significantly different ($\chi^2 = 0.1926$, df = 1, $p = 0.6608^{n.s.\cdot}$). Both open and close vowels have a higher error rate than mid vowels, and this is significant under a chi-square test. Close and mid vowels are significantly different ($\chi^2 = 7.5132$, df = 1, $p = 0.006125^{**}$), and open and mid vowels are significantly different ($\chi^2 = 9.8679$, df = 1, $p = 0.001682^{**}$).

A sonority/acoustic energy account cannot capture this pattern, since it would predict that open vowels are the least confusable, followed by mid vowels and, finally, close vowels, because the size of jaw aperture correlates positively with the amount of acoustic energy.

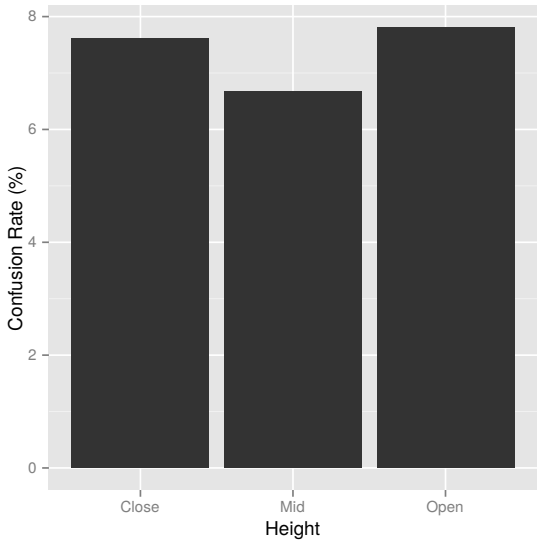


Figure 3.15: Confusion rate of vowel height: the confusion rates are shown as bar charts in percentages, with one bar per height level.

To examine this pattern further, a confusion matrix of vowel height in proportions is shown in Table 3.5.

Close vowels are perceived as mid vowels at a rate of 6.4%, and as open vowels at 1.2%. Open vowels are perceived as mid vowels at a rate of 6.1% and as close vowels at 1.7%. Mid vowels are perceived as close vowels at a rate of 3.7% and as

Stim.\Resp.	Close	Mid	Open
Close	0.924	0.064	0.012
Mid	0.037	0.933	0.029
Open	0.017	0.061	0.922

Table 3.5: Confusion matrix of vowel height in proportions: the labels on the left (stimulus) are the intended height, and the labels on the top (response) are the perceived height.

open vowels at 2.9%. These proportions indicate that close vowels are more likely to be perceived as mid vowels (6.4%) than as open vowels (1.2%), and the difference in proportions is substantial and statistically significant ($\chi^2 = 323.1237$, df $= 1$, $p = 2.2 \times 10^{-16***}$). Open vowels are more likely to be perceived as mid vowels (6.1%) than as close vowels (1.7%), and the difference in proportions is substantial and statistically significant ($\chi^2 = 186.1061$, df $= 1$, $p = 2.2 \times 10^{-16***}$) (just as the confusions of close vowels). The mid vowels, however, do not have a strong perpetual bias towards open vowels (3.7%) or close vowels (2.9%), with similar proportions of confusions; even though this difference is statistically significant ($\chi^2 = 14.0532$, df $=$ 1, $p = 0.0001777^{***}$), the difference is a lot smaller than the confusions of close/open vowels, as indicated by the smaller χ^2 value (14.0532 compared to 323.1237/186.1061) or the smaller p-value (0.0001777 compared to 2.2×10^{-16}). These confusion patterns can be explained with Steriade's (2001) account of perceived similarity.

Under this account, the probability of a confusion between two phones is a function of the perceived similarity of the two phones. This can be formulated as $Pr(Input \rightarrow Output) = f(Perceived\ Similarity(Input,\ Output))$. The acoustic distances between close and mid vowels, and open and mid vowels is shorter than the distances between close and open vowels. The confusion rates between close and mid vowels, and open and mid vowels are higher than the confusion rates between close and open vowels; this pattern can be explained using the relative perceived similarity between vowel height by assuming that acoustic distances are correlated

with perceived similarity. Under this account, the Perceived Similarity of {Open, Mid} should be similar to that of {Close, Mid}, and both should be more similar than that of {Close, Open}. This perfectly predicts our proportions in Table 3.5. The proportion (which is the same as probability) of [Close → Mid] is 6.4% which is similar to that of [Open → Mid] (6.1%); both proportions are higher than that of [Close → Open] (1.2%) and that of [Open → Close] (1.7%). In sum, Steriade's (2001) account of perceived similarity can predict the probability of confusion from a given vowel height to other vowel heights.

However, the perceived similarity account cannot explain the fact that mid vowels are less confusable than close and open vowels. From the proportion matrix, an asymmetrical pattern can be observed, such that close vowels are perceived as mid vowels (6.4%) more often than the reverse (3.7%); similarly, open vowels are perceived as mid vowels (6.1%) more often than the reverse (2.9%). The low substitution error rate of mid vowels is a reflection of this asymmetrical pattern. I have no immediate explanation for this perceptual bias from open/close vowels to mid vowels. An explanation for this pattern is provided later in Chapter 4, Section 4.2.3.

3.4.3.2 Backness

Figure 3.16 shows the confusion rate of vowels by backness. Back vowels are the most confusable, followed by central vowels and finally front vowels. Again, the sonority account fails to capture this trend, as it would predict that back vowels are less confusable, followed by mid vowels and finally front vowels, because of the size of the oral cavity (back vowels have the largest, front vowels have the smallest).

The statistical significance of this trend is examined. The difference between each level is compared with the mean of the subsequent levels. The following two contrasts are tested under a chi-squared test: Front vs. [Central, Back] and Central vs. Back.

The chi-squared test results are summarised below:

- Front vs. [Central, Back]: $\chi^2 = 17.7034$, df $= 1$, $p = 2.582 \times 10^{-05***}$
- Central vs. Back: $\chi^2 = 8.2595$, df $= 1$, $p = 0.004054^{**}$

The chi-squared test results confirm the sonority trend, since both contrasts are statistically significant.

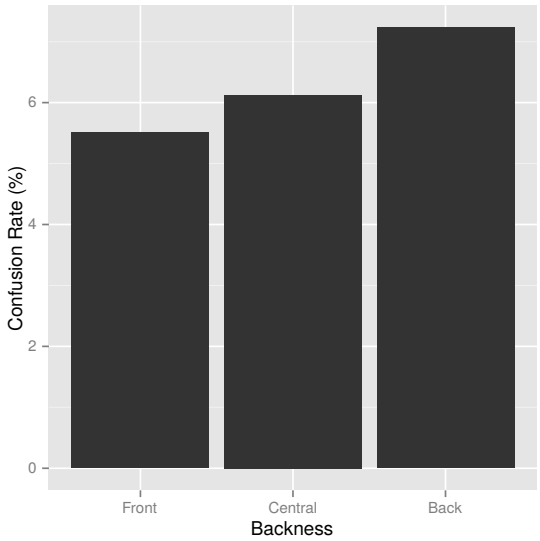


Figure 3.16: Confusion rate of vowel backness: the confusion rates are shown as bar charts in percentages, with one bar per backness level.

Stim.\Resp.	Front	Central	Back
Front	0.945	0.027	0.028
Central	0.040	0.939	0.021
Back	0.048	0.024	0.928

Table 3.6: Confusion matrix of vowel backness in proportions: the labels on the left (stimulus) are the intended backness, and the labels on the top (response) are the perceived backness.

To examine this pattern further, a confusion matrix of vowel backness in proportions is shown in Table 3.5. The back vowels are perceived as front vowels at a rate of 4.8%, and as central vowels at 2.4%. This difference is statistically significant

($\chi^2 = 65.726$, df $= 1$, $p = 5.182 \times 10^{-16***}$); therefore, there are significantly more back-to-front errors than back-to-central errors.

Central vowels are perceived as front vowels at 4%, and as back vowels at 2.1%. This difference is statistically significant ($\chi^2 = 54.7024$, df $= 1$, $p = 1.402 \times 10^{-13***}$); therefore, there are significantly more central-to-front errors than central-to-back errors.

Front vowels are perceived as central vowels at 2.7% and as back vowels at 2.8%. This difference is statistically *in*significant ($\chi^2 = 0.6749$, df $= 1$, $p = 0.4113^{n.s.}$); therefore, the number of front-to-central errors is similar to the number of front-to-back errors.

Overall, we see a pattern of vowel fronting, with back vowels being perceived as front vowels, and central vowels being perceived as front vowels more often than the reverse direction (backing). Unlike the analyses with vowel height, the perceived similarity account does not seem to play a role here, since we could otherwise expect the rate of [Front → Central] to be higher than that of [Central → Back], the rate of [Central → Front] to be of a similar rate as that of [Central → Back], and the rate of [Back → Central] to be higher than that of [Back → Front]. However, none of these predictions are correct. This observation of perceptual vowel fronting is examined in depth in a later section, Section 3.8.4, in which I argue that this fronting pattern supports the third principle of vowel chain shifts (Labov, 1994a, p. 116).

3.4.4 Conclusion

In this section, a descriptive analysis of the naturalistic confusion matrix was conducted. Crucially, phonetic patterns in the rates of segmental confusions on a featural level were identified.

In the first section, the overall error rates were analysed. There are more substitution errors than insertion and deletion errors, and deletion errors in turn are more

frequent than insertion errors. This pattern is true when considering both consonants and vowels together, or separately. The pattern of having more substitutions than insertions + deletions can be explained by how the presence of a segment is supported by both multiple cues (acoustics, phonotactics, and others). Since listeners are unlikely to imagine a non-existent segment, there are more deletions than insertions. Furthermore, there are more consonant errors than vowel errors. This difference is driven mainly by a higher insertion and deletion rate of the consonants than of the vowels, thereby highlighting the perceptual salience of vowels, as well as how consonants rely hugely on transitional cues in the vowels (Wright, 2004).

The substitution rate of consonants was then analysed by place, manner and voicing. In terms of place, the confusion rates have the following trend (low to high) Coronal > Dorsal > Labial + Dorsal > Labial > Glottal, which is best explained using a combination of the underspecification of coronal (Lahiri and Reetz, 2002) and Lombardi's (2002) place markedness scale. In terms of manner, after restricting the adjacent segments, the confusion rates have the following trend (low to high) Glide > Liquid > Nasal > { Fricative, Stop } > Affricate. With the exception of Affricate, the trend matches the sonority scale with more sonorous manners being more confusable. In terms of voicing, voiceless consonants are more confusable than voiced consonants, which again can be explained using a sonority/acoustic energy account.

Finally, the substitution rate of vowels was analysed by height and backness. In terms of height, open and close vowels are more confusable than mid vowels, and, more specifically, the open and close vowels are mainly confused with the mid vowels. This can be explained with Steriade's (2001) account of perceived similarity, and not with sonority. In terms of backness, back vowels are the most confusable vowels, followed by mid vowels, then finally front vowels. A closer analysis of the confusions in proportion shows that there is an overall perceptual bias of back/central vowels

being perceived as front vowels, and this is addressed in a later section of this chapter (Section 3.8.4).

3.5 Analyses of phonetic bias in vowel confusions

To establish the phonetic bias in the vowel misperception in the naturalistic corpus, it would be informative to compare them with distances that are phonetically based.

3.5.1 Acoustic distances

To obtain a baseline of vowel distances, we will look for acoustic measurements obtained from two classic acoustic studies of American English vowels.

The first set of large-scaled measurements were obtained by Peterson and Barney (1952). In this study, they recorded 76 speakers, including 33 men, 28 women and 15 children. Each speaker produced the ten vowels [i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ] in the context of /hVd/ twice. This amounts to 1,520 recorded words. In terms of the demographics of the speakers, the majority of the women and children grew up in the Middle Atlantic speech area, while the male speakers were more heterogeneous demographically and the majority spoke General American.

The second mega study was a follow-up study by Hillenbrand et al. (1995), conducted ≈ 40 years after Peterson and Barney (1952). They recorded 139 speakers, including 45 men, 48 women, and 46 children. Each speaker produced the twelve vowels [i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɝ] in the context of /hVd/, twice for the children, and three times for the men and the women. Notably, 87% of the speakers were raised in Michigan's lower peninsula. The speakers underwent an extensive dialect screening process to ensure that their production of the vowels in questions are contrastive.

The F1 and F2 formant values (Hz) of the vowels recorded in these two studies

were extracted from the publications. Only the average values for each of the three populations (men, women and children) were available. In order to obtain an average representation of American English vowels, the formant values of both studies are further averaged across studies for men, women and children separately, with the exception of the vowels that exist only in Hillenbrand et al. (1995), which can only be taken directly from one study. These formant values averaged across studies are further averaged across the three populations. The resultant values are therefore an averaged representation of twelve American English vowels [i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɝ], spoken by 215 speakers.

To obtain acoustic distances from formant values, we first converted the formant values from Hertz to Mels. The choice of the Mel scale over the Bark scale (Traunmüller, 1990) is arbitrary. The resultant conversion will be similar with either scale. Crucially, they can both capture the fact that human perception of frequency is non-linear (Stevens, Volkmann, and Newman, 1937). There are multiple versions of the Mel-scale formula; a popular version published in O'Shaughnessy (1987) was chosen, $m = 1127.01048 \times ln(1 + (f/700))$, where f is a formant value in Hertz and m is the converted value in Mels. A distance matrix of the vowels was then calculated using the Euclidean distance, which is an established distance metric for converting relative formant values to distances (Yilmaz, 1967; Yilmaz, 1968).

3.5.2 Perceptual distances

To obtain perceptual distances of American English vowels, I extracted a vowel confusion matrix from the naturalistic data. Only vowel to vowel aligned pairs were considered. Furthermore, on top of the monophthongs (long or short), only the nucleus, and not the offglides, of the diphthongs were included, as shown in Figure 2.2, with 16 vowels [i, ɪ, e, ɛ, æ, a, ɑ, ɒ, ɔ, o, u, ɯ, ɜ, ʌ, ʊ, ə] . The vowel confusion matrix is therefore 16 by 16.

Since the acoustic distances (from Hillenbrand et al., 1995; Peterson and Barney, 1952) were primarily calculated from vowels of the General American accent, I excluded [ʉ] and [ɒ] from the vowel set as they are primarily from the British accents in the corpus. The resultant confusion matrix is 14 by 14. This confusion matrix was then converted into a distance matrix using the procedure described in Section 3.3.1. Crucially, this distance matrix is perceptually grounded because it is based on Shepard's law which best captures the relationship between perceptual distance and confusion similarity.

3.5.3 Comparison – acoustic and perceptual distances

This section compares the acoustic and perceptual distances, obtained from acoustic measurements of formant values from experimental studies, and confusability of vowels from the naturalistic corpus.

Since the naturalistic distance matrix contains more vowels than the acoustic distance matrix, we removed the additional vowels [ə] and [a] from the naturalistic distance matrix. Furthermore, the vowel [ɜˠ] in the acoustic distances is treated as [ɜ] in order to match with the NURSE vowel in naturalistic distance matrix. With these two treatments, both matrices now contain the same set of vowels.

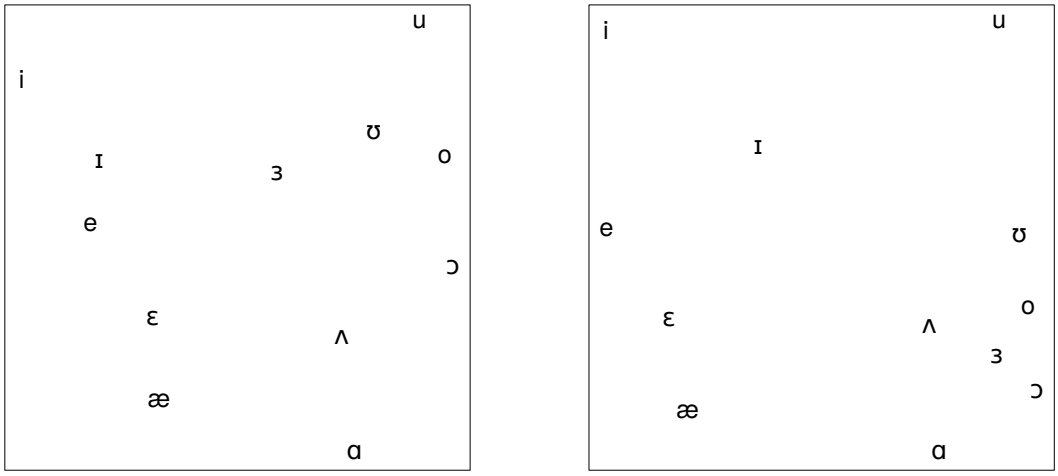
3.5.3.1 Global similarity

To analyse the global similarity of the acoustic and perceptual distances, we employ the Mantel test, as described in Section 3.3.2.1. With Pearson coefficient (a parametric test), the correlation is $r = 0.6831$ ($p = 9.999 \times 10^{-5}$ with 10,000 permutations). With Kendall rank coefficient (a non-parametric test), the correlation is $\tau = 0.4713$ ($p = 9.999 \times 10^{-5}$ with 10,000 permutations). Both parametric and non-parametric correlation tests show that the perceptual distances correlate positively with acoustic distances, at a strong and statistically significant level.

3.5.3.2 Structural similarity

To analyse the structural similarity of the acoustic and perceptual distances, we employ MDS, as described in Section 3.3.3.2.

MDS places each vowel at an optimal position relative to the rest of the vowels in a two-dimensional space. This is applied to the two distance matrices separately. Figure 3.17a and 3.17b show the relative positions of the vowels using acoustic distances and perceptual distances respectively. The acoustic distances can be projected perfectly into a two-dimensional space because it is originally based on two dimensions: two formants, F1 and F2. This means the visualisation using the acoustic distances can explain 100% of the variance in the original acoustic distances. Relative to the acoustic distances, the perceptual distances can explain 52% of variance using a two-dimensional space.



(a) Acoustic distance visualisation **(b)** Perceptual distance visualisation

Figure 3.17: Two-dimensional projection of the relative positions of American English vowels using their acoustic (a) and perceptual (b) distances (naturalistic). The visualisation in (a) explains 100% of the variance in the original acoustic distances, while the visualisation in (b) explains 52% of the variance in the original perceptual distances.

Both visualisations share strikingly similar structures with the IPA chart as shown

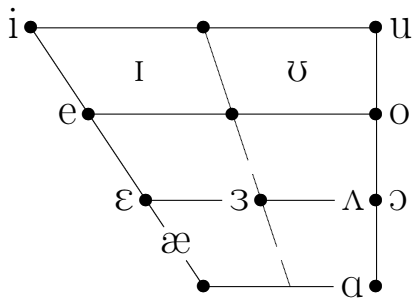


Table 3.7: IPA vowel chart containing twelve American English vowels [i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɜ]: only the nucleus portion is shown for diphthongs and long vowels.

in Figure 3.7. The acoustic visualisation has a stronger resemblance than the perceptual visualisation, which is not surprising as it is based on formant measurements. First of all, the perceptual visualisation has highly acceptable positions of four corner vowels [i, æ, ɑ, u]. Second of all, excluding the central vowels [ɜ, ʌ], the relative heights are perfectly projected for the front vowels with [i, ɪ, e, ɛ, æ], and the back vowels with [u, ʊ, o, ɔ, ɑ], ordered in decreasing height.

The perceptual visualisation nonetheless has a number of divergences from the IPA chart (and indeed the acoustic visualisation). Firstly, the most striking divergence is the position of [ɜ] relative to [ʌ], as it is both lower and backer than the expected. Secondly, [ʊ] is too far back, relative to [u]. Thirdly, [o] and [ɔ] are both lower than expected, especially relative to [ʌ]. Looking more closely, the overall divergence can be seen as [ɜ, ʊ, o, ɔ] having a more compact perceptual space "shrunk" to the bottom right. This overall divergence cannot be readily explained. The first possibility is that [ɜ] in rhotic accents often carries retroflexion, [ɜ˞], which is absent in the other vowels; therefore, the inclusion of [ɜ] could have an impact on the relative distances with the rest of the vowels beyond frontness and height. The second possibility is that [ɜ] could, in fact, be rounded in specific accents, and this feature was not captured in the transcription due to symbolic simplifications (see Section 2.2.7 for the simplifications of the vowel sets of various accents found in the corpus); therefore, the rounded [ɜ] is found to be perceptually closer to the other

rounded vowels [ʊ, o, ɔ]; this means that one of the two dimensions in the projection is not frontness, but a mixture of frontness and roundedness or simply roundedness, because front vowels in English are generally not rounded, whereas the back vowels are almost all rounded. Finally, there is a divergence with the pair /ɪ/ and /e/, such that the distance between /ɪ/ and /e/ is much closer acoustically than perceptually; however, I have no immediate explanation for this divergence.

3.5.4 Conclusion

This section computed the perceptual distances of American English vowels calculated from the perceptual confusability in the naturalistic corpus, and subsequently compared them with acoustic distances using two metrics.

Firstly, the global similarity between acoustic and perceptual distances were analysed using the Mantel test and there is a strong and significant relationship between the two ($r = 0.6831$ and $\tau = 0.4713$, $p < 0.0001$ with 10,000 permutations). Secondly, the structural similarity between two were analysed using MDS. By projecting distances into a two-dimensional plane, the relative positions of the vowels were shown. The projected plane with the perceptual distances bears a strong resemblance with both the acoustic space and the IPA chart in terms of their front/backness and height.

Together, this indicates that on both global and structural levels perceptual confusions in a naturalistic setting has a strong phonetic (acoustic) bias. This is a surprising finding, since the confusion matrix was extracted regardless of any contexts; for instance, the vowels could be from any words, and they could have any phonological environments (suprasegmental and segmental). Despite all the top-down influences (which are examined in Chapter 4), listeners still heavily rely on phonetic/acoustic similarity between vowels in everyday speech processing, whether it be successful or not.

3.6 Analysis of phonetic bias in consonant confusions

To establish the phonetic bias in the consonant misperception in the naturalistic corpus, it would be informative to compare them with distances that are phonetically based.

3.6.1 Featural distances

Unlike the vowel analyses, the phonetic relationship between consonants are complex and cannot be acoustically measured with only a few phonetic measurements such as formant values in the case of vowels, because consonants have multiple acoustic cues, and cues are different for each type of consonants; therefore, it is not easy to compare them directly. Alternatively, we could describe the phonetic properties of a consonant using distinctive feature systems, and use them to compute the relative similarity (and therefore distances) of consonants.

Hayes's feature set was used as the chosen feature system. This feature set was extracted from the software *Pheatures Spreadsheet* (van Vugt, Hayes, and Zuraw, 2012). The feature set contains the following features: [syllabic, stress, long, consonantal, sonorant, continuant, delayed release, approximant, tap, trill, nasal, voice, spread glottis, constricted glottis, labial, round, labiodental, coronal, anterior, distributed, strident, lateral, dorsal, high, low, front, back, tense]. It is worth noting that these features are primarily articulatory, and not acoustic nor perceptual. A feature set that is acoustically/perceptually grounded might be more appropriate for establishing a baseline of phonetic distances for comparing with perceptual distances.

The simplest way of computing phonetic distances would be to count the number of distinctive features a given pair of segments have in common – in other words, feature counting. However, it has been suggested that this is an empirically inad-

equate metric (Frisch, 1996). Alternatively, Frisch (1996) and Frisch, Broe, and Pierrehumbert (1997) proposed a natural class based metric of similarity between two phones. Using natural classes makes intuitive sense, because featural representation of phones were originally used to describe natural classes, which is a category consisting of a group of phones that plays a role in phonological patterns (Kenstowicz, 1994; Padgett, 2002). Frisch's metric is dependent on the phonological inventory (which phones exist in the language) and the structure of the inventory (as described by a given feature set). For a given pair of phones, the similarity is defined as the number of natural classes shared by the two phones, divided by the sum of the number of shared and unshared natural classes, as shown in Figure 3.18. To compute Frisch's similarity, the *Segmental similarity calculator* program by Albright (2006) was used with the following settings: 1) The type of natural class descriptions was chosen to be fully specified (rather than contrastive underspecification) and 2) The maximal superclass (the class that includes all known segments) was excluded. Frisch's similarity has been tested against behavioural data, such as English speech errors, phonotactic constraints in Arabic, and acceptability ratings of non-words in Arabic. Frisch found that this metric is superior to other feature-counting metrics across a range of behavioural data; however, it is unclear whether it would be equally superior for perceptual data. Beyond the empirical validity, Frisch's similarity has an advantage of being insensitive to feature redundancy. In sum, Frisch's similarity is an adequate, if not ideal, choice of similarity metric that uses feature sets.

$$Similarity = \frac{Shared\ \ Natural\ \ Classes}{Shared{+}Unshared\ \ Natural\ \ Classes}$$

Figure 3.18: Frisch's similarity (Frisch, 1996; Frisch, Broe, and Pierrehumbert, 1997)

For the present analysis, 26 phones were included: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ], excluding [j] and [w]. The reason for excluding the glides is that in the naturalistic corpus, glides are used as offglides of diphthongs, which means some of glides are consonantal and some are vocalic (in the

sense that they form part of a vowel). Given that our analysis focuses on consonants, these two ambiguous phones were excluded.

The computed similarity matrix of the 26 phones was then converted into distances using Shepard's distance metric (Figure 3.3).

3.6.2 Perceptual distances

To obtain perceptual distances of English consonants, I extracted a consonant confusion matrix from the naturalistic data. Only consonant to consonant aligned pairs were considered. I considered the consonants used in the transcription as shown in Figure 2.1, excluding [j] and [w] for the reason mentioned in the previous section (Section 3.6.1). In total, 26 phones were considered: [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ].

This confusion matrix (26 by 26) was then converted into a distance matrix using the procedure described in Section 3.3.1.

3.6.3 Comparison – featural and perceptual distances

3.6.3.1 Global similarity

To analyse the global similarity of the featural and perceptual distances, we employ the Mantel test as described in Section 3.3.2.1. With the Pearson coefficient, the correlation is $r = 0.2709$ ($p = 1.999 \times 10^{-4}$ with 10,000 permutations). With the Kendall rank coefficient, the correlation is $\tau = 0.2469$ ($p = 9.999 \times 10^{-5}$ with 10,000 permutations). Both parametric and non-parametric correlation tests show that the perceptual distances correlate positively with featural distances, at a modest and statistically significant level.

3.6.3.2 Structural similarity

To analyse the structural similarity of the featural and perceptual distances, we employ the agglomerative hierarchical clustering technique, as described in Section 3.3.3.1. This method seeks to merge phones hierarchically from bottom to top. The resultant "tree" reflects the hierarchical structure of the consonant distances. We apply all three common linkages (clustering strategies): Complete, Average and Single.

The hierarchical trees for featural and perceptual distances are visualised together in a single plot per linkage method. Figure 3.19, Figure 3.20 and Figure 3.21 are the plots with Complete, Average and Single linkages respectively. These plots are so-called tanglegrams, plotted using the R package *dendextendRcpp* (Galili, 2014). In each of these tanglegrams, the tree on the left is based on featural distances, and the tree on the right is based on perceptual distances. To best visualise the difference between the two trees, lines connecting the leaves (the individual phones) between the two trees are drawn in the middle. Furthermore, the edges of the branches that are unique to each tree are shown as dotted lines, thus highlighting the differences between the two trees.

Beginning with complete linkage (Figure 3.19), we will first examine the perceptual tree on the right. The two major clusters (the first split) of the tree are [ʃ, b, ð, dʒ, f, h, tʃ, v, ʒ, θ, p^h, t^h, k^h] and [ɹ, g, d, k, l, m, n, ŋ, p, s, t, z, ɾ]. It is clear that, on the whole, they are divided into phones with frication and those without, with the exceptions of [s, z, b]. Therefore, the first split of the tree indicates that frication (or perhaps the distinctive feature [±continuant]) plays a important role in perception. Interestingly, the aspirated voiceless stops are amongst the frication group, highlighting that the aspiration of the stops is, in fact, a form of frication, reinforcing the idea of distinguishing between the aspirated voiceless stops and the unaspirated/voiced stops in the data. The non-frication branch is then split into two more clusters further down the tree [ɹ, l, m, n, ŋ, s, z] and [g, d, k, p, t, ɾ],

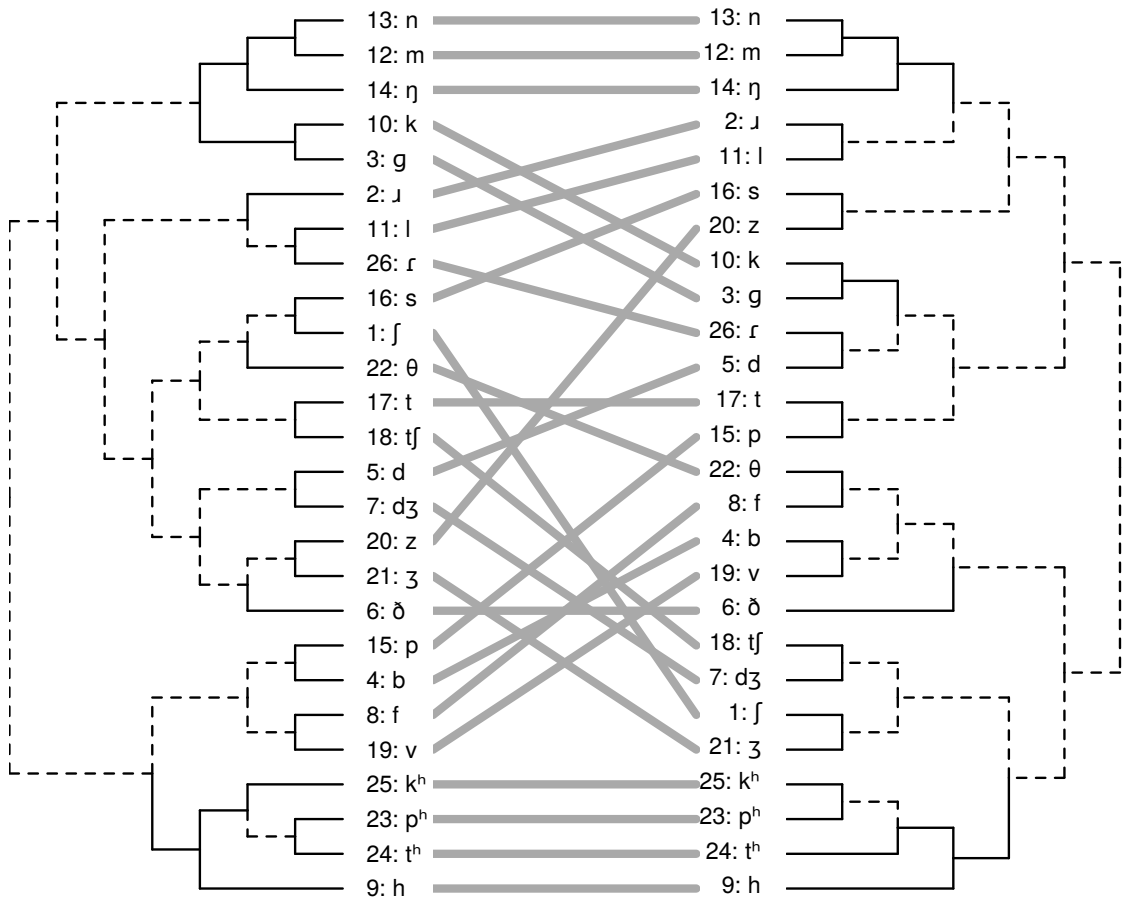


Figure 3.19: Hierarchical clustering of featural distances (left) and perceptual distances (right) with *Complete* linkage: distances are represented as trees; the lines in the middle are drawn to connect the leaves (the individual phones) between the two trees; the edges of the branches that are unique to each tree are shown as dotted lines.

which can be categorised as sonorants and non-sonorants, with the exception of [s, z]. Among the sonorant branch (ignoring [s, z]), there is a clear division between nasals [m, n, ŋ] and liquids [ɹ, l]. The further splits under the non-sonorant branch are not immediately interpretable. Switching to the frication branch, the immediate split creates two groups [b, ð, v, θ, f] and [ʃ, dʒ, h, tʃ, ʒ, p^h,t^h, k^h]. The first group contains "weak"/non-sibilant fricatives, while the second group contains affricates

and post-alveolar sibilants, as well as aspirated voiceless stops and glottal plosive. One interpretation of these two groups is duration (long and short) because, firstly, affricates and aspirated stop are more complex, both containing a plosive phone followed by a fricative phone and, secondly, sibilants [ʃ, ʒ] have longer noise duration than non-sibilants. Under the long duration branch, there is a further division between the affricates and sibilants, and the aspirated stops and [h], which can be interpreted as [± spread glottis]. Finally, all the affricates and sibilants (including [s, z], which clustered with the sonorant branch) are clustered together by [± voice] on the finest cluster level: [tʃ, dʒ], [ʃ, ʒ] and [s, z].

From examining the structure of the perceptual tree, it is clear that its clustering pattern is a function of phonetic similarities. We identified multiple phonetic dimensions, such as sonorant, spread glottis, voicing, frication, nasality, liquid, sibilancy, and duration. Interestingly, some of these dimensions were also identified in the classic experimental study by Miller and Nicely (1955). In Miller and Nicely (1955), they proposed and analysed five articulatory dimensions: voicing, nasality, affrication, duration and place of articulation. These dimensions were chosen in order to reasonably summarise the pattern of confusion in their data. Their affrication dimension is frication, distinguishing between [f, θ, s, ʃ, v, ð, z, ʒ] from the stops and nasals. Their *duration* dimension differs from ours; they classified [s, z, ʃ, ʒ] as being longer than other phones in their study, without including aspirated stops. Of the five dimensions proposed by Miller and Nicely (1955), our hierarchical clustering reflect at least three (*voicing, nasality, (af)frication*), and maybe four if we were to include *duration*, which has a different definition from ours. Finally, place of articulation was found to be a poor dimension in Miller and Nicely (1955), and this is also reflected by its absence in the hierarchical structure.

In comparison to the featural tree, only three sub-branches have a near-direct correspondence: the nasal branch [m, n, ŋ], the liquid branch [ɹ, l] and the [+ spread

glottis] branch [h, p^h,t^h, k^h]. To better quantify the similarity between the featural tree and the perceptual tree, multiple correlation analyses are applied. First two variants of cophenetic correlation test are applied. With the Pearson coefficient, the correlation is $r = 0.1149$ ($p = 0.036$ with 1,000 permutations). With the Kendall rank coefficient, the correlation is $\tau = 0.1286$ ($p = 0.02$ with 1,000 permutations). Furthermore, we apply Baker's Gamma Index, which has the value $r = 0.1560$ ($p = 0.013$ with 1,000 permutations). All three correlation measures (parametric cophenetic correlation, non-parametric cophenetic correlation and Baker's Gamma) show that the hierarchical tree of perceptual distances correlates positively with the tree of featural distances, at a modest and statistically significant level.

Moving onto the trees with average linkage (Figure 3.20), we will again focus on the perceptual tree on the right. The structure of the tree is nearly identical to that with complete linkage (Figure 3.19) with one major divergence of the branch containing [g, d, k, p, t, ɾ]. This branch was clustered with the sonorant branch in the tree with complete linkage, whereas in the current tree, it is clustered with [b, v, θ, f]. To assess the similarity between the two perceptual trees (complete and average), I again applied the three correlation measures, which yielded the following correlations: cophenetic (Pearson) $r = 0.5561$ ($p = 0$ with 1,000 permutations), cophenetic (Kendall rank) $\tau = 0.4518$ ($p = 0$ with 1,000 permutations), and Baker's Gamma Index, $r = 0.5062$ ($p = 0$ with 1,000 permutations). The correlation measures all indicate that there is a positive correlation between the two linkages, at a strong and statistically significant level. The similarity between the trees with the two linkages, complete and average, indicates that the resultant structure is relatively stable.

Comparing the perceptual tree to the featural tree with average linkage, we see that only one sub-branch has a direct correspondence, which is the nasal branch. To quantify the similarity between the featural tree and the perceptual tree with

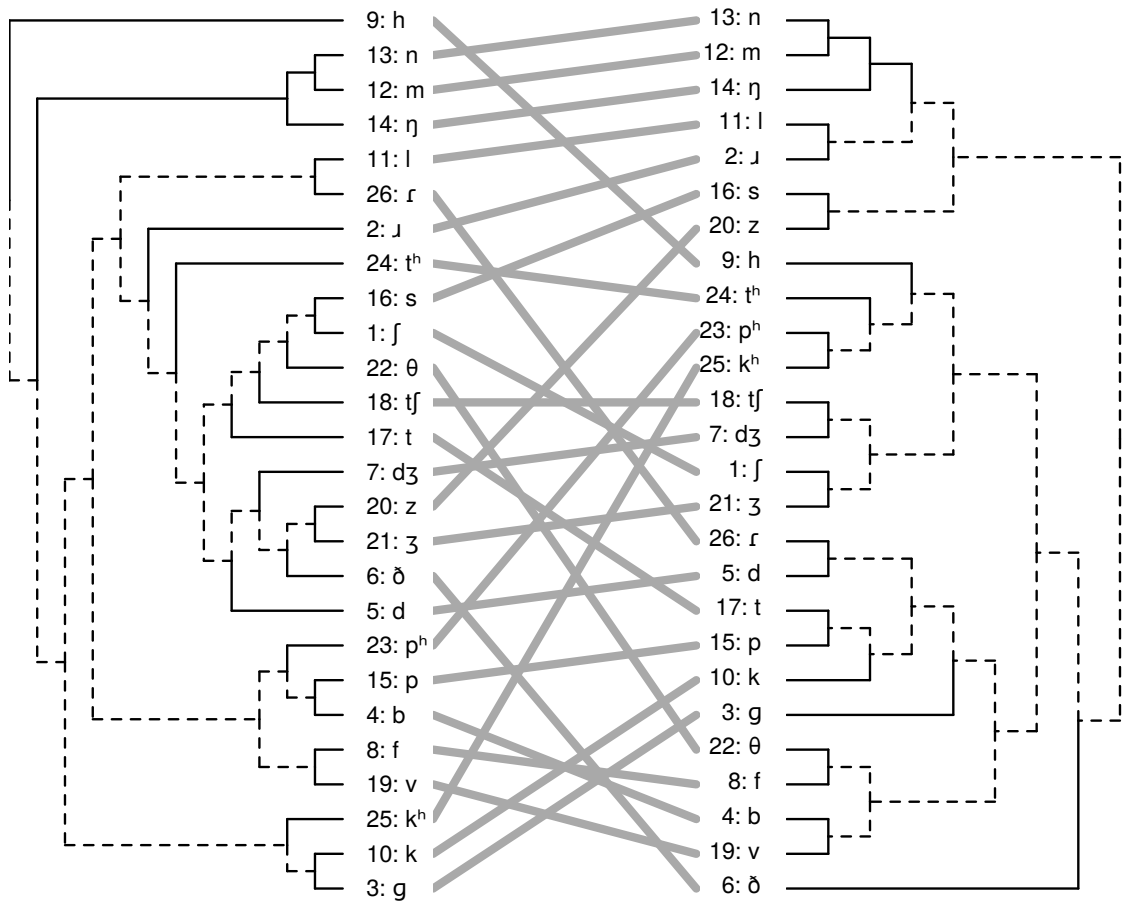


Figure 3.20: Hierarchical clustering of featural distances (left) and perceptual distances (right) with *Average* linkage: the lines in the middle are drawn to connect the leaves (the individual phones) between the two trees; the edges of the branches that are unique to each tree are shown as dotted lines.

average linkage, we applied multiple correlation analyses, which yielded the following correlations: cophenetic (Pearson) $r = 0.1439$ ($p = 0.047$ with 1,000 permutations), cophenetic (Kendall rank) $\tau = 0.1284$ ($p = 0.077$ with 1,000 permutations), and Baker's Gamma Index, $r = 0.1658$ ($p = 0.074$ with 1,000 permutations). All three measures show that the hierarchical tree of perceptual distances correlates positively with the tree of featural distances, at a modest and statistically near-significant level.

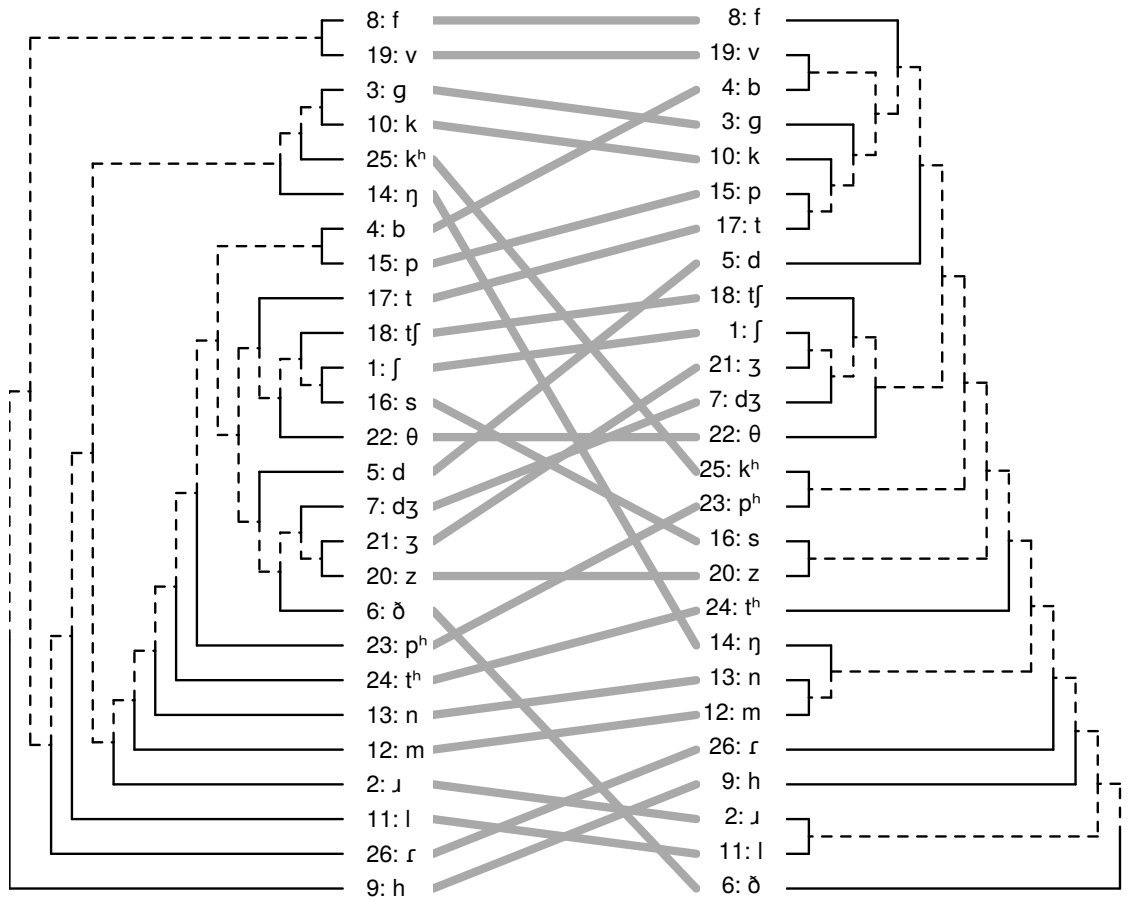


Figure 3.21: Hierarchical clustering of featural distances (left) and perceptual distances (right) with *Single* linkage: the lines in the middle are drawn to connect the leaves (the individual phones) between the two trees; the edges of the branches that are unique to each tree are shown as dotted lines.

Finally, we analyse the trees with single linkage. Focusing on the perceptual tree, it is immediately clear that it has a very different structure compared to those with complete and average linkages. Concretely, the larger clusters (from the top of the tree) have no obvious interpretation, since each split created an imbalanced number of phones per cluster: e.g. at the first split, one cluster contains only [ð], whereas the other cluster contains the rest of the phones. Going further down the tree, we can

identify a nasal cluster [m, n, ŋ], a liquid cluster [ɹ, l] and a large cluster grouped by manner [f, v, b, g, k, p, t, d], which contains essentially all stops (except for [f, v]).

Visually, the perceptual tree has no direct correspondence with the featural tree. To better quantify the similarity between the featural tree and the perceptual tree with single linkage, we applied multiple correlation analyses, which yielded the following correlations: cophenetic (Pearson) $r = 0.3046$ ($p = 0.026$ with 1,000 permutations), cophenetic (Kendall rank) $\tau = 0.2303$ ($p = 0.048$ with 1,000 permutations), and Baker's Gamma Index, $r = 0.3007$ ($p = 0.04$ with 1,000 permutations). All three measures show that the hierarchical tree of perceptual distances correlates positively with the tree of featural distances, at a modest/moderate and statistically significant level.

All three linkages were able to extract clusters of segments that resemble natural classes from the perceptual distances. The *complete* and *average* linkages were able to extract natural classes at a higher level of the tree (closer to the root) (e.g. sonorants), while the *single* linkage was only able to extract natural classes at a lower level of the tree (closer to the leaves) (e.g. nasals), and the higher levels of the tree were hard to interpret in terms of natural classes. However, the *single* linkage managed to yield the highest correlation between the tree of perceptual distances and the tree of featural distances, compared to the *complete* and *average* linkages. Therefore, all three linkages were useful for revealing hierarchical structures in the data.

Interestingly, the hierarchical structures resemble the contrastive hierarchy by Dresher (2008). In our hierarchical structures, starting from the root, each split can be seen as a feature being used to *contrast* the two sets of phones under its leaves. Therefore, the order of the splits is essentially the order of features being used to contrast the entire set of phones, such as the contrastive hierarchy. However, it is worth noting that not every split in our hierarchical structures can be perfectly interpreted in terms of distinctive features. For instance, a split created two sets of

phones, [n,m,ŋ,p] and [b,d,g,k,t], and [p] is an outlier in the assumed nasal group. In fact, the structures projected with the *single* linkage (Figure 3.21) suffer from this problem. Therefore, it can be difficult to directly interpret the order of the splits in terms of the contrastive hierarchy. One potential solution is to compute a score for how good a given distinctive feature is at distinguishing the two sets of phones at each split, e.g. a classification score (F-score). Concretely, at each split, there would be a set of classification scores, one for each distinctive feature; and the distinctive feature that has the highest classification score is therefore the most representative distinctive feature for that split. All of the most representative distinctive features are then ordered by the order of the splits. This could then be interpreted as a contrastive hierarchy of the perceptual structure.

3.6.4 Conclusion

This section computed the perceptual distances of American English consonants calculated from the confusion data in the naturalistic corpus and subsequently compared them with featural distances on both global and structrual levels.

Firstly, the global similarity between featural and perceptual distances was analysed using the Mantel test and there is a modest and significant relationship between the two ($r = 0.2709$ and $\tau = 0.2469$, with $p < 0.0001$ with 10,000 permutations). Secondly, the structural similarity between the two was analysed using Hierarchical Clustering. By merging phones from the bottom to the top, using three different common linkages, different hierarchical structures of the consonants were discovered. Particularly with complete and average linkages, the structures of the perceptual distances revealed multiple phonetic dimensions, such as sonorant, spread glottis, voicing, frication, nasality, liquid, sibilancy and duration; and the two structures using complete and average linkages were extremely similar (and indeed strongly correlated, $r/\tau = 0.4$ to 0.6, $p = 0$), suggesting that it is a stable structure. Further-

more, the phonetic dimensions identified in the perceptual tree matched those found in the experimental confusion study by Miller and Nicely (1955), thus highlighting that phonetic biases that exist in consonant confusions in experimental settings are robust, and can be found in naturalistic settings. Different correlation measures (Cophenetic and Baker's gamma) were applied to compare between the hierarchical structures based on featural distances and perceptual distances. Overall, the correlations were positive, at a moderate/modest and statistically significant level (r/τ = 0.1 to 0.3).

At both global and structural levels, there are significant similarities between the featural and perceptual distances. Furthermore, individual examinations of the hierarchical structures based on perceptual distances reflected clear phonetic dimensions. However, the strength of the similarity was only at a modest level. This is surprisingly poor compared to the strong similarity found in the comparisons between acoustic and perceptual distances in the vowel analyses (Section 3.5.3). Multiple explanations can be put forth for this poor similarity with the featural distances.

Firstly, the featural distances were computed using primarily articulatory features which do not reflect perceptual dimensions; therefore, the poor correlation between articulatory-based featural distances and perceptual distances is expected. In fact, the question of whether features grounded in one domain – be it articulation, acoustics or perception – can capture meaningful structures in another domain is related to the correlates between features grounded in different domains. The feature relationships between domains have been previously examined, for instance, by Fant (1962), Delattre (1967), Stevens (1972), Geumann (2001), and Stevens (2002). The overall conclusion is that correlates of the features across domains are extremely complex and the relationships are many-to-many, rather than simply one-to-one. These complex many-to-many relationships between correlates of features in different domains suggest that none of the established feature sets contain features that can only be

defined in one domain. Concretely, in a comparative study of consonant feature system in speech errors (speech misproducton) by Broecke and Goldstein (1980), several established feature sets were compared. The feature sets compared were defined primarily in articulation terms (e.g. Chomsky and Halle (1968) and Ladefoged (1975)), and primarily in acoustical terms (e.g. Jakobson, Fant, and Halle (1952)), in terms of their ability to capture speech error patterns. They attempted to answer the question of whether domain specific feature sets (in their case, articulatory-based ones) are better than those of other domains. However, they found that the feature sets based on different domains are all capable of capturing consonantal structures in speech errors, and crucially, none of the feature sets examined were superior. Broecke and Goldstein (1980) concluded that to truly find a preference for a particular domain in speech errors, one would have to compare feature sets that are grounded *purely* in one domain, such as acoustically based ones used by speech synthesisers. This comparative study on speech errors therefore suggests that our choice of established feature sets should not have an effect on the level of similarity and that even feature sets that are based primarily on perceptual terms should do equally well/poorly because they all contain correlates of features between them. Furthermore the ideal feature set should be purely grounded in one domain (in our case, it should be acoustics or perception).

Secondly, the poor similarity could suggest that phonetic biases in consonant confusions in perception are relatively small compared to those in vowel confusions. Let us consider consonant confusions modelled as an equation. Consonant confusions can be seen as a function of phonetic biases, but they carry relatively less weight compared to other terms/components in the equation. Furthermore, several speculations can be made. Firstly, this could suggest that there are more components in the equation of consonant confusions than that of vowel confusions; in other words, consonant confusions are more complex. Secondly, consonant confusions are more

susceptible to non-phonetic factors, such as top-down factors (e.g. lexical selections), than vowel confusions; this speculation is confirmed in Chapter 4, Section 4.2.

3.7 Analysis of ecological validity

This section attempts to address the ecological validity of two experimental conditions that are often manipulated in order to induce perceptual confusions. The first condition is the signal-to-noise ratio (henceforth SNR), which is the ratio between the amount of noise and the amount of the signal. The second condition is bandwidth filtering. Firstly, I compare the different experimental controls and outline the experimental misperception corpora of English that were used to address the two main experimental conditions described in Section 3.7.1. Secondly, I document the method for assessing the ecological validity of experimental conditions in Section 3.7.2. Thirdly, the analyses of the two conditions are reported in Section 3.7.3 and Section 3.7.4. Finally, I conclude the findings in Section 3.7.5.

3.7.1 Experimental English corpora

Three experimental corpora of speech misperception errors of American English are analysed: Miller and Nicely (1955), Wang and Bilger (1973), Cutler et al. (2004), and Phatak and Allen (2007). These corpora were chosen for their diversity, primarily in the unit of the stimuli. Concretely, the stimuli have different levels of complexity across these four studies. In Miller and Nicely (1955), the unit stimuli were in nonsense CV syllables with the same vowel /a/, with the 16 different consonants. Subsequently, Wang and Bilger (1973) presented a similar study with the full consonant inventory (24 phonemes) of English (c.f. only 16 in Miller and Nicely (1955)). They embedded the consonants in both CV and VC nonsense syllables, as opposed to just CV. Furthermore, they used three corner vowels /a/,/i/ and /u/ in

these syllables; therefore, they covered a wider range of syllables than Miller and Nicely (1955). Cutler et al. (2004) went one step further by testing all phonologically permissible CV and VC syllables in English, thus allowing us to examine richer contextual effects on errors. Finally, Phatak and Allen (2007) tested 64 different CV syllables (16 consonants and four vowels /ɑː, ɪ, æ, eɪ/ with a close-set response task with 64 possible responses, unlike Cutler et al. (2004) which restricted the experiment to either only consonant confusions or vowel confusions by allowing only consonant or vowel responses.

Beside the unit of the stimuli, ranging from a nonsense CV syllable with only one vowel to all possible CV and VC (all vowels and consonants), these studies are different in a number of ways: a) the kind of noise manipulations (masking with different signal-to-ratio ratios), b) frequency band filtering with different cut-off levels and c) noise type. For an overview of the four corpora, please see Table 3.8. A comparison of the different experimental controls used by these studies will be made in the first section (Section 3.7.1.1). The details of the studies are then explained in subsequent sections.

Source	Unit of stimuli	Seg. Error	Syllable Types	Cons:Vowel	Noise Type	SNR(dB)	Speakers: Listeners
Miller and Nicely (1955)	CV	C	16	16:1	White	-18 to +12	5:5
Wang and Bilger (1973)	CV, VC	C	129	24:3	White	-10 to +15 Quiet	1:16
Cutler et al. (2004)	CV, VC	C,V	645	24:15	six-talker babble	0 to 18	1:16
Phatak and Allen (2007)	CV	C,V	64	16:4	Speech shaped	-22 to -2 Quiet	14:32

Table 3.8: An overview of the four experimental corpora for English: the "Source" column indicates the name of the experimental corpora; the "Syllable Types" column indicates the syllable type tested by each study; the "Seg. Error" column represents whether each study tested the confusion of consonants (C) or vowels (V), or both (C, V); the "Noise Type" column represents the noise type used by each study to mask their stimuli, each as white noise, six-talker babble noise, and speech-shaped noise; the "SNR(dB)" column represents the SNR levels of the stimuli and Quiet means no noise was added; the "Speakers:Listeners" column represents the number of speakers the stimuli were produced by and the number of listeners tested; and the notation $x : y$ denotes the number of speakers x and the number of listeners y.

3.7.1.1 Comparison of experimental controls

This section compares the different experimental controls (as summarised in Table 3.8) in the above experimental studies, focusing on the noise type, the frequency bandwidth, the number of speakers and listeners, and the number of syllables/consonants/vowels.

3.7.1.1.1 Noise type Noise in the context of being a masker can be defined as sounds that are not the voice of the person we are trying to hear. Noise has two main effects on the speech signal, which are *energetic masking* (Pollack, 1975) and *informational masking* (Pollack, 1975; Watson, Kelly, and Wroton, 1976; Freyman et al., 1999). Energetic masking is when the noise interferes with the speech signal in the acoustic environment (Lidestam, Holgersson, and Moradi, 2014). Informational masking is when noise interferes with the speech signal in the perceptual process (Lidestam, Holgersson, and Moradi, 2014); and more broadly, it can also be defined as any masking that cannot be attributed to energetic masking.

The corpora mentioned above used different noise maskers: white noise in Miller and Nicely (1955) and Wang and Bilger (1973), six-talker babble noise in Cutler et al. (2004), and speech-shaped noise in Phatak and Allen (2007). The differences between these noise maskers are discussed below.

White noise is a stationary noise, and it is regarded as an energetic masker. It masks the entire frequency spectrum of human hearing in equal amounts. However, human perception of volume is logarithmic (doubling the frequency is perceived as doubling the volume); therefore, it masks high frequency content more than low frequency content. In other words, it is particularly effective in masking higher formants and frication noise.

Multi-talker babble noise masks the speech signal by means of adding competing speech signals. It is basically the listening condition in the cocktail party problem

(Cherry, 1953). Babble noise is less effective as an energetic masker than white noise (Festen and Plomp, 1990; Simpson and Cooke, 2005), but it is an effective informational masker. The amount of informational masking and energetic masking of babble noise is a function of how many competing speech signals are added to the original speech signal (i.e. the number of competing talkers). Informational masking on the linguistic component is most effective when the number of talkers is two to three (Carhart, Johnson, and Goodman, 1975; Freyman, Balakrishnan, and Helfer, 2004), and the effect is almost absent when the number of of talkers is higher than ten (Freyman, Balakrishnan, and Helfer, 2004); while energetic masking increases with the number of talkers because as the number of talkers increases, the babble becomes less speech-like and more noise-like.

Speech-shaped noise is when there are infinite competing talkers. It has the spectrum that approximates the average long term spectrum of speech. Therefore, similar to multi-talker babble noise, it masks frequency regions that are most speech-relevant.

In terms of their relative effect on masking the speech signal, speech-shaped noise tends to mask low frequency regions more than white noise, and white noise tends to mask high frequency regions more than speech-shaped noise. At a given SNR level, four to eight talker babble noise tends to mask consonants more than speech-shaped noise (Simpson and Cooke, 2005). In any case, their relative masking effect depends on the other experimental controls, such as the stimuli (e.g. if the stimuli contain mostly fricative consonants, then white noise would lower the overall accuracy more than speech-shaped noise would).

In terms of ecological validity, white noise is arguably less ecological than multi-talker babble noise and speech-shaped noise, because the latter two are more speech-like and resemble real life situations (e.g. in a noisy cocktail party with multiple people talking at the same time). Crucially, multi-talker babble noise and speech-

shaped noise mask speech relevant frequency regions as opposed to the entire frequency spectrum. Therefore, one might expect that Cutler et al. (2004) and Phatak and Allen (2007) are more ecologically valid than Miller and Nicely (1955) and Wang and Bilger (1973) in terms of the masker.

3.7.1.1.2 Frequency bandwidth Of the four studies, only Miller and Nicely (1955) examined the effect of bandwidth filtering. By systematically varying the cut-off upper or lower band, the authors examined the effect of low-pass filtering and high-pass filtering. Using transinformation analysis (Attneave, 1959; Shepard, 1972), the amount of information successfully transferred from the signal to the listeners' system was computed using the confusion matrices. Five phonetic features were examined: voicing, nasality, (af)frication, duration, and place of articulation. It is worth noting that the duration feature is to distinguish between [s,ʃ,z,ʒ] (which are classified as long) and the other consonants; therefore, it is essentially a sibilancy feature; and the affrication feature is to distinguish fricatives from non-fricatives. The differences between low-pass filtering and high-pass filtering are discussed below.

Low-pass filtering filters the high frequency components. The authors found that it reduces the amount of transferred information more severely for place, followed by duration (i.e. sibilancy), then affrication (i.e. frication), than other features. That is, the voicing, and nasality are robust in low-pass filtering.

This relative robustness is apparent by comparing the information transferred between the widest bandwidth and the narrowest bandwidth. With SNR being fixed at +12dB, the amount of place information is 1.090 bits with the widest bandwidth (200–6500 Hz), and the information reduces to 0.025 bits with the narrowest bandwidth (200–300 Hz); this is a 98% reduction in information. Similarly, duration (i.e. sibilancy) has 0.751 bits with the widest bandwidth and 0.042 bits with the narrowest bandwidth: a 94% reduction. Affrication (i.e. frication) has 0.853 bits with the widest bandwidth and 0.159 bits with the narrowest bandwidth: a 81%

reduction. Nasality has 0.555 bits with the widest bandwidth and 0.371 with the narrowest bandwidth: a 33% reduction. Finally, voicing has 0.956 bits with the widest bandwidth and 0.623 bits with the narrowest bandwidth: a 34% reduction. In sum, voicing and nasality are three times more robust than place, affrication and duration.

This pattern is to be expected, with place, affrication and duration being particularly vulnerable to low-pass filtering. Low-pass filtering removes the high frequency components. The affrication and duration features cover all the fricatives, of which the primary phonetic cues lie in the high frequency regions (Wright, 2004). Similarly, the main place cues of fricatives lie within the spectrum of the frication noise; one of the place cues of stops is the release bursts, which consist of short intervals of frication noise (Wright, 2004). Therefore, the place cues of fricatives and stops are mostly absent due to low-pass filtering.

Let us move on to high-pass filtering. High-pass filtering filters the low frequency components. Miller and Nicely (1955) found that it reduces the information transferred almost equally for all features, with duration being the least reduced. The robustness of duration (i.e. sibilancy) can be explained by the fact that the sibilants are characterized by their high frequency energy, and since high-pass filtering only removes low frequency components, the sibilants are largely unaffected. The reason for why all the other features (i.e. non sibilants) are equally affected by high-pass filtering is that it removes most of the acoustic information of the consonants. Therefore, the consonants are no longer audible, and the listeners have to guess. Consequently, the confusion patterns are equally random. Comparing it to low-pass filtering, we can see that low-pass filtering affects the linguistic features differently, and the removal of high frequency components still leaves the consonants audible because low frequency components contain most of the acoustic information.

In sum, low-pass filtering particularly masks fricatives, while high-pass filtering

masks all consonants but the sibilants. Low-pass filtering affects linguistic features (which generates non-random confusions) differently while high-pass filtering affects them equally (which generates random confusions). Therefore, low-pass filtering is a more meaningful manipulation than high-pass filtering, and therefore likely to be more ecologically valid.

In addition, Miller and Nicely (1955) also observed that the pattern of information transferred is similar between the manipulation of low-pass filtering and that of SNR levels. The reason given for this correspondence is that white noise masks high frequency components of speech more than low frequency components, and this is effectively what low-pass filtering does; low-pass filtering removes the high frequency components. This again reinforces the idea that low-pass filtering is a useful and ecologically valid manipulation.

3.7.1.1.3 The number of speakers and listeners The number of speakers and listeners was different across the four studies. Miller and Nicely (1955) used five speakers to produce the stimuli and tested more listeners. Wang and Bilger (1973) and Cutler et al. (2004) used only one speaker, but three times more listeners (16 listeners) than Miller and Nicely (1955). Phatak and Allen (2007) tested more speakers and listeners than the other three studies, with 14 speakers and 32 listeners.

In terms of ecological validity, one could expect that with more speakers and listeners tested, the confusion results can be better generalised to the population. With a small number of listeners or speakers, the confusion patterns can be skewed by individual differences. Recent work has shown that speech perception is a function of individual differences (Yu et al., 2011; Yu, 2013), such as sex and autistic traits. Therefore, of the four studies, Phatak and Allen (2007) is expected to be the most ecologically valid (and therefore reliable) study.

3.7.1.1.4 The number of syllables/consonants/vowels

The number of syllables/consonant/vowels tested was different for each study. Miller and Nicely (1955) tested 16 consonants preceding one vowel in a CV syllable; therefore, 16 syllable types were tested. Wang and Bilger (1973) tested 24 consonants with three vowels in both CV and VC syllables, and this amounts to 129 syllable types. Cutler et al. (2004) was the most comprehensive study in terms of the number of syllables/consonants/vowels tested; all possible CV and VC syllables (645 syllables) were tested which covers 24 consonants and 15 vowels. Phatak and Allen (2007) tested 16 consonants preceding four vowels, which accounts to 64 syllable types. Therefore, Cutler et al. (2004) is expected to be the most ecologically valid study in terms of the coverage of syllables/consonants/vowels.

3.7.1.2 Miller and Nicely (1955)

Miller and Nicely (1955) examined 16 English consonants: approximately three quarters of the consonants and 40 percent of all phonemes. The 16 consonants are [p], [t], [k], [f], [θ], [s], [ʃ], [b], [d], [g], [v], [ð], [z], [ʒ], [m] and [n]. Five female Americans acted as both the listeners and the speakers. They reported that the recordings did not have any noticeable dialect. These consonants were embedded in a nonsense CV syllable with the /a/ vowel in *father*, which I assumed to be [ɑː].

The stimuli were frequency distorted by applying low-pass and high-pass filters with different cut-off levels and masked with noise at different signal-to-noise ratios. Concretely, three kinds of stimuli manipulations were employed. Firstly, they masked the stimuli with white noise with different signal-to-noise ratios of the following values, -18, -12, -6, +0, +6 and +12 dB, while keeping the full bandwidth 200–6,500 Hz. Secondly, the upper frequency band was manipulated with the values (in Hz) 300, 400, 600, 1,200, 2,500 and 5,000; the lower frequency band was kept at 200 Hz and the signal-to-noise ratio was kept at +12 dB. Thirdly, the lower frequency

Signal-to-Noise Ratio (dB)	Lower band (Hz)	Upper band (Hz)
-18	200	6,500
-12	200	6,500
-6	200	6,500
0	200	6,500
+6	200	6,500
+12	200	6,500
+12	200	300
+12	200	400
+12	200	600
+12	200	1,200
+12	200	2,500
+12	200	5,000
+12	1,000	5,000
+12	2,000	5,000
+12	2,500	5,000
+12	3,000	5,000
+12	4,500	5,000

Table 3.9: 17 conditions tested by Miller and Nicely (1955) of different Signal-to-Noise Ratios (dB), lower bands (Hz) and the upper bands (Hz)

band was manipulated with the values (in Hz) 1,000, 2,000, 2,500, 3,000 and 4,500; the upper frequency band was kept at 5,000 Hz, and the signal-to-noise ratio was kept at +12 dB. In total, 17 confusion matrices were extracted from the study (see Table 3.9 for a summary of the 17 conditions). There were 4,000 observations at each condition, with each syllable judged 250 times under every condition, making 68,000 trials in total.

3.7.1.3 Wang and Bilger (1973)

Wang and Bilger (1973) examined 24 consonants embedded in a CV syllable, and 19 consonants embedded in a VC syllable. Due to technical constraints on the number of responses, they created two sets of consonants in a CV syllable, covering 24 consonants with overlapped consonants between the two sets, and two sets in a VC syllable, covering 19 consonants with overlapped consonants between the two sets

(see Table 3.10 for a detailed breakdown of phonemes in each of the four syllable sets.) Unlike Miller and Nicely (1955), this study embedded the consonants with three vowels [ɑː],[iː] and [uː] and tested all phonologically permissible CV and VC syllables, which makes 129 syllables. The stimuli were recorded by one male adult speaker who was presumably an American. The listeners were six males and ten females, who were presumably Americans. The task was forced-choice, with 16 possible responses. Three kinds of stimuli manipulations were performed. Firstly, the signal-to-noise ratio was varied with the following values in dB: -10, -5, 0, +5, +10 and +15. Secondly, for each signal-to-noise ratio the signal level was varied with the following values in dB SPL: 50, 65, 80 and 95. Thirdly, the signal level was varied *without* masking noise, with the signal levels ranging from 20 to 45 dB SPL in 5-dB steps, and from 55 to 115 dB SPL in 10-dB steps.

Syllable Set	Consonant phonemes
CV-1	[p], [t], [k], [b], [d], [g], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ]
VC-1	[p], [t], [k], [b], [d], [g], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ]
CV-2	[p], [t], [tʃ], [dʒ], [l], [ɹ], [f], [s], [v], [m], [n], [h], [hʷ], [w], [j]
VC-2	[p], [t], [g], [ŋ], [m], [n], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ]

Table 3.10: 4 syllable sets tested by Wang and Bilger (1973): CV-1 and VC-1 have the same set of consonants; CV-2 and VC-2 contains consonants that are not in CV-1 and VC-1.

Syllable Set	Signal-to-Noise Ratio (dB)	Signal Levels (dB SPL)
CV-1 VC-1 CV-2 VC-2	-10, -5, 0, +5, +10, +15	50, 65, 80, 95
CV-1 VC-1 CV-2 VC-2	N/A	20 to 45 dB SPL (in 5-dB steps) and 55 to 115 dB SPL (in 10-dB steps)

Table 3.11: The conditions of the eight confusion matrices published in Wang and Bilger (1973) of different syllable types, Signal-to-Noise Ratio (dB), and Signal Levels (dB SPL)

The published data were eight confusion matrices. They consist of the four syllable sets summed over all signal-to-noise ratios and all signal levels, and the four syllable sets summed over all signal levels without any masking noise. (See Table 3.11 for a summary of the conditions for the eight matrices available).

3.7.1.4 Cutler et al. (2004)

Cutler et al. (2004) tested all possible standard American English CV and VC sequences using 24 consonants and 15 vowels (excluding schwa). /ŋ/ and /ʒ/ were excluded in the CV syllables. /h/, /w/ and /j/ were excluded in the VC syllables. Although /ʒ/ is a possible onset in American English, it was not tested, perhaps because it is a relatively rare onset. In total, there were 645 syllables. The stimuli manipulations involved adding multi-talker babble noise with the following signal-to-noise ratios: 0, 8 and 16 dB. The experiments had two sessions: a consonant session and a vowel session. In the consonant session, the participants listened to these nonsense syllables and were then asked to identify only the consonant and not the vowel. Similarly, in the vowel session, they had to identify only the vowel and not the consonant. The response type was forced-choice. The participants were 16 native American listeners (and 16 native Dutch listeners, but their data were not considered in this thesis). Each of them completed both the consonant and vowel sessions, with 1,935 trials per session, making 3,870 trials per participant, 96 trials per syllable and 61,920 trials overall. The complete data were published online, with which various confusion matrices could be computed.

3.7.1.5 Phatak and Allen (2007)

Phatak and Allen (2007) tested 64 CV syllables. The 64 syllables contain 16 consonants [p], [t], [k], [f], [θ], [s], [ʃ], [b], [d], [g], [v], [ð], [z], [ʒ], [m] and [n] – the same set as Miller and Nicely (1955), and four vowels [ɑː, ɪ, æ, eɪ].

The stimuli manipulations involved adding speech-shaped noise with the following signal-to-noise ratios: -22, -20, -16, -10 and -2dB as well as in *Quiet* (without any masking noise). A speech-shaped noise has a long-term average spectrum of speech signals, and it would mask a speech signal uniformly across frequencies, while white noise (as used by Miller and Nicely (1955) and Wang and Bilger (1973)) would mask high frequencies more than low frequencies.

Each CV syllable was spoken by 14 native speakers of American English, and 14 listeners were recruited. All listeners are native speakers of English, in which ten have American accents, and one has a Nigerian accent. The response type was forced-choice, with 64 possible syllables as well as a *Noise* option (if the listener only heard noise). In total, 5,376 trials (16 consonants $\times$ 4 vowels $\times$ 14 speakers $\times$ 6 SNR levels (including Quiet)) were presented to each participant. The experiment was split into 42 tests, each with 128 sounds. On average, the experiment took 15 hours to complete by each listener. A subsequent study by Singh and Allen (2012) recruited additional listeners, and the details are reported in Toscano and Allen (2014). The data was kindly made available to me by Prof. Jont B. Allen. The final dataset contained 101,760 completed trials by 32 listeners.

3.7.2 Method

Three experimental English corpora were used as the reference data sets (see Section 3.7.1 for details): Miller and Nicely (1955), Wang and Bilger (1973), and Cutler et al. (2004). In order to identify the ecological validity of experimental conditions, the naturalistic confusion matrices are taken as the baseline; that is, they are most ecologically valid. They are then systematically compared with experimental confusion matrices of various experimental conditions. Using correlation tests, experimental conditions that correlate most with the naturalistic matrices would be the most ecologically valid.

3.7.2.1 Pre-processing

A few adjustments were made to the representation of the phones in the experimental corpora in order to match those in the naturalistic corpus. Firstly, all voiceless stops [p, t, k] in CV conditions were treated as aspirated voiceless stop [p^h, t^h, k^h], but not in VC conditions. This adjustment was therefore applied to all the [p, t, k] in the confusion matrices of Miller and Nicely (1955) (since they only tested CV), the two CV syllable sets in Wang and Bilger (1973), and the CV condition in Cutler et al. (2004). Secondly, in the vowel confusion data in Cutler et al. (2004), we considered only the starting point of the diphthongs: [aʊ] and [aɪ] as [a], [ɔ] as [ɔ], [oʊ] as [o] and [eɪ] as [e]. Similarly, all long vowels were treated as short [iː] as [i], [ɑː] as [ɑ], [ɔː] as [ɔ] and [uː] as [u]. Finally, the NURSE vowel was treated as [ɜ].

3.7.2.2 Extraction of matrices

3.7.2.2.1 Naturalistic confusions Two matrices from the naturalistic corpus were extracted. They are the same matrices in previous vowel analyses (Section 3.5.2) and consonant analyses (Section 3.6.2).

The full vowel confusion matrix was first extracted, and it contains 16 phones [i, ɪ, e, ɛ, æ, a, ɑ, ɒ, ɔ, o, u, ʉ, ɜ, ʌ, ʊ, ə]. I then excluded [ʉ] and [ɒ] from the matrix as they are not part of the General American accent. This confusion matrix (14 by 14) was then converted into a distance matrix using the procedure described in Section 3.3.1.

The consonant matrix was extracted and it contains 28 phones [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ, j, w]. I then excluded [j] and [w] for the reason mentioned in the Section 3.6.1. This confusion matrix (26 by 26) was then converted into a distance matrix using the procedure described in Section 3.3.1.

3.7.2.2.2 Experimental confusions Different experimental confusion matrices were extracted for the noise level analyses and the bandwidth analyses.

For the noise level analyses of the consonants, six matrices of Miller and Nicely (1955) were extracted from the publication; they differ in terms of their SNRs, with the following values: -18, -12, -6, +0, +6 and +12 dB, all with the full bandwidth 200–6,500 Hz.

Nine matrices were extracted from Cutler et al. (2004) with three syllable conditions (CV, VC, and CV + VC), and three SNR levels (0, +8, +16).

Fourteen matrices were extracted by Wang and Bilger (1973) with seven different syllable conditions and two noise levels. Since Wang and Bilger (1973) tested two syllable sets, one was to match phones tested in Miller and Nicely (1955) (CV1, VC1), and the other was to cover the missing phones from the first syllable set (CV2, VC2). Out of the seven syllable conditions, four were simply CV1, CV2, VC1 and VC2, and the other three were generated by merging these syllable sets, which are CV1 + VC1, CV2 + VC2 and CV1 + CV2 + VC1 + VC2. The reason for merging CV and VC syllables is to test whether the findings are stable across syllable types (CV, VC) and a context-free condition (CV + VC).

Finally, six matrices were extracted from Phatak and Allen (2007); they each have a different SNR level at -22, -20, -16, -10 and -2dB and a Quiet condition. Given that the unit of response in Phatak and Allen (2007) is at the level of the syllable, it is possible that some of the consonant confusions occurred with vowel confusions – that is both the consonant and the vowel were confused in a response. To better match the consonant confusions with the other three experimental corpora which allowed only consonant confusions, I extracted only the trials that contain correctly perceived vowels; that is, in trials that contain an error, only the consonant was misperceived.

For the noise level analyses of the vowels, nine matrices were extracted from

Cutler et al. (2004) with three syllable conditions (CV, VC, and CV + VC), and three SNR levels (0, +8, +16). The vowel confusions from Phatak and Allen (2007) were not included because they only tested four vowels.

For the bandwidth analyses, two sets of matrices were extracted from Miller and Nicely (1955). The first set of matrices has the SNR fixed at +12 and the lower frequency band fixed at 200 Hz, but with a different upper frequency band per matrix. The upper bands are 300, 400, 600, 1,200, 2,500, 5,000 and 6,500; therefore, six matrices were extracted. The second set of matrices has the SNR fixed at +12, and the upper frequency band fixed at 5,000 Hz, but with a different lower frequency band per matrix; the lower bands are 1,000, 2,000, 2,500, 3,000 and 4,500.

3.7.2.3 Comparative methods

The similarity between confusion matrices was analysed in terms of both global and structural similarities. The Mantel test (Section 3.3.2.1) was used to evaluate the global similarity for both consonant and vowel confusions. Hierarchical clustering techniques were used in Section 3.3.3.1 to evaluate the structural similarity for only consonant confusions.

With the hierarchical technique, we applied all three common linkages as before (Complete, Average and Single) with all three correlation measures (cophenetic correlation with Pearson coefficient, cophenetic correlation with Kendall Rank coefficient, and Baker's gamma index). Therefore, for the structural comparison, nine correlation values with p-values were obtained for each comparison. With the Mantel test, both the Pearson and Kendall Rank coefficients were tested; therefore, with the global comparison, two correlation values with p-values were obtained for each comparison.

In each comparison, one naturalistic confusion matrix (consonant or vowel) was compared with one experimental confusion matrix (consonant or vowel). Any phones

that do not exist in both matrices were excluded. The correlation values are only considered if the significance level is below a critical α-level, as they provide us with an indication of the validity of the correlation values. α was selected to be 0.1, which would allow us to include cases where the correlation values are near-significant. The significant correlation values would then be aggregated by different experimental conditions in each experimental dataset. Finally, the aggregated data were visualised with box-plots for evaluation.

In the following sections, I will first present the analyses of the Noise level condition for both consonants and vowels on the global level and for consonants on the structural level. I then will present the analyses of the bandwidth condition for the consonants on both global and structural levels.

3.7.3 Noise levels

This section attempts to identify whether specific SNR levels in experimental settings can generate consonant (and vowel) confusions that are most similar to those in naturalistic settings. We compare the six SNR levels in Miller and Nicely (1955), two noise conditions (with and without masking noise) in Wang and Bilger (1973), and six SNR levels (five with masking noise and one without any noise (i.e. in Quiet)) in Phatak and Allen (2007) for consonant confusions, as well as three SNR levels in Cutler et al. (2004) for both consonant and vowel confusions. We will analyse each experimental dataset, starting from Miller and Nicely (1955).

3.7.3.1 Miller and Nicely (1955)

To assess the global similarity, we perform the Mantel test. The significant correlation values between the consonant confusions in the naturalistic corpus and those in Miller and Nicely (1955) at different SNR levels are shown in Figure 3.22. The figure shows that at -18dB and -12dB the correlations are around 0.17 and the lowest correlation

is 0.1282 at -6dB. From -6dB onwards, there is a steady increase in correlation until +6dB and a sharp increase from +6dB to +12dB with a correlation of 0.2762. Overall, the figure shows +12dB correlates most strongly with the naturalistic corpus, and there is a steady increase in global correlation with an increase in SNR.

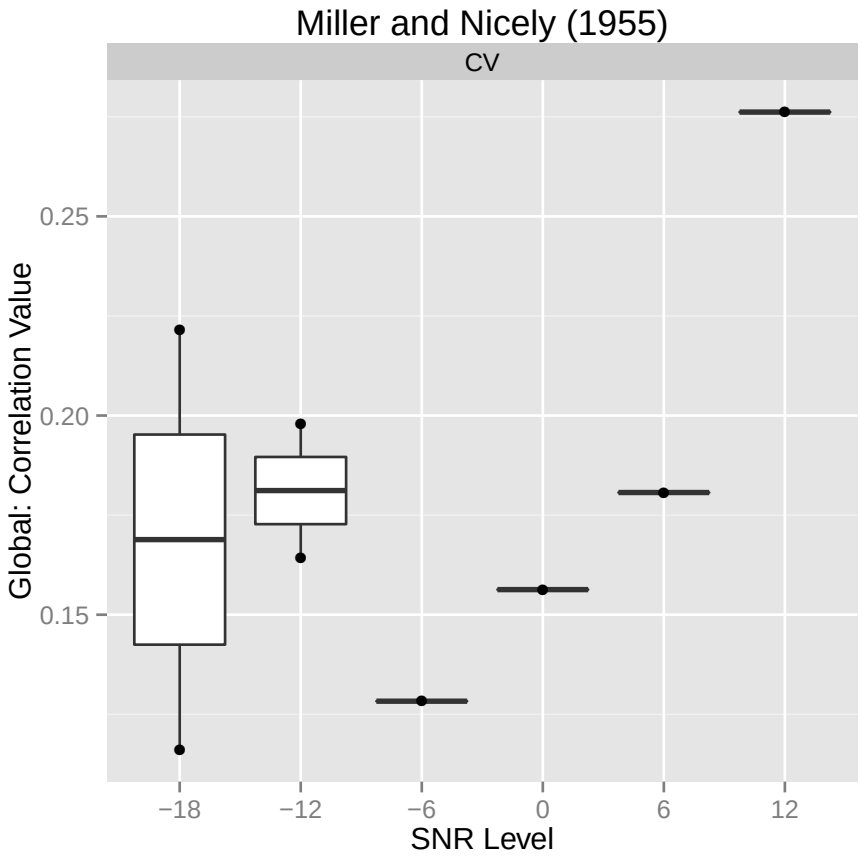


Figure 3.22: Global similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

Next, we assess the structural similarity with the same set of matrices. The significant correlation values between the hierarchical clusters of consonant confusions in the naturalistic corpus, and those in Miller and Nicely (1955) at different SNR levels are shown in Figure 3.23. Similar to the pattern on the global level (Figure 3.22), the figure shows a steady increase in structural correlation with an increase in SNR from -18dB to +12dB. The peak correlation is ≈ 0.45 at +12dB. The steady

increase has an obvious outlier at -6dB, which has a correlation as high as +6dB.

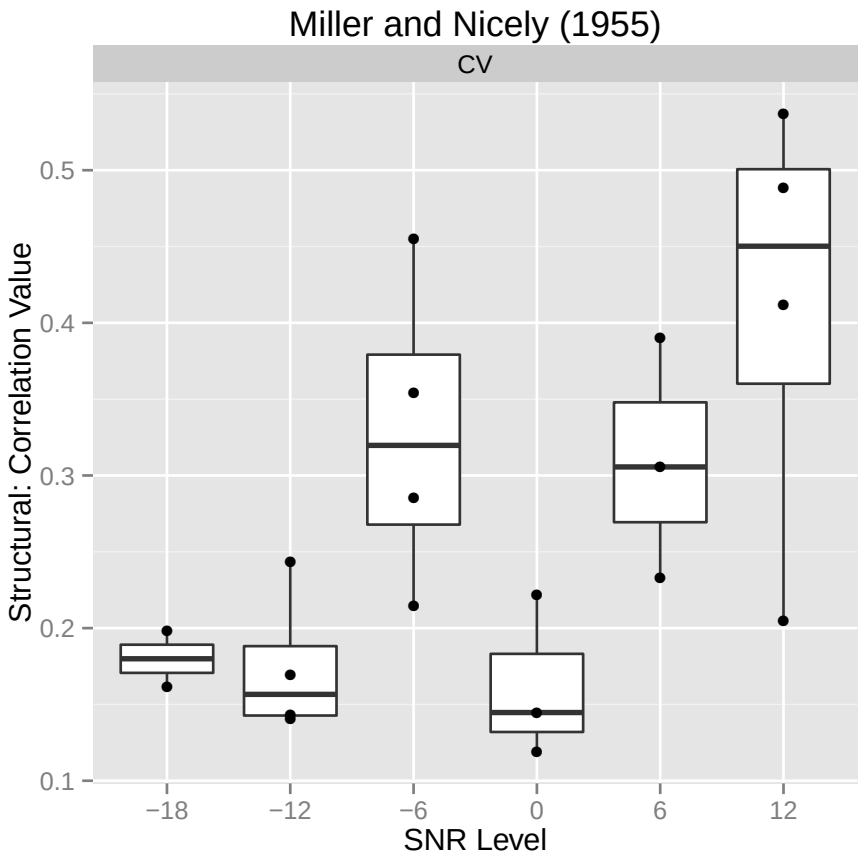


Figure 3.23: Structural similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

Having analysed the global and structural similarity between consonant confusions in the naturalistic corpus and those in Miller and Nicely (1955) at different SNR levels, we found that the best correlation is at +12dB and the pattern suggested that the low SNR levels are less ecologically valid – naturalistic confusions occur at relatively high SNR levels. To reinforce these findings, we will now analyse Wang and Bilger's (1973) consonant confusions.

3.7.3.2 Wang and Bilger (1973)

Unlike Miller and Nicely's (1955) study, Wang and Bilger (1973) did not provide the individual confusion matrices at each SNR level. However, there are two noise conditions that we could test. One of the noise conditions is the pooled data across all SNR levels, as well as the signal levels being manipulated from 50 to 95dB SPL in 15dB-steps. The other noise condition is that no masking noise was added to the signal and only the signal level was manipulated from 20 to 45dB SPL in 5-dB steps and from 55 to 115dB SPL in 10dB-steps. I labelled these two conditions as *Noise* and *Quiet* respectively.

Similar to the analyses presented in the previous section, we first analyse the global similarity between Wang and Bilger (1973) and the naturalistic corpus, and the correlation values of both noise conditions across seven syllable conditions are shown in Figure 3.24. Six out of seven of the syllable conditions suggest that the *Quiet* condition correlates better with the naturalistic corpus than the *Noise* condition, with the exception of the syllable condition *CV2*, which has *Noise* being more correlated. While *Quiet* is more correlated than *Noise* consistently, it is worth noting that the difference in correlation values is small (the largest difference is merely 0.03).

Let us move on to the analyses of structural similarity between Wang and Bilger (1973) and the naturalistic corpus. The results are summarised in Figure 3.25. Unlike the findings with the global similarity, the pattern is less robust. The results of the syllable conditions CV1, VC1, and CV1 + VC1 suggest that *Quiet* is better correlated than *Noise*. Of these three conditions, VC1 has the largest difference of almost 0.2 in correlation, and thus the large difference in CV1 + VC1, was driven by VC1. Focusing on the three conditions related to the second syllable set, CV2, VC2 and CV2 + VC2, there is no visible difference in the CV2 condition, and there is a slight difference in favour of *Noise* over *Quiet* in VC2; therefore, unsurprisingly, CV2

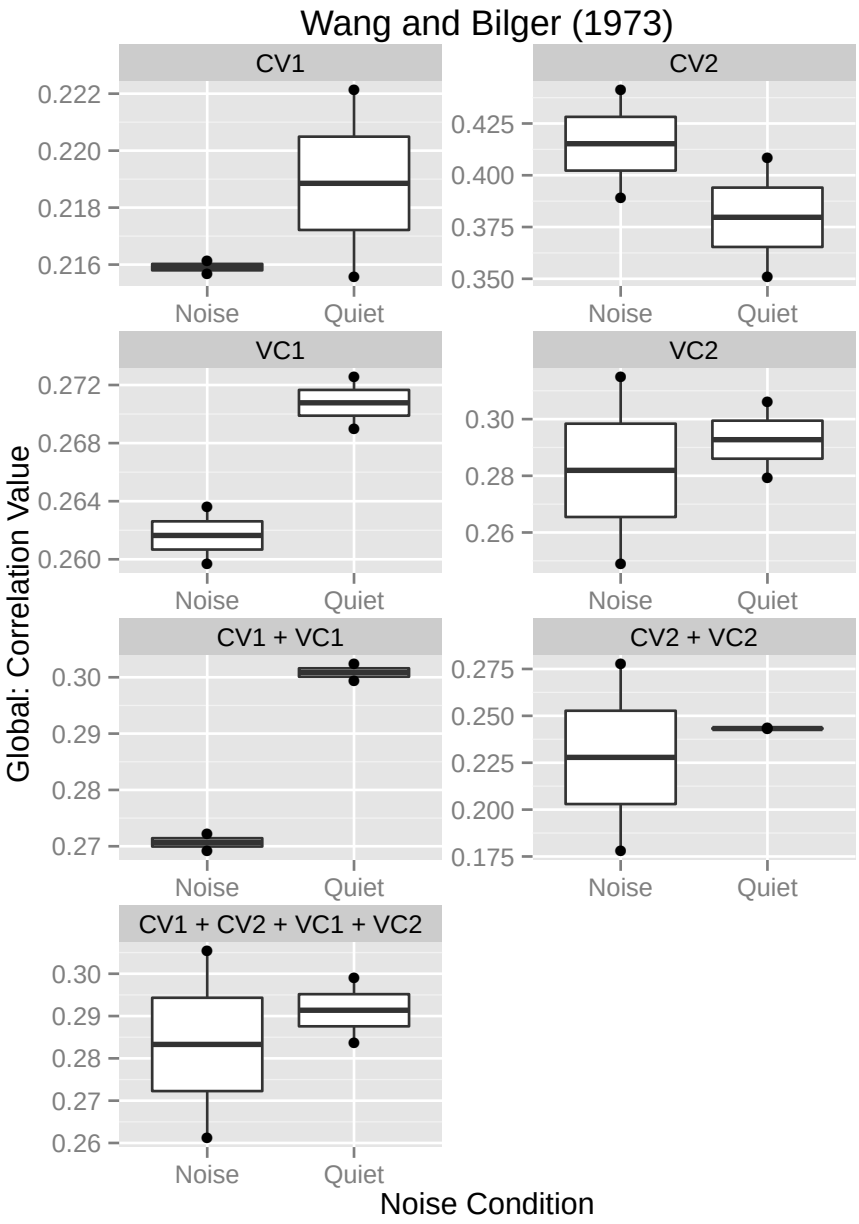


Figure 3.24: Global similarity of consonants between Wang and Bilger (1973) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

+ VC2 also has a slight difference in favour of *Noise*. Finally, the merged condition of all four matrices (CV1 + CV2 + VC1 + VC2) suggests that *Noise* is better correlated than *Quiet*, with a difference of 0.2. To sum up, there is a trend of *Quiet* being more correlated than *Noise* in the first syllable set CV1, VC1 and the CV1

+ VC1, but there is a reverse trend, although a weak one, with VC2. Furthermore, even though the difference is small with VC2, it was able to overcome the reverse pattern in CV1 and VC1, when all syllable sets (CV1 + CV2 + VC1 + VC2) were merged.

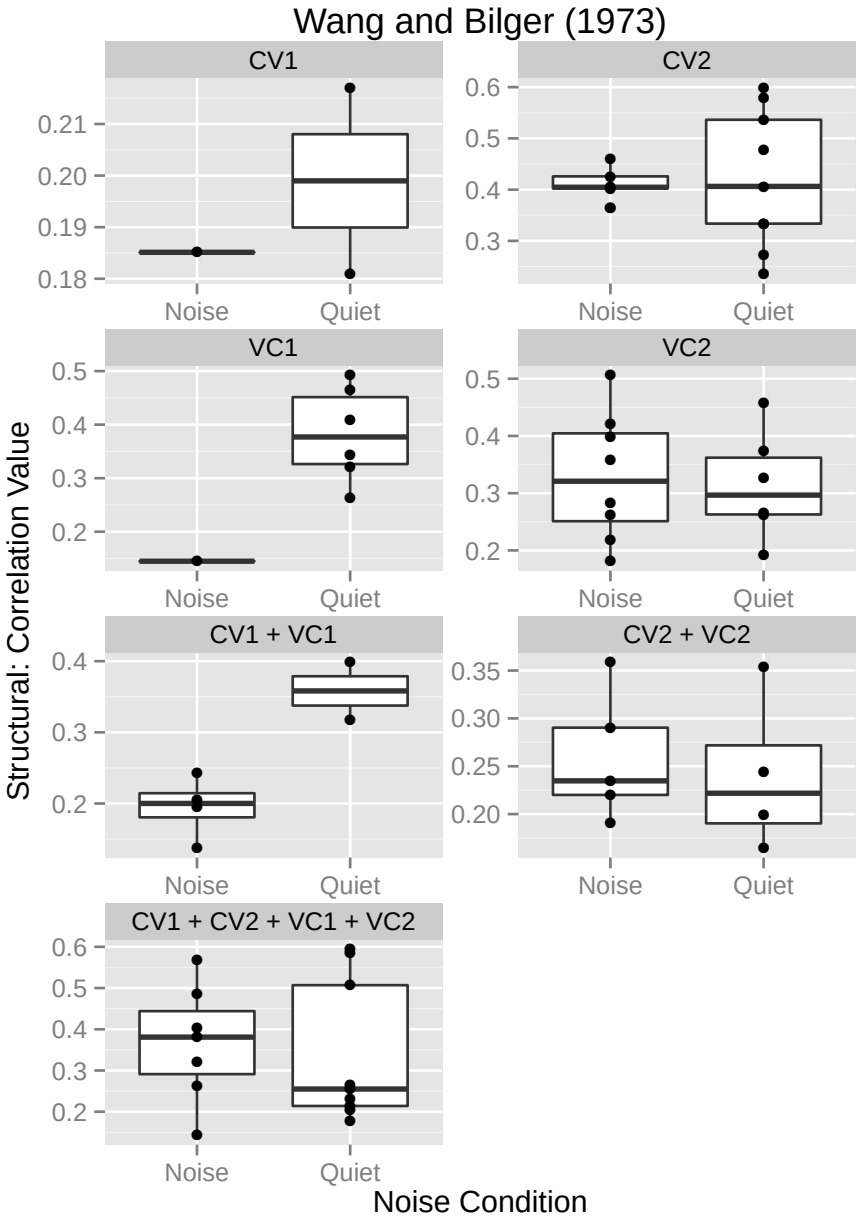


Figure 3.25: Structural similarity of consonants between Wang and Bilger (1973) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

The small difference between the *Quiet* and *Noise* in either direction needs to be explained. One possible factor is that in both conditions the signal levels were manipulated. In the *Noise* condition, the signals were less degraded as the signal levels were high (50 to 95dB SPL), while in the *Quiet* condition, the signals were more degraded (as low as 20dB SPL) and were manipulated with more levels (13 levels). Therefore, the two conditions were not only different in terms of the presence or absence of masking noise; thus, the differences in correlation were reduced. A second possible factor is that the *Quiet* condition might not be as natural as one would expect. Since in naturalistic settings masking noise definitely exists, the lack of any masking noise actually makes the condition *un*natural, which weakens the correlation.

Overall, the patterns identified in these analyses with Wang and Bilger (1973) are in line with the ones found with Miller and Nicely (1955): that is, naturalistic confusions occur at high SNR levels, when the signals are not severely degraded. Finally, we will perform similar analyses with Cutler et al.'s (2004) confusions.

3.7.3.3 Cutler et al. (2004)

Just as the analyses above, the global similarity between Cutler et al. (2004) and the naturalistic corpus was analysed. Figure 3.26 shows the significant correlation values using the Mantel test at three different SNR levels – 0, +8, and +16dB, and at three different syllable conditions – CV, VC and CV + VC. Across all three syllable conditions, the peak correlation is at +8dB which is in the middle of the SNR range, and the correlations at 0dB and +16dB are of a similar level.

Similarly, the structural similarity was analysed and summarised in Figure 3.27. Across all syllable conditions, 0dB has the highest correlation, +16dB came second, and finally +8dB has the lowest correlation. Interestingly, +8dB has the lowest correlation in terms of structural similarity, whereas it has the highest correlation in

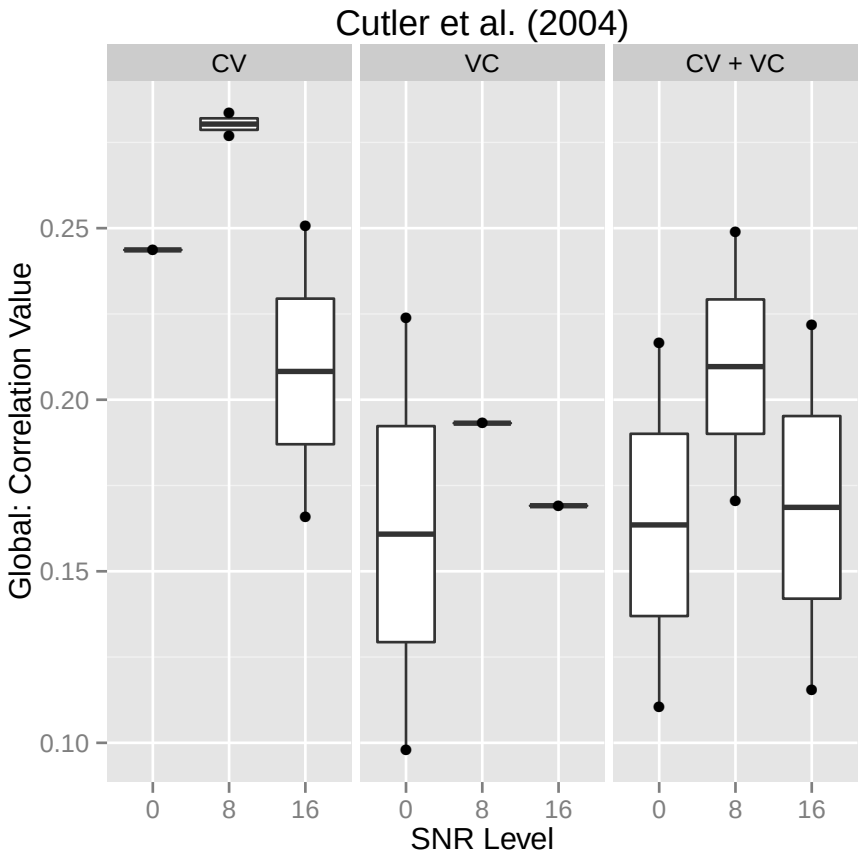


Figure 3.26: Global similarity of consonants between Cutler et al. (2004) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

terms of global similarity. I do not have an immediate explanation for this reversal in correlation between global and structural similarities.

Overall, the analyses of the global and structural similarity do not correspond with previous findings with Miller and Nicely (1955). The peak correlation was found at +8dB in terms of global similarity and 0dB in terms of structural similarity. Together, this suggests that within the positive range of SNR levels, the lower the SNR, the higher the correlation. With only three SNR levels, it is unclear whether this pattern is the result of a random peak or trough in correlation or not.

To clarify the pattern at hand, we perform the same analyses of global similarity on *vowel* confusions to see whether we would again find +8dB being the best condi-

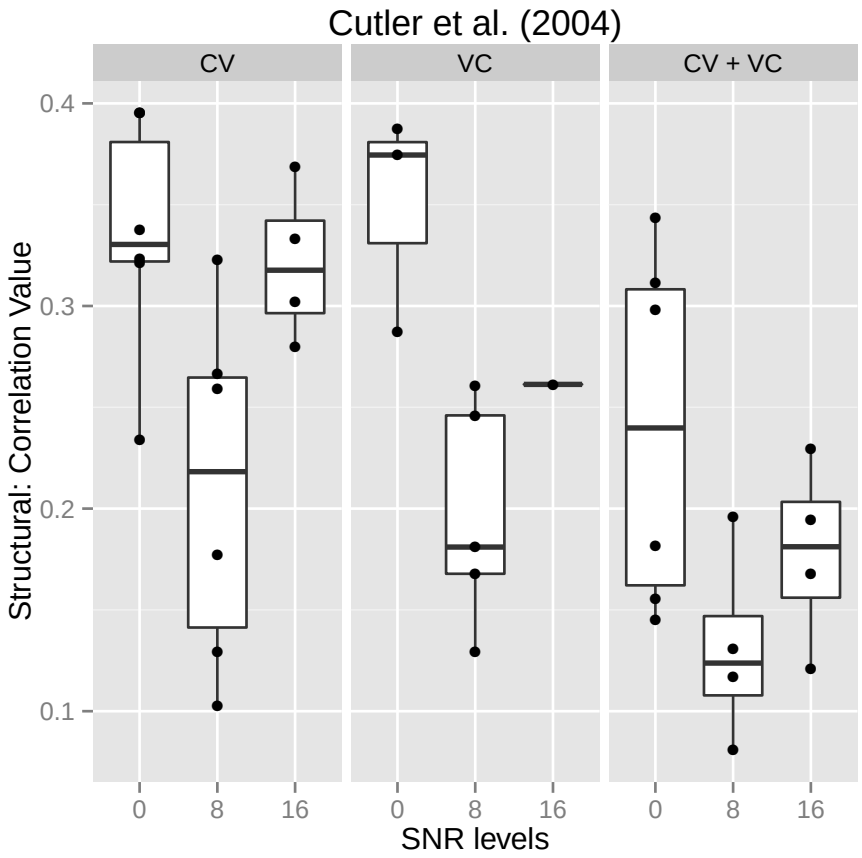


Figure 3.27: Structural similarity of consonants between Cutler et al. (2004) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

tion. The results are summarised in Figure 3.28. The first observation is that the correlation at +8dB again varies across syllable type. In CV, +8dB gives the highest correlation, while in VC it gives the lowest correlation. Together with the previous inexplicable pattern found with the +8dB condition, it is possible that this condition is affected by experimental artefacts. Overall, across all three syllable conditions, there is a trend of an increase in correlation with an increase in SNR, which is consistent with the patterns found with Miller and Nicely's (1955) consonant confusions.

Let us considering all the analyses above on Cutler et al. (2004). Consonant confusions appear to favour lower SNR, while vowel confusions appear to favour higher

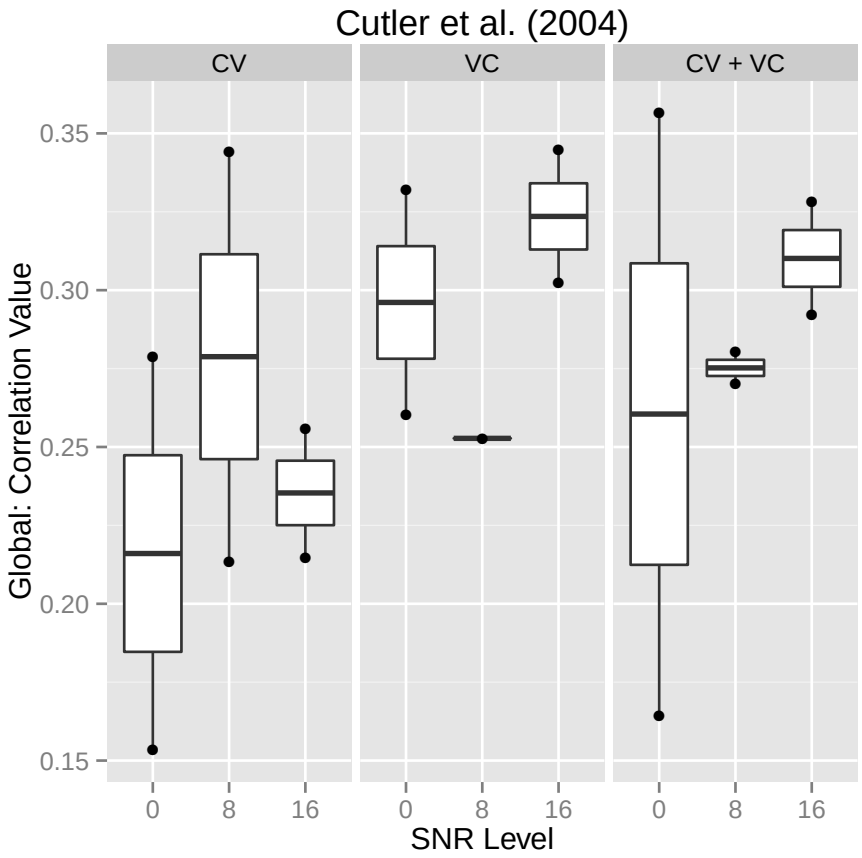


Figure 3.28: Global similarity of vowels between Cutler et al. (2004) and the naturalistic corpus at different SNR levels: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

SNR. The patterns are highly inconsistent, varying across consonants and vowels, and across global and structural similarities. This inconsistency could be due to the ceiling effect, because all the SNR levels tested were positive and the correlation values were simply fluctuating. Another possible explanation is that the masking noise was multi-talker babble noise which is considerably different from white noise, as used in Miller and Nicely (1955) and Wang and Bilger (1973), and that the two masking noise types are known to mask speech differently (as summarised in Section 3.7.1.1). Multi-talker babble noise masks speech-relevant frequency bands, while white noise masks high frequencies more than low frequencies in speech. Furthermore, multi-talker babble noise generated from a small number of talkers is a form

of informational and energetic masking, while white noise is purely energetic masking. Therefore, we might not expect to find similar correlations between Miller and Nicely (1955) and Cutler et al. (2004).

So far, our analyses of the data from Miller and Nicely (1955), Wang and Bilger (1973) and Cutler et al. (2004) have raised three further questions. In Wang and Bilger (1973), we found that *Quiet* (without masking noise) has a slight advantage over *Noise* (with masking noise); however, the two conditions are *not* perfectly matched in other aspects, e.g. the signal level manipulations were different. So it remains unclear as to whether conditions without masking noise really outperform conditions with masking noise. In Miller and Nicely (1955), we found that the higher the SNR, the better the correlation; however, the findings from Cutler et al. (2004) suggest that this is not the case with multi-talker babble noise. Together, they raised two further questions: a) whether the difference between Cutler et al. (2004) and Miller and Nicely (1955) is due to the use of different masking noise or not; and b) whether the fluctuating correlation across the three SNR levels in Cutler et al. (2004) is a result of ceiling effects or not. In fact, the next study, Phatak and Allen (2007), might be able to shed some light on these three questions. Firstly, it covers a wide range of SNR levels (five levels with masking noise), which could address the question regarding potential ceiling effects. Secondly, it uses speech-shaped noise, which is different from white noise and could address the question of whether the use of different masking noise has an effect on the correlation pattern. Finally, it also has a *Quiet* condition, which is matched in other aspects in the conditions that are masked with noise; this could therefore clarify the findings with Wang and Bilger (1973).

3.7.3.4 Phatak and Allen (2007)

As with the analyses above, the global similarity between Phatak and Allen (2007) and the naturalistic corpus is analysed. Figure 3.29 shows the significant correlation

values using the Mantel test at five different SNR levels: -22, -20, -16, -10 and -2dB as well as in *Quiet*. It is clear the *Quiet* condition has the lowest correlation of all ($r = 0.1356$). At the other extreme, -22dB shows similarly low correlation as *Quiet*. The correlation value increases as we moved away from the extreme SNR levels, with an upside-down U-shaped pattern. The correlation is highest ($r = 0.3199$) at -16dB, and followed by -10dB ($r \approx 0.29$). The highest and the lowest correlations have a sizeable difference of 0.2 in correlation.

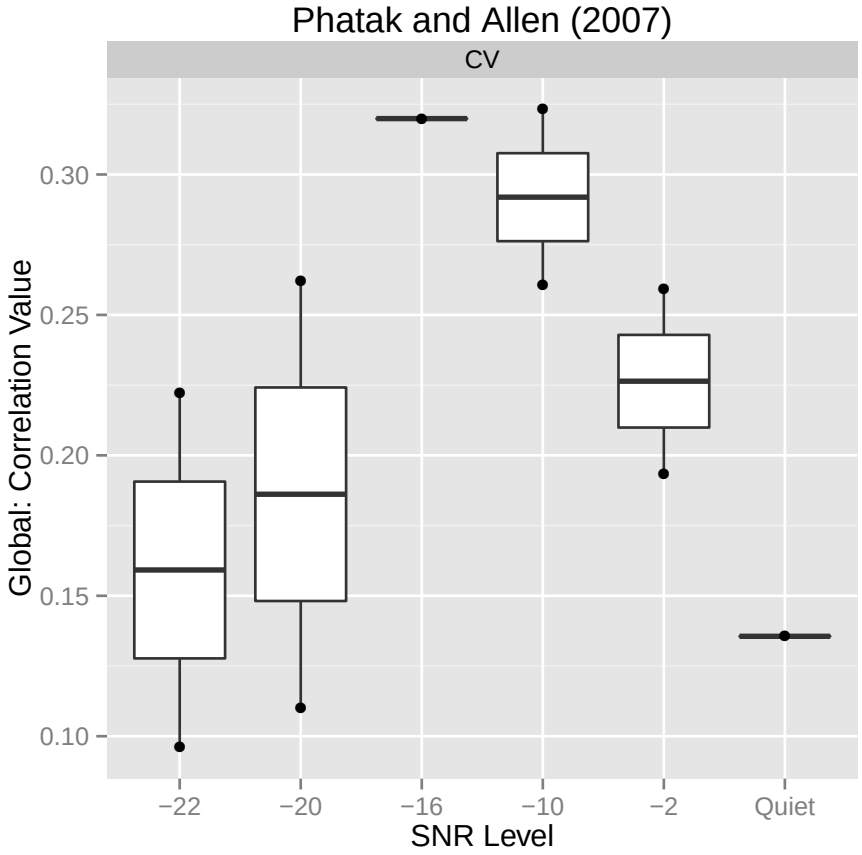


Figure 3.29: Global similarity of consonants between Phatak and Allen (2007) and the naturalistic corpus at different SNR levels and in Quiet: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

In fact, a similar pattern can be found in the analyses of structural similarity, as shown in Figure 3.30. Again, we see that the *Quiet* condition has the lowest correlation ($r \approx 0.15$), and -10dB has the highest correlation ($r \approx 0.41$), which was

the second highest in terms of global similarity. All the other SNR levels are of a similar correlation value at around 0.32.

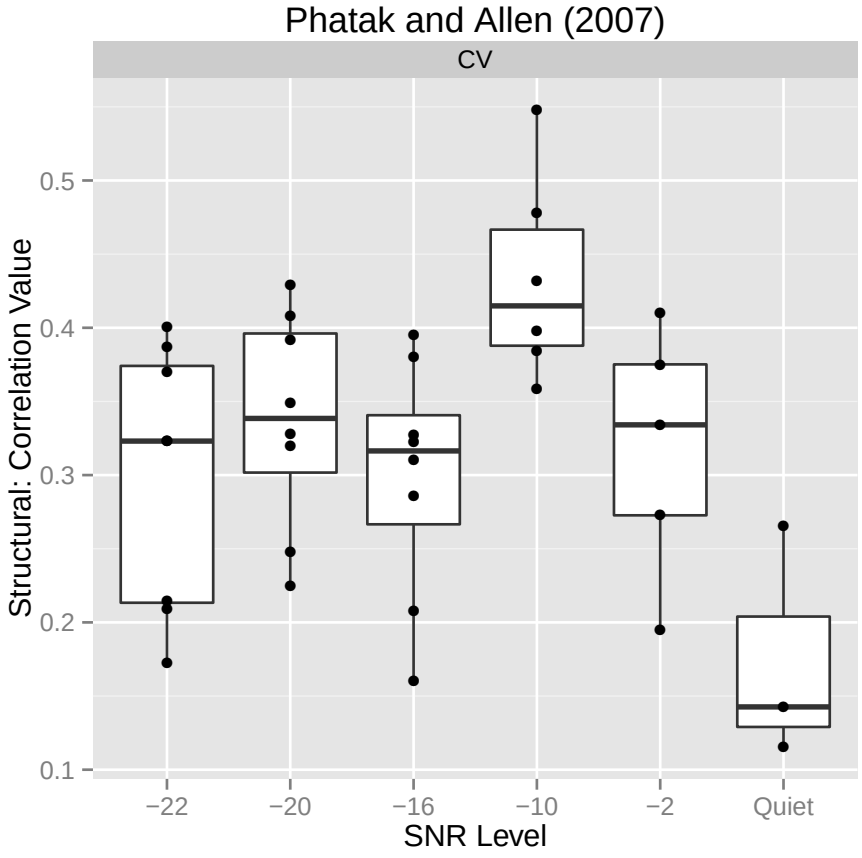


Figure 3.30: Structural similarity of consonants between Phatak and Allen (2007) and the naturalistic corpus at different SNR levels and in Quiet: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

The findings derived from the data reported by Phatak and Allen (2007) have indeed clarified a number of questions we raised earlier. Firstly, a *pure Quiet* condition without any experimental manipulation of adding masking noise and of changing the signal level yielded the worst correlation in both global and structural similarities. This is likely due to the lack of confusions. Compared to the findings with Wang and Bilger (1973), the consistent but small advantage of *Quiet* over *Noise* is likely due to the signal level manipulation (which generated confusions), and not the lack

of masking noise.

Secondly, recall that with consonant confusions in Cutler et al. (2004) the lowest SNR 0dB has higher correlation than the highest SNR +16dB with the CV syllable condition (Figure 3.26 and Figure 3.27). We speculated that this is simply the result of ceiling effects and the difference is due to random fluctuations. This pattern can, in fact, be explained with what we found with the data from Phatak and Allen (2007), such that an intermediate SNR level is more correlated than the extreme SNR levels, and that the high correlation at 0dB and low correlation at +16dB lie on the right of the upside-down U-shaped pattern. We therefore rejected the speculation of the ceiling effects.

Finally, recall that with Miller and Nicely (1955), we found that the higher the SNR, the higher the correlation. However, in the analyses of both Cutler et al. (2004) and Phatak and Allen (2007), we found that the best correlation is not necessarily achieved at high SNR. Taking into consideration the low correlation with the Quiet condition, it is clear that "the higher the SNR, the higher the correlation" is untrue, because the Quiet has extremely high SNR, and if we were to increase the SNR beyond +12dB (the highest level in the experiment) in Miller and Nicely (1955), we should expect to see an increase and a definite decrease when the SNR is too high to generate a sufficient amount of confusions, therefore revealing the upside-down U-shaped pattern. Together, this suggests that the use of different masking noise types has an effect on the resultant correlation pattern, such that the peak correlation lies at different SNR levels. We found white noise has a peak correlation at higher SNR (in the positive range) than speech-related noise (speech-shaped noise and multi-talker babble noise), which have a peak correlation at a lower SNR (in the negative range).

3.7.4 Frequency bandwidth

This section attempts to identify whether specific bandwidth manipulation in experimental settings can generate consonant confusions that are most similar to those in naturalistic settings. Two bandwidth manipulations were examined in Miller and Nicely (1955). The first manipulation is the adjustment of the upper frequency band: that is, to apply different low-pass filters. The second manipulation is the adjustment of the lower frequency band: that is, to apply different high-pass filters. First, we will analyse the low-pass filter condition.

3.7.4.1 Low-pass filter

To assess the global similarity, we perform the Mantel test between the consonant confusions in the naturalistic corpus and those in Miller and Nicely (1955) at different SNR levels with six different low-pass filters: 300, 400, 600, 1,200, 2,500, 5,000 and 6,500 Hz with a fixed high-pass filter at 200 Hz and +12dB SNR. The significant correlation values are summarised in Figure 3.31; 200–400 Hz and 200–6,500 Hz correlate equally well at around 0.275. The other high-pass filters have low correlation values that fluctuate from 0.15 to 0.20. The high correlation at 200–400 Hz is surprising, considering that it is extremely narrow, and therefore should not reflect naturalistic settings. For this reason, it is possible that 200–400 Hz is an outlier, and that there is a trend of an increase in correlation with an increase in bandwidth. Analyses of the structural similarity could clarify this speculation, and the results are summarised in Figure 3.32.

The pattern became clearer in terms of structural similarity. Overall, there is a positive trend as speculated in the global similarity analyses, with the highest correlation ($r \approx 0.45$) at 200–6,500 Hz, and the second highest correlation ($r \approx 0.43$) at 200–5,000 Hz, and the lowest correlation ($r \approx 0.19$)at 200–300 Hz. However, the trend is not a steady one. Among the high-pass filters in between, those at 400, 600

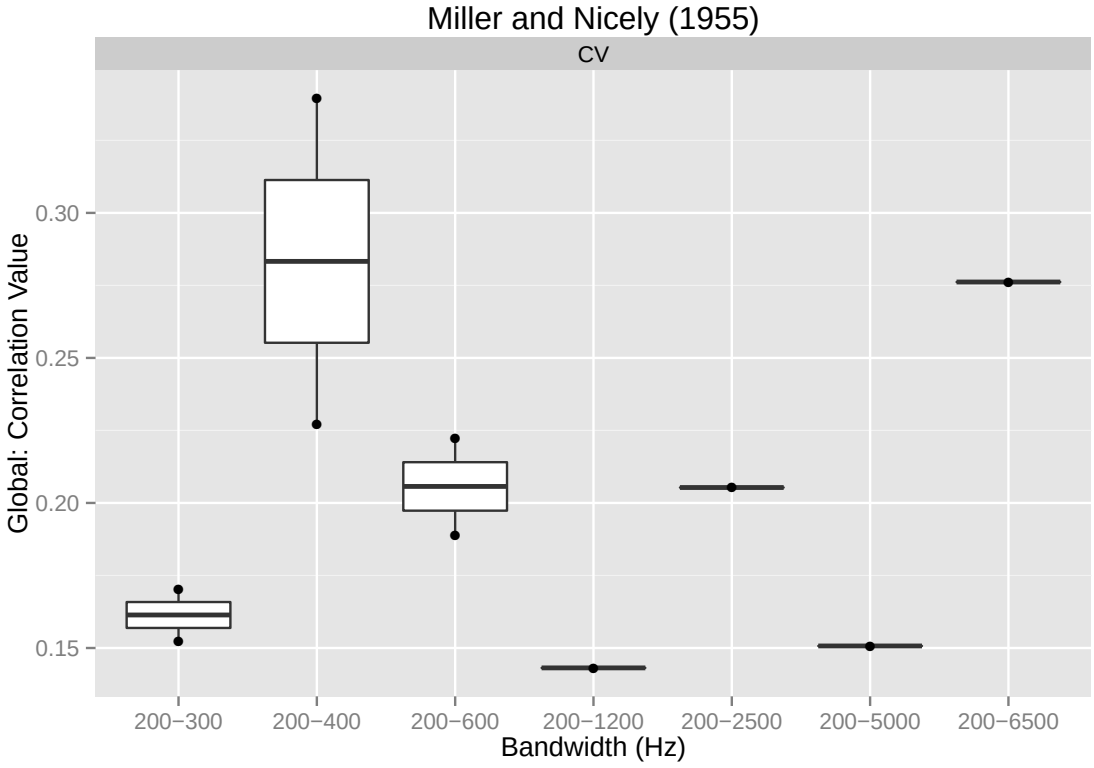


Figure 3.31: Global similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus with different low-pass filters: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

and 1,200 Hz have higher correlation than 300 Hz and 2,500 Hz. It is possible that this "peak" in correlation is merely noise in the data, given that 200–400 Hz was also a potential outlier in the global similarity analyses.

As discussed in Section 3.7.1.1, Miller and Nicely (1955) observed low-pass filtering has a similar effect on misperception as manipulating the SNR levels using white noise. Therefore, the global and structural correlations between the naturalistic data and experimental data could be similar between the experimental data manipulated with low-pass filtering and those manipulated with noise levels. This speculation is, in fact, confirmed by comparing Figure 3.31 (the global correlations with different low-pass filtering) with Figure 3.22 (the global correlations with different noise levels); and similarly Figure 3.32 (the structural correlations with different low-pass

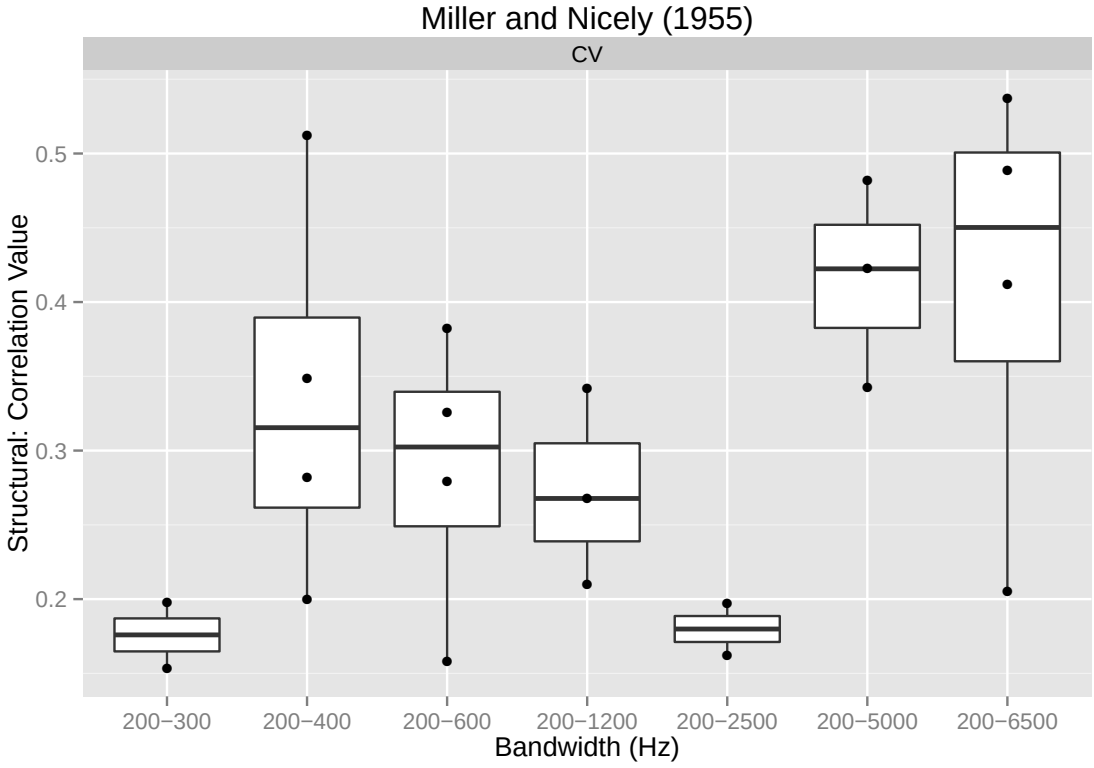


Figure 3.32: Structural similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus with different low-pass filters: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

filtering) with Figure 3.23 (the structural correlations with different noise levels). These comparisons show that the correlation patterns are similar between the manipulation of low-pass filtering and that of noise levels, with the highest correlation at the highest SNR level and the widest bandwidth.

3.7.4.2 High-pass filter

Let us move on to the analyses of the high-pass filter condition. The global similarity analyses are summarised in Figure 3.33 and the structural similarity analyses are summarised in Figure 3.34.

In terms of global similarity, the two widest bandwidth conditions (1,000–5,000 Hz and 200–5,000 Hz) ($r \approx 0.16$) have higher correlation values than the two narrowest

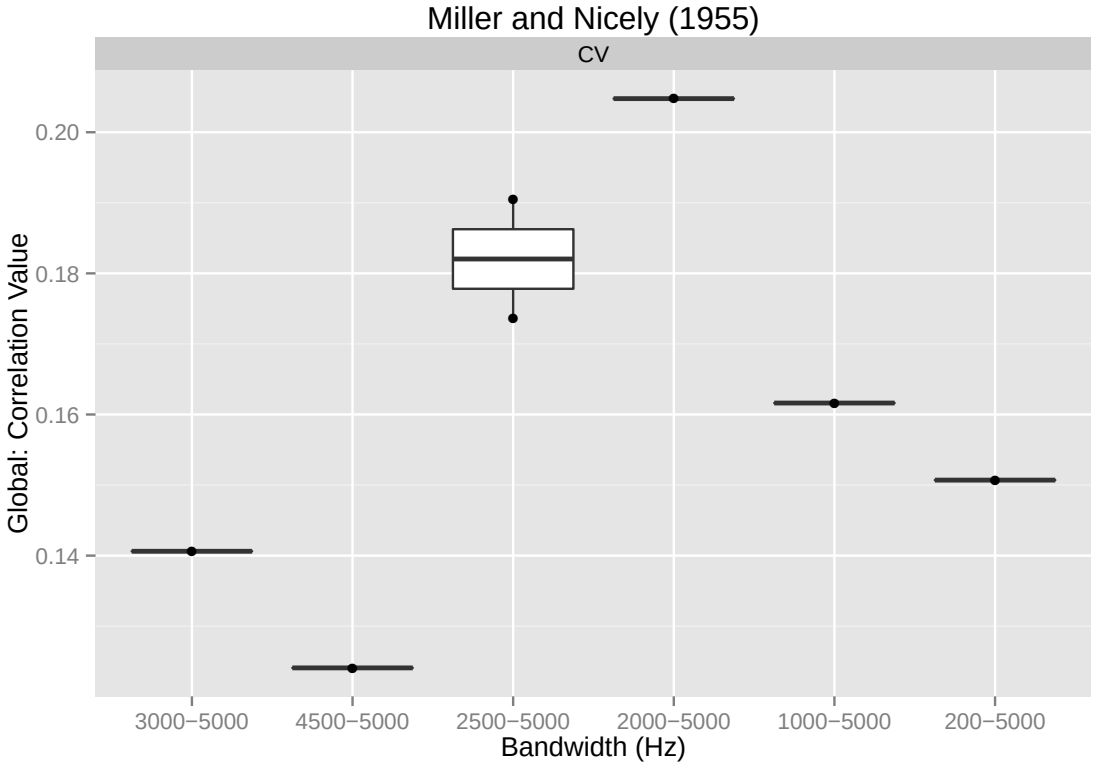


Figure 3.33: Global similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus with different high-pass filters: the points represent the significant correlation values (Mantel's correlation), aggregated with boxplots.

bandwidth conditions (3,000–5,000 Hz and 4,500–5,000 Hz) ($r \approx 0.13$), which is expected from the low-pass filter analyses. However, the two bandwidth conditions in between (2,500–5,000 Hz and 2,000–5,000 Hz) have the two highest correlation values ($r \approx 0.18$ and 0.20), which is unexpected. In terms of structural similarity, there is a solid trend showing consistently low correlation values (r from 0.25 to 0.3) at 3,000–5,000 Hz, 4,500–5,000 Hz, 2,500–5,000 Hz and 2,000–5,000 Hz, and a sharp increase in correlation at 200–5,000 Hz, which is the widest bandwidth. (N.B. The 1,000–5,000 Hz condition has no significant correlation values to report.)

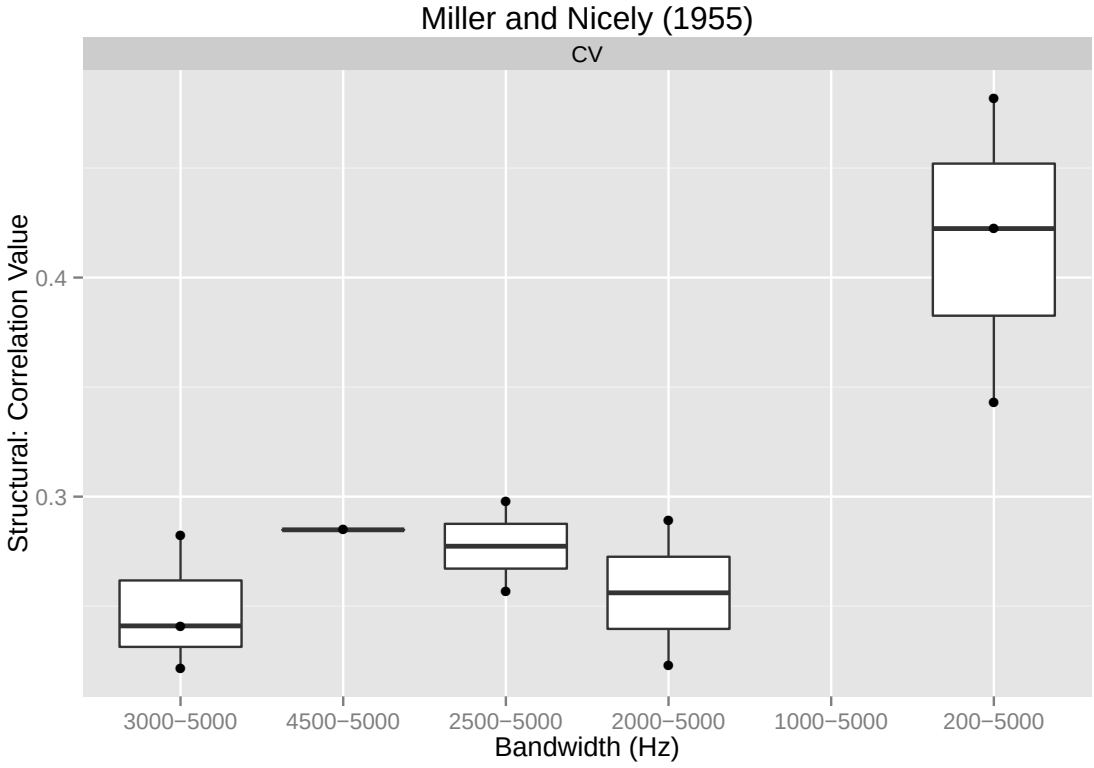


Figure 3.34: Structural similarity of consonants between Miller and Nicely (1955) and the naturalistic corpus with different high-pass filters: the points represent the significant correlation values (cophenetic correlation and Baker's gamma index), aggregated with boxplots.

3.7.5 Conclusion

Section 3.7.3 attempted to address whether specific SNR levels can generate consonant confusions that are most similar to those in the naturalistic corpus. By examining four different experimental corpora, we discovered a number of findings.

Firstly, extreme SNR levels (too much noise to no noise (i.e. Quiet)) are least similar to naturalistic conditions, because at extremely low SNR levels the signals are too degraded, and the resultant confusions are randomly generated (i.e. by chance), and at extremely high SNR levels (i.e. Quiet) the signals are not degraded enough to generate sufficient amount of confusions for any patterns to emerge.

Secondly, three noise-types were used across the experimental corpora: white noise, multi-talker babble noise and speech-shaped noise. Our finding suggests that

the SNR level that is most similar to the naturalistic corpus lies in a different range for different masking noise. With white noise, the "optimal" level ("optimal" is defined as the level with the best correlation with the naturalistic corpus) lies in the positive SNR range; and with speech-shaped noise it lies in the negative SNR range. Furthermore, the vowel analyses using the data from Cutler et al. (2004) suggest that the optimal SNR level lies in a different range between consonant confusions and vowel confusions. In terms of vowel confusions with multi-speaker babble noise, the optimal level is likely to be in the positive range. The optimal level is further dependent on other factors such as syllable type, and perhaps other factors that were not examined, such as knowledge of the listeners (native versus. non-native), speech rate, lexicality (some CV and VC syllables could be a word) and many others.

Thirdly, the findings using the data from Wang and Bilger (1973) suggest that confusions generated by manipulating the signal level without any masking noise have the potential to generate confusions that are similar to the naturalistic confusions. In fact, manipulation of the signal level has been used to generate confusions that are also found in naturalistic corpora. For instance, Cutler and Butterfield (1992) generated mis-segmentation of word boundaries using faintly heard speech. That is, speech presented at a low signal level with an error rate of 50% and the resultant mis-segmentation pattern matched that found in a naturalistic corpus (Bond's corpus).

Section 3.7.4 examined another experimental condition: bandwidth filtering. Using Miller and Nicely's (1955) data, it was found that the wide/full bandwidth produces confusions that are most similar with naturalistic confusions. Even though narrowing the bandwidth can generate more confusions, those confusions are not realistic, and therefore bandwidth manipulation as an experimental condition is not ecologically valid.

In sum, while it is possible to induce misperception using a number of experimental manipulations, such as masking the signal with noise of different types at

different SNR levels, bandpass filtering, adjusting signal levels and many others, some conditions are better correlated with naturalistic misperception. When designing a perception experiment, researchers have to decide what manipulations to apply. However, in most cases, such decisions are simply arbitrary, picking manipulations that are not too extreme, or holistic – trying to cover a wide range of conditions. This is because there is simply no benchmark for whether a given manipulation is valid or not. If a given hypothesis is not consistently rejected across the manipulations tested, then the researchers will have to somehow justify why this is the case. Given these problems, I argue that our naturalistic corpus can serve as a benchmark corpus. Future researchers could compare their confusion matrices generated with different experimental conditions to the benchmark matrix to determine whether certain conditions are more ecologically valid than others.

3.8 Analysis of asymmetrical patterns

This section will examine the asymmetrical patterns in consonant and vowel confusions. It is well-known that both historical sound change and perceptual confusions show asymmetries (Ohala, 1989). Under Ohala's framework, the listener is a source of sound change (Ohala, 1981). Perceptual experiments controlled specifically to test this hypothesis (Plauché, Delogu, and Ohala, 1997; Chang, Plauché, and Ohala, 2001) complement classic perceptual experiments such as Miller and Nicely (1955) and those mentioned in Section 3.7.1. However, it remains an open question as to whether asymmetrical patterns found in perceptual experiments can also be found in naturalistic settings. If sound change is indeed motivated by perceptual confusions, then it is crucial that these confusions occur beyond laboratory settings, and can be found in the naturalistic corpus.

Three asymmetrical patterns are selected for analyses: 1) TH-fronting, 2) velar

nasal fronting, and 3) back vowel fronting. These patterns are selected because previous work on sound change has considered them in terms of a perceptual-based account as well as their high confusion rates. In evolutionary phonology, Blevins (2004, pp. 134–135) considers TH-fronting as a context-free sound change that is motivated by perceptual asymmetries and the confusion rate between [θ] and [f] is one of the highest consonant pairs in classical confusion studies like Miller and Nicely (1955). While most accounts of velar nasal fronting has been either articulatory or historical (Houston, 1985), the confusion rate between [ŋ] and [n] is also relatively high, being the 6th most confusable pair in the naturalistic corpus (see Table 3.12 for the top ten most confusable pairs). Given the high confusion rate and the fact that it is a well-known asymmetrical pattern, it is worth examining if a perceptual-based account can also be made, as in the case of TH-fronting. Finally, back vowel fronting is a well-established sound change pattern for chain shifts (Labov, 1994a), and vowel confusions in experiments have been used to motivate this specific sound change (Benkí, 2003). It is therefore worth re-examining this specific pattern to see if the asymmetries can also be found in the naturalistic corpus.

Segment Pairs	Rank
[ʃ, ʒ]	1st
[dʒ, ʒ]	2nd
[z, s]	3rd
[n, m]	4th
[dʒ, tʃ]	5th
[n, ŋ]	6th
[tʃ, ʃ]	7th
[k^h, p^h]	8th
[t, p]	9th
[p, k]	10th

Table 3.12: The top ten most confusable consonant pairs in naturalistic misperception: the "Rank" column indicates the rank for the top ten consonant pairs with 1st being the most confusable.

For each pattern, I will examine the strength and direction of the asymmetries

in both the naturalistic corpus and the experimental corpora (Section 3.7.1). This indirectly reinforces the ecological validity of the experimental studies, and reveals whether any specific experimental conditions would fail to generate the asymmetrical patterns.

In the following sections, I will first describe the method for quantifying asymmetries. After this, I will focus on each of the three asymmetrical patterns. Finally, I will summarise the findings in the conclusion section.

3.8.1 Method

To quantify the asymmetries from confusion matrices, we employed a bias measure from signal detection theory, called c (short for *criterion*) (Macmillan and Creelman, 2004, pp. 27–31).

To calculate c, we first need to extract a 2 by 2 matrix (containing the correct and incorrect responses for a given pair of phones) from a full matrix. This procedure relies on the *constant ratio rule*. According to the constant ratio rule (Clarke, 1957), response bias is presumed to be independent of the number of stimuli; that is, the frequency ratios of the responses should be approximately constant in both a full matrix and a sub-matrix. This rule performs better on multidimensional stimuli than unidimensional ones (Hodge and Pollack, 1962; Hodge, 1967). Given that phones are multidimensional, this rule is especially appropriate for our confusion matrices. However, it has been suggested by Luce that this rule is a strong and uncongenial assumption (Macmillan and Creelman, 2004, p. 249); therefore, the findings in this section should be subjected to alternative methods in the future. In any case, we will accept the constant ratio rule in the present study. Following this rule, we can therefore extract a subset 2 by 2 matrix from the full matrix in counts (i.e. not converted into proportions) for a given pair of phones.

The resultant 2 by 2 matrix is illustrated in Figure 3.13. The confusion matrix

Stimulus \Response	x	y
x	HIT	MISS
y	FALSE ALARM	CORRECT REJECTION

Table 3.13: An illustration of Hit, Miss, False Alarm and Correction Rejection in a 2 by 2 confusion matrix

contains two dummy segments: x and y. In signal detection theory, we treat the confusion matrix as a Yes-and-No task: where x is the target stimulus, the response can either be Yes or No. A Yes response is a successful identification of x, and this is called a HIT. A No response is an incorrect identification of x as y, and this is called a MISS. In the second row of the matrix, when the stimulus y is perceived as x, it is called a FALSE ALARM, which is a false alarm of selecting x, and when the stimulus y is perceived as y, it is called a CORRECT REJECTION, which is a correct rejection of x. The HIT and FALSE ALARM counts can then be converted into proportions as HIT rate and FALSE ALARM rate. HIT rate is HITS divided by the sum of HITS and MISSES. FALSE ALARM rate is FALSE ALARMS divided by the sum of FALSE ALARMS and CORRECT REJECTIONS. Finally, the HIT rate and the FALSE ALARM rate are converted into z scores. The z scores of the two rates are then summed and multiplied by -0.5 to give the bias measure c. In sum, the bias measure c is defined as:

$$c = -0.5 \times (z(HIT\ rate) + z(FALSE\ ALARM\ rate))$$

A negative c indicates a bias in favour of x, in our dummy matrix (Figure 3.13), i.e. y is perceived as x more often than x as y. A positive c indicates a bias in favour of y over x, i.e. x is perceived as y more often than y as x. A zero c indicates there is no bias in either direction.

The sparse matrix issue (as discussed in Section 3.3.1.5) is also a problem for methods in signal detection theory. When computing the bias measure c, if either the FALSE ALARM rate or the HIT rate were 1 or 0, then we would get infinity during

the conversion from proportion to z scores. In most cases, the issue is either the FALSE ALARM rate was 0 or the HIT rate was 1 (or both). For instance, in our dummy matrix, if there were no x being perceived as y, then the HIT rate would be 1; if there were no y being perceived as x, then the FALSE ALARM rate would be 0.

A common correction technique in signal detection theory is to convert any rate with the value of 0 to $0.5/N$, where N is the count sum of the row on which the zero rate lies. In other words, this is saying we have half a count. Based on this concept, any rate with a value of 1 is also converted to $1 - 0.5/N$. In other words, it is the count sum minus half a count (Macmillan and Kaplan, 1985). This is similar to the additive smoothing technique (as discussed in Section 3.3.1.5), but without correcting the count sum from N to $N + 0.5$.

However, if both the MISS rate (1 - HIT rate) and the FALSE ALARM rate are both 0, then there is simply no confusion in either direction, and the resultant c bias value is dependent purely on the count sum of each row in the matrix due to smoothing. I regard these values to be dubious; therefore, if both the MISS rate and the FALSE ALARM rate were 0, then I would set the c bias value to 0, assuming no bias in either direction. A final remark is that c bias values are unit-less like z scores.

3.8.2 TH-fronting

TH-fronting is a well-known sound change in progress in some varieties of English. The dental fricatives [θ, ð] are relatively unnatural phones and rare across languages. Indeed, children when acquiring these sounds would tend to substitute them with the labial fricatives [f, v] respectively. In some accents, such as New York, the dental fricatives are replaced by alveolar plosives [t, d] instead (Wells, 1982b, pp. 96–97).

This phenomenon is widespread among a variety of English accents, e.g. New Zealand English (Wood, 2003), London and Edinburgh English (Schleef and Ramsammy, 2013), to name a few. It is sensitive to lexical frequency, phonotactics and

morphological complexity (Stuart-Smith and Timmins, 2006; Clark and Trousdale, 2009), and it has been suggested that its prevalence is partly due to media exposure (Stuart-Smith, 2005; Stuart-Smith, 2007).

A perceptual account of the sound change [θ] > [f] is more natural than an articulatory account (Blevins, 2004, pp. 134–135). In terms of the acoustic cues, one difference between [θ] and [f] is that the intensity range is lower in [f] (3,000–4,000 Hz) than that in [θ] (7,000–8,000 Hz) (Yavaş, 2011, p. 111). However, this frequency cue is, in fact, not the main perceptual cue (Levitt et al., 1988). The crucial cue is the difference in the formant transition preceding or following the two phones; however, this difference is small and therefore not robust (Blevins, 2004, pp. 134–135). Furthermore, perception studies with infants showed that while they can distinguish between all other segmental contrasts, they have difficulties distinguishing [θ] and [f] (Vihman, 1996, p. 60). Indeed, in experimental studies of perception, such as Miller and Nicely (1955), [θ] and [f] has one of the highest confusion rates because they are perceptually similar. More specifically, in studies such as Miller and Nicely (1955), [θ] is perceived as [f] more often than [f] as [θ] (Johnson, 2012, Ch. 5); this suggests that while the two phones are perceptually similar, there is a asymmetry. However, it is yet to be confirmed whether this pattern holds water in naturalistic settings and if this pattern is robust across a wide range of experimental manipulations of SNR levels, bandpass filters, response types, and syllable types (CV vs. VC).

To address these questions, in the following sections, I will first conduct a general analysis, comparing the confusion pattern in the naturalistic corpus with those in experimental matrices. I then will focus on the experimental matrices, examining the effects of two experimental conditions – noise levels and frequency bandwidth (and the syllable types whenever possible) – on the asymmetrical pattern.

3.8.2.1 Overview

The naturalistic data being examined is the same context-free consonant confusion matrix. The experimental data being examined are those mentioned in Section 3.7.1. The matrices extracted are similar to those extracted for the ecological analyses in Section 3.7.2.2.2, but without all the combined matrices by syllable types. The experimental matrices are:

- 17 matrices from Miller and Nicely (1955) with different SNR levels and band-pass filters (all CV).
- Six matrices from Wang and Bilger (1973) with one CV set, two VC sets, and two noise conditions (Noise and Quiet).
- Six matrices from Cutler et al. (2004) with CV and VC syllable types across three SNR levels.
- Six matrices from Phatak and Allen (2007) at six SNR levels, but strictly consonant confusions (the vowels in the CV syllables are correctly perceived).

For the analyses below, I will only address the voiceless pair [θ] and [f]. This is because the sound change with the voiced pair is dependent on its phonological environment, being predominantly in medial and final positions (Kerswill, 2003). Since our experimental studies do not cover any medial consonants, detailed comparisons cannot be made.

Before looking at any asymmetrical patterns, we could look at the perceptual similarity of the two phones in question. In the hierarchical clustering analyses in Section 3.6, we projected the consonant confusion into three clustering trees, each with a different linkage: complete, average and single. Interestingly, the complete and average trees (Figure 3.19 and Figure 3.20) revealed that at the finest clustering levels, [θ] and [f] are of the same hierarchical cluster in terms of perceptual similarity. At this

point, it should be recalled that hierarchical clustering is a structural representation of the confusion patterns, which often recapitulates natural classes. The consistent clustering of [θ] and [f] suggests that the two phones are perceptually most similar to each other.

To examine the asymmetries, we calculate the c bias measure from the naturalistic matrix, and each of the experimental matrices. The corresponding c bias values are plotted in Figure 3.35 as boxplots, separated by studies in the x-axis, with c bias values plotted on the y-axis as both individual points and box plots. The first observation is that both naturalistic and experiment studies have their median c bias values above zero (as indicated by the black line in the boxplots being above the dotted line). This shows that there is a robust perceptual bias for [θ] being perceived as [f] more often than the reverse. Its robustness is indicated by how this bias is consistent across four experimental studies, each of which has very different experimental conditions (see Table 3.8 for a summary). Furthermore, we analyse all the c bias values from the experimental studies ($N = 35$) by applying a one-sample Wilcoxon signed rank test, testing the hypothesis that the true value is zero. We found that the c bias values of the experimental studies (the mean value: 0.278) are significantly different from zero, with $p = 7.74 \times 10^{-5}$ (two-tailed), indicating that it is a significant positive c bias. Most importantly, this positive c bias also exists in the naturalistic corpus, therefore indicating the bias (and indeed TH-fronting) cannot only be found in experimental settings, but also in the "wild".

The second observation is regarding the variability of the c bias within each of the experimental studies, as indicated by the size of each box plot. Some studies have higher variability than others. The study with the highest variability is Miller and Nicely (1955), followed by Wang and Bilger (1973), then Cutler et al. (2004) and Phatak and Allen (2007). The variability can possibly be explained by the number of experimental conditions tested. For instance, Miller and Nicely (1955) tested both

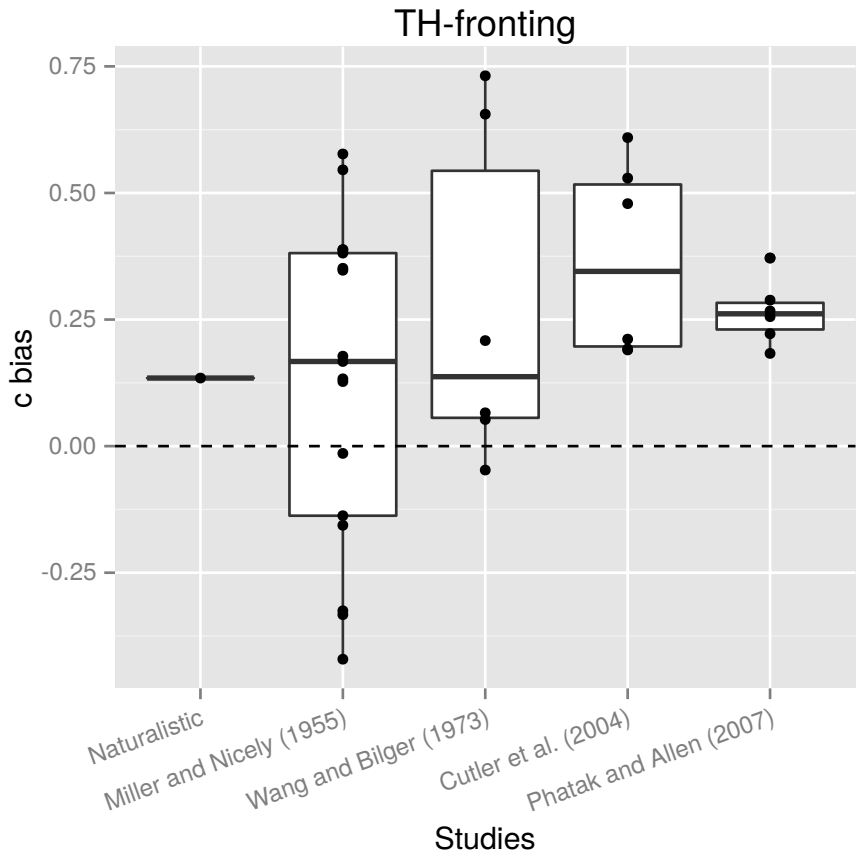


Figure 3.35: Strength of TH-fronting across naturalistic and experimental studies: the points represent the c bias values, aggregated with boxplots.

SNR levels and bandpass filters, while the others tested only SNR levels; similarly, Wang and Bilger (1973) and Cutler et al. (2004) both tested CV and VC syllables, while Phatak and Allen (2007) tested only CV syllables. Another explanation is the number of speakers and listeners involved; Phatak and Allen (2007) has the highest number of speakers ($N = 14$) and listeners ($N = 32$), while the others have one to five speakers and five to 16 listeners. More speakers and listeners used in a study could iron out any potential individual variations, thus giving a low variability of the c bias values.

Finally, some of the c bias values are negative, indicating that [f] is perceived as [θ] more often than the reverse pattern in those cases. They are six out of 17 c bias values from Miller and Nicely (1955) and one out of six values from Wang and

Bilger (1973). To clarify the variability and these negative c bias values, in the next two subsections, I will examine whether the c bias values are conditioned by specific SNR levels, and bandpass filters, or perhaps the variability and negative values are simply random noise in the data.

3.8.2.2 Noise levels

I will present the c bias value for each experimental data, and highlight and describe any potential patterns. After presenting all the experimental data, I will take all the patterns into consideration in a discussion section.

3.8.2.2.1 Miller and Nicely (1955) I will first examine Miller and Nicely (1955). Figure 3.36 shows the c bias values at different SNR levels at a fixed 200–6,500 Hz bandwidth. At 200–6,500 Hz, the bias remains positive across all SNR levels tested (-18dB to 12dB). The weakest bias is at the hardest SNR level, -18dB, with $c = 0.1329$. The strong bias is at -6dB, an immediate SNR level, with $c = 0.5457$. The plot suggests a tendency of an upside down U-shaped pattern, with extreme SNR levels (at both ends) having weaker bias; however, there is a discrepancy at +12dB or at +6dB.

3.8.2.2.2 Wang and Bilger (1973) Moving onto Wang and Bilger (1973), we compared the Noise and Quiet conditions in one CV matrix, and two VC matrices (VC1 and VC2). Recall that the two VC syllable sets have some consonants being different. The c bias values for TH-fronting are shown in Figure 3.37.

Firstly, the Noise condition in VC2 has a negative c bias, while the rest are positive. However, it is relatively small (compared to the Noise condition in VC1, c ≈ 0.20) with an absolute value of around 0.050; therefore, the negative c bias in any case is weak at best. In both VC syllable sets (VC1 and VC2), the Quiet condition also has a weak (though positive) bias which is four times lower than VC1 in the

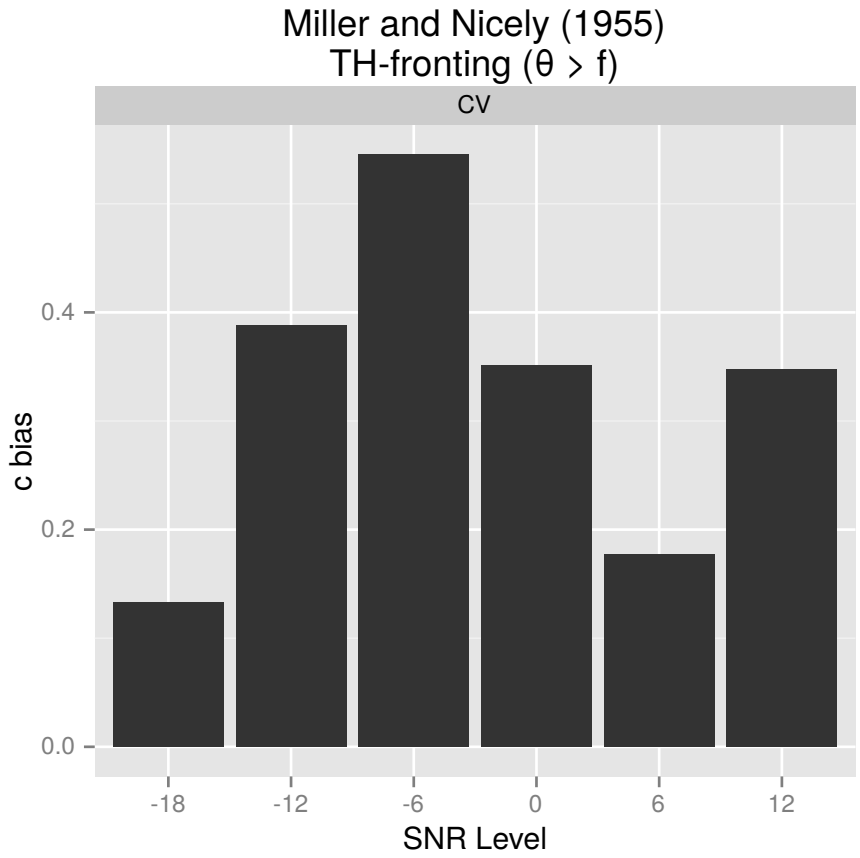


Figure 3.36: Strength of TH-fronting in Miller and Nicely (1955) with different SNR levels: the bars represent the c bias values.

Noise condition. This suggests that the TH-fronting has a weak bias in the Quiet condition in a VC syllable. Secondly, the biases in CV1 are at least three times higher than those in VC syllables (compared against the highest value in VC of c $\approx$ 0.20). This would suggest that the phenomenon is stronger in CV than VC in general. Thirdly, the values in CV1 and VC1 show that the c bias is stronger in the Noise condition than the Quiet condition.

3.8.2.2.3 Cutler et al. (2004) The c bias values in Cutler et al. (2004) are shown in Figure 3.38 for TH-fronting, comparing two syllable types, CV and VC, across three SNR levels (0, +8, 16dB). All the biases are positive across syllable types and SNR levels. However, there does not seem to be any obvious pattern with

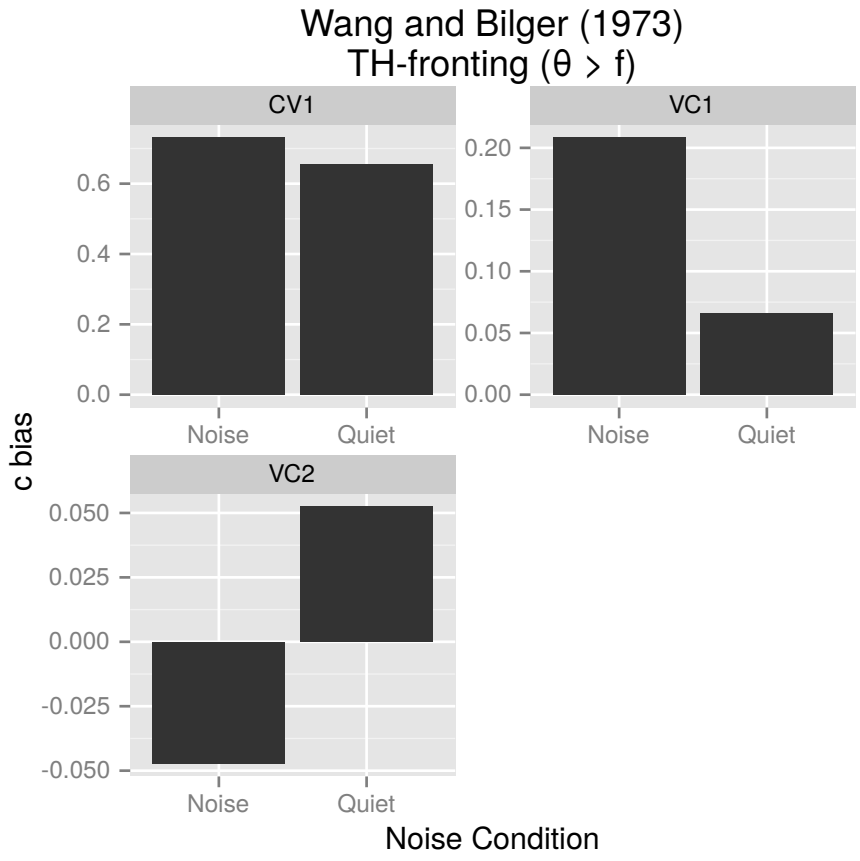


Figure 3.37: Strength of TH-fronting in Wang and Bilger (1973) with different noise conditions: the bars represent the c bias values.

the SNR levels, which could be due to the fact that it is difficult to see a pattern from only three levels of SNR. The most striking difference is the effect of the syllable type, with the VC syllable having biases that are two to three times stronger than those in the CV syllable. This is surprising because this is the opposite of what I found with Wang and Bilger (1973).

3.8.2.2.4 Phatak and Allen (2007) The c bias values in Phatak and Allen (2007) are shown in Figure 3.39 for TH-fronting, comparing six SNR levels. The c bias values are positive at all SNR levels. The c bias value appears to be dependent on the SNR levels. From the plot, there appears to be two peaks: the highest c bias value is at the Quiet condition and the second highest is at -20dB. From -22dB, the

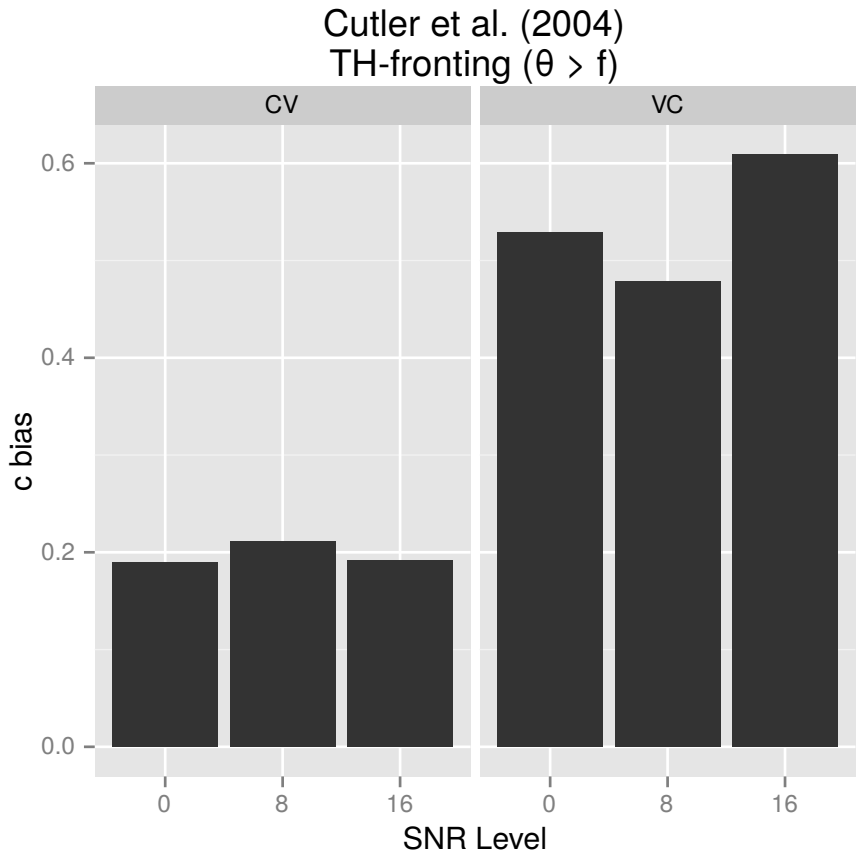


Figure 3.38: Strength of TH-fronting in Cutler et al. (2004) with different SNR levels: the bars represent the c bias values.

value raises to the smaller peak at -22dB, and then it decreases as SNR increases until -10dB; after -10dB, the value increases again all the way to Quiet.

3.8.2.2.5 Discussion Using the matrices from Miller and Nicely (1955) and Phatak and Allen (2007), our finding suggests that the c bias value for TH-fronting is dependent on the SNR levels, as shown in Figure 3.36 and Figure 3.39. Figure 3.39 suggests at least two maxima of c bias values: one in the negative SNR range and one in the positive range. This would explain the discrepancy at +6 and +12dB, such that +6dB was the inflection point and the c bias value increases to +12dB and possibly higher should Miller and Nicely (1955) have tested higher SNR levels. Further investigations are necessary to clarify this apparent pattern with experimental

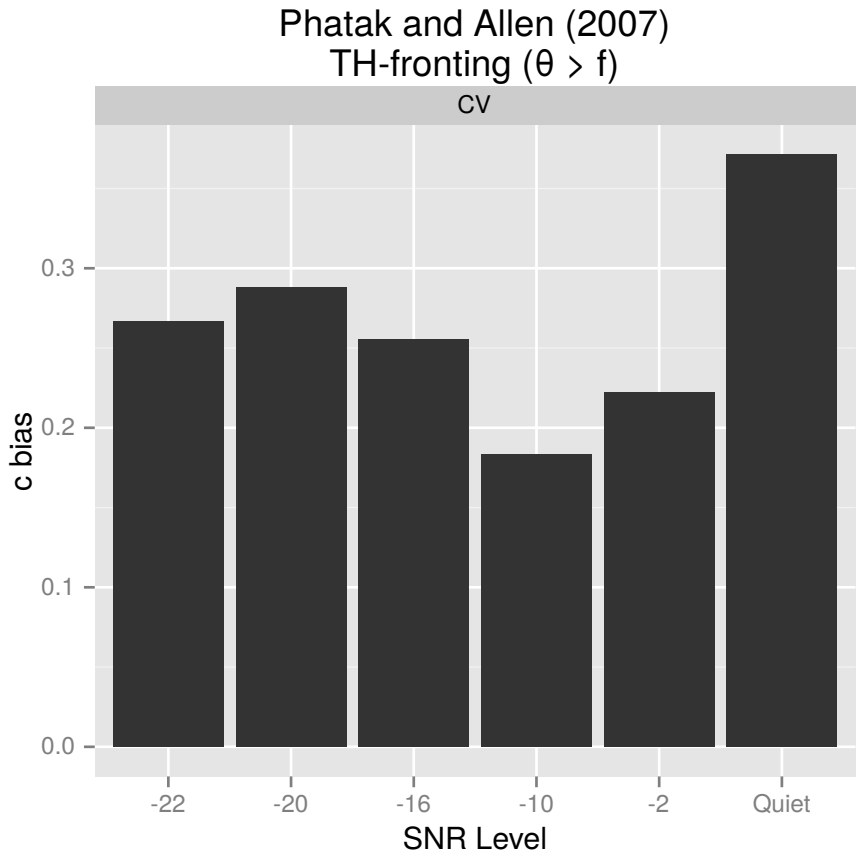


Figure 3.39: Strength of TH-fronting in Phatak and Allen (2007) with different SNR levels: the bars represent the c bias values.

studies that tested a broader SNR range.

Recall that the Noise condition has a stronger bias than the Quiet condition with Wang and Bilger (1973). This does not match with the pattern found with Phatak and Allen (2007), where the Quiet condition has the strongest bias of all of the SNR levels. One explanation for this mismatch is that the Quiet condition in Wang and Bilger (1973) has an additional manipulation, which is the signal levels, and this manipulation also presented in the Noise condition, but in a different range. Given such a confound, the difference between Noise and Quiet in Wang and Bilger (1973) is less reliable than that in Phatak and Allen (2007).

Finally, unrelated to noise levels, we found that in Wang and Bilger (1973), the bias is stronger in CV syllables than in VC syllables (Figure 3.37). However, the

pattern was reversed in Cutler et al. (2004) (Figure 3.38). It is not immediately clear why there is such a mismatch, considering that in many aspects Cutler et al. (2004) is more complete, and therefore more reliable. Firstly, Cutler et al. (2004) tested all possible CV and VC syllables (645 syllables with 24 consonants and 15 vowels), while Wang and Bilger (1973) tested only a small subset (129 syllables with 24 consonsants and three vowels). Secondly, Wang and Bilger (1973) restricted the possible consonant set to 16 phones due to technical limitations at the time. This means that if the listeners misheard sound A as sound B, but sound B is not one of the 16 allowable responses, then the listener was forced to choose a different response. This redistribution of out-of-set responses could have an effect on the asymmetries.

There is independent evidence from varieties of English and across languages that support TH-fronting being stronger at VC than CV. Firstly, the typological evidence in varieties of English suggests that, in production, TH-fronting favours word-final positions over word-initial positions. In some varieties of Scottish English, when /θ/ is in a syllable-initial/word-initial position, [θ] is favoured over [f], and when /θ/ is in a syllable-final/word-final position, [f] is favoured (Stuart-Smith and Timmins, 2006; Clark and Trousdale, 2009). Furthermore, in African American Vernacular English, TH-fronting occurs strictly in medial and word-final positions (Sneller, 2014).

Secondly, the typological evidence from various languages (Turkish, German, many Slavic languages and others) suggests that marked features for place and manner are likely to be neutralised as the unmarked values in coda positions. Kiparsky (2008) attributed this *coda neutralisation* to perceptual saliency (Steriade, 2001), such that featural distinctions have a lower perceptual salience in coda than in onset.

Having analysed the effect of noise levels (and syllable types) on TH-fronting, I will now analyse the effect of bandpass filters on TH-fronting.

3.8.2.3 Frequency bandwidth

Figure 3.40 shows the c bias values of TH-fronting with seven different low-pass filters. The filters are ordered in the plot from low to high (left to right). First of all, we see that from the left of the plot, the 200–300 Hz and 200–400 Hz bandwidths have negative c bias values, and as the bandwidth widens from 200–600 Hz upwards to 200–6,500 Hz, the bias values became positive again. It is clear that at extremely narrow bandwidths in the low frequency range, the TH-fronting bias is reversed.

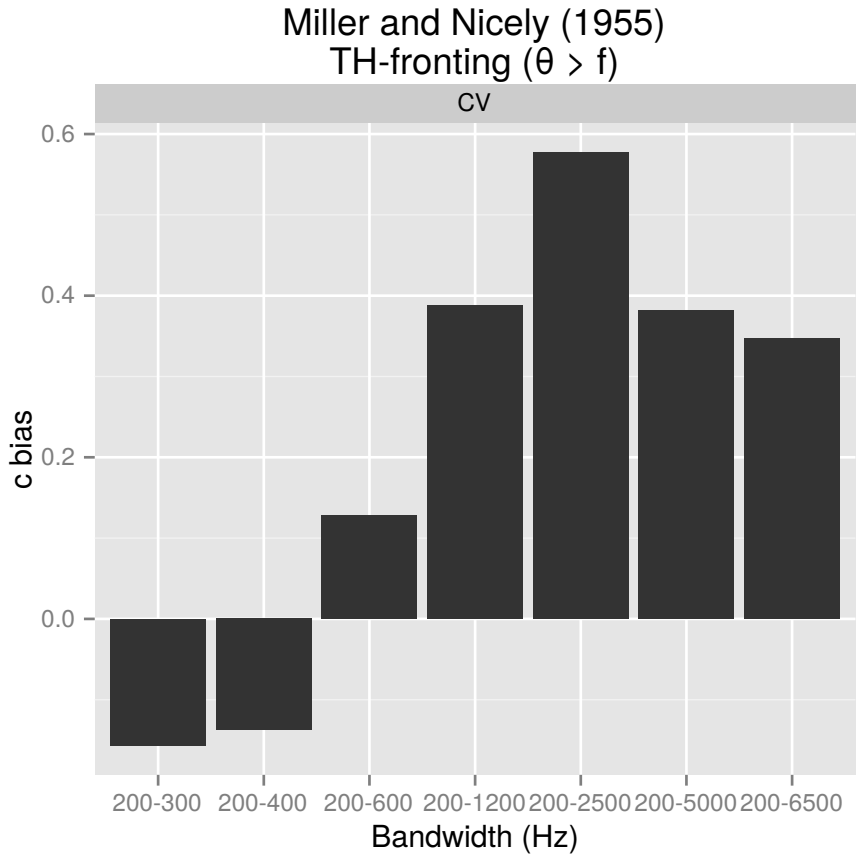


Figure 3.40: Strength of TH-fronting in Miller and Nicely (1955) with different low-pass filters: the bars represent the c bias values.

Similar observations can be made in Figure 3.40 showing the c bias values with six different high-pass filters. Three out of six of the filters resulted in a negative c bias, and they are the three narrowest ranges (3,000–5,000, 4,500–5,000, and 2,500–5,000

Hz). The next wider bandwidth, 2,000–5,000 Hz, has a weak negative c bias. Finally, the two widest bandwidths, 1,000–5,000 Hz and 200–5,000 Hz, have a positive c bias.

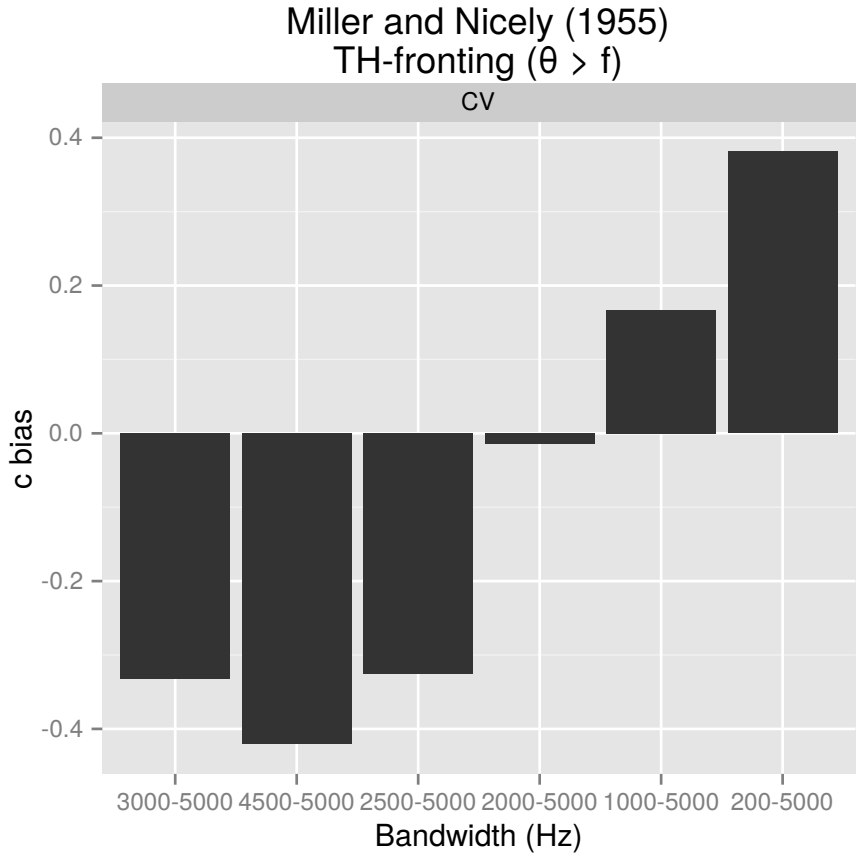


Figure 3.41: Strength of TH-fronting in Miller and Nicely (1955) with different high-pass filters: the bars represent the c bias values.

Overall, the c bias value is dependent on bandwidth manipulation. Specifically at narrow bandwidths, the direction of the bias is reversed. Given that in the naturalistic matrix and most other experimental matrices the c bias is positive, these negative c biases generated by narrow bandwidths indicate that bandwidth manipulation is less ecologically valid than adding masking noise, which resulted in positive c biases across SNRs in almost all the experimental matrices (apart from the VC2 matrix in Wang and Bilger (1973)).

Finally, it is worth considering the conditions that diverged from the overall pattern, and the possible causes for the divergence. The high-pass filtering generated

negative c bias values at 3,000–5,000, 4,500–5,000, and 2,500–5,000 Hz (see Figure 3.41). Recall that [f] has a lower intensity range (3,000–4,000 Hz) than [θ] (7,000–8,000 Hz). Given that the broadest bandwidth (200 – 6500 Hz) used by Miller and Nicely (1955) already excluded the intensity range of [θ], the fact that listeners prefer [θ] over [f] is puzzling, since the intensity cue for [θ] is missing.

One explanation is that listeners rely on the contrast of the absence of low frequency components and the presence of high frequency components. The effect of high-pass filtering is that only high frequency components can be found in the signal and no low frequency components can be found. The bandwidths that created the divergence (3,000–5,000, 4,500–5,000, and 2,500–5,000 Hz) are in the high frequency range. Together, the presence of high frequency components provides evidence for [θ] that has a high intensity range (regardless of the actual range of [θ]), and the absence of low frequency components provide no evidence for [f] which has a low intensity range (regardless of the actual range of [f]). Therefore, listeners have a preference for [θ] over [f] in the narrow frequency bandwidths that are in the high range.

The low-pass filtering has two diverged bandwidths, 200–300 and 200–400 Hz. Although I have no immediate explanation for their negative c bias values, they are likely to be negligible, because their values are relatively small (around -0.1) compared to the other diverged bandwidths (3,000–5,000, 4,500–5,000, and 2,500–5,000 Hz) which have values that are three times lower (around -0.3).

3.8.3 Velar nasal fronting

Velar nasal fronting is the process of a velar nasal /ŋ/ being realised as an alveolar nasal [n]. This phenomenon has been extensively studied in sociolinguistics (commonly known as the ING variable (Chambers, 2003)). It has been extensively studied across a variety of English dialects (see Wagner (2008, Ch. 4) for an extensive list

of references). Velar nasal fronting is a stable variable. Its application rate has been posited by Labov (1994a) to follow a hierarchy (verb > adjective > gerund > noun), with the verb forms (namely, the suffix -ing) showing that highest rate.

Houston (1985) summarised a historical account for velar nasal fronting. The process is said to reflect a historical morphological alternation between the verbal noun suffix <ing> and the present participle suffix <inde>. The historical account comprises two core concepts. The first is a theory of phonetic levelling and the second is a theory of functional shift (a process of syntactic syncretism). I will review the theory of phonetic levelling below and leave the theory of functional shift aside.

Within the theory of phonetic levelling, there are multiple explanations. One of which is that the present participle suffix underwent a stop deletion; therefore, <inde> [ind] was pronounced as [in], and by assuming that <ing> had the pronunciation [iŋ], we could say that the stop deletion of [d] triggered the confusion between <ind> and <ing>, which led the merging of both suffixes as <ing> [iŋ].

Another final stop deletion account is that both suffixes had a final stop before the merger, such that <inde> was pronounced as [ind] and <ing> as [ing]. After the process of final stop deletion, both suffixes had the pronunciation [in], which serves as a trigger for the merging of both suffixes as <ing>.

Alternative to the final stop deletion accounts, some researchers considered that <ing> had multiple pronunciations [ng, nk, n, ŋ] and <inde> also had multiple pronunciations [nd, nt, n]. The confusions between the two suffixes were due to the overlapping pronunciations, which were likely to involve [n] and [ŋ], because the stop variants were too distinctive for a confusion to occur.

In addition, an articulatory explanation is that the high vowel in <ing> pulled the velar nasal [ŋ] forward, thus turning it into an alveolar nasal [n] as a result of co-articulation. However, this cannot explain why the change was never completed, as otherwise we should see <in> as the standard orthography variant and [in] as the

standard pronunciation.

In sum, the theory of phonetic levelling is inconclusive. First of all, the pronunciations of <ind> and <ing> are unclear, given we only had orthographical evidence; therefore, we cannot be sure if they had final stops [d] and [ɡ]. This casts doubts on those accounts that rely on the deletion of the stops as the initial confusion trigger. Furthermore, some researchers have suggested that the change in either direction [ŋ] > [n], and [n] > [ŋ], or simply not having an explanation for the direction.

As an alternative to the inconclusive historical account of velar nasal fronting, either as a consequence of independent processes (such as stop deletion) or co-articulation (fronting of the velar nasal due to the front vowel), velar nasal fronting could perhaps be explained in terms of final place neutralisation found in diachronic sound change. Diachronically, final place neutralisation is a common phenomenon. When nasals are involved, all the nasals [m, n, ŋ] have a tendency to be neutralised as a velar nasal [ŋ]. To support this, Blevins (2004, pp. 120–122) gives evidence from Chinese dialects, where some/all of the final nasals [m, n, ŋ] in Middle Chinese are velar nasals [ŋ] in other sister dialects, such as Fuzhou. Blevins suggests that the neutralisation is a kind of perceptual-based sound change. Firstly, due to lenition, the three nasals are often produced as nasal glides. Secondly, listeners have a strong perceptual bias to perceive nasal glides as velar nasals rather than nasals at other places of articulation. Together, this led to a perceptual-based sound change – [m, n, ŋ] > a nasal glide, due to lenition in production; and the nasal glide > [ŋ] due to their perceptual similarity. However, this perceptual-based sound change cannot explain the direction of velar nasal fronting, which is from [ŋ] to [n], and not the reverse.

Perhaps this perceptual-based sound change of [m, n, ŋ] > [ŋ] is not comparable to velar nasal fronting because it is a diachronic change, while velar nasal fronting is a synchronic change; therefore, it could be beneficial to examine synchronic evidence of

velar nasal fronting in other languages. Velar nasal fronting can be found in varieties of Mandarin Chinese where [ŋ] is produced as [n], and this is especially common in Southern China and Taiwan (Yang, 2010). Yang (2010) conducted a production study, testing the production errors between [ŋ] and [n] in coda positions in Mandarin. This study also compared two groups of participants: those from Taiwan and those from mainland China. Their results showed that, firstly, the production errors are conditioned by the preceding vowel [i, ə] and, secondly, there is a group difference between the Taiwanese participants and Mainland participants. On the one hand, the participants from Taiwan have above 95% error rate where /ŋ/ is produced as [n] after [i, ə], suggesting that there is a complete merger of /ŋ/ and /n/ as /n/ in coda positions after [i, ə]. On the other hand, the participants from mainland China had ≈ 40% error rate in *both* directions after [i], and the rate is negligible after [ə]. The error patterns from both groups together suggest that velar nasal fronting in Mandarin Chinese is conditioned by a similar phonological environment as that in English, because in English it occurs most frequently with <ing> which contains also a front high vowel; although in English, it is further conditioned by the grammatical class. This recurrent phonological condition for velar nasal fronting in both English and Mandarin Chinese lends support to the articulatory explanation mentioned above, which states the velar nasal [ŋ] is pulled forward into an alveolar nasal [n] by the high front vowel. Furthermore, in Mandarin Chinese, the confusion between [ŋ] and [n] in production is asymmetrical from [ŋ] to [n], which matches the velar nasal fronting phenomenon in English. Although the error rates are approximately equal in both directions for the mainland participants, I consider this as a sound change in process (/ŋ/ and /n/ as /n/), and that the errors of /n/ as [ŋ] were the result of hypercorrection.

In sum, previous accounts of the velar nasal fronting phenomenon in English considered it a consequence of final stop deletion in the verbal noun suffix <ing> and

the present participle suffix <inde>, and a result of co-articulation (fronting of the velar nasal due to the front vowel). The phenomenon is not compatible with accounts of diachronic sound change of final place neutralisation as they predict the change being in the opposite direction. Besides English, Mandarin Chinese also has velar nasal fronting, and the phenomenon shares a similar phonological environment as that in English. Although the change can move towards completion in some varieties of Mandarin Chinese (such as those in Taiwan), velar nasal fronting in English was previously suggested to be a stable process that is unlikely to be completed.

The fact that velar nasal fronting exists in both Mandarin Chinese and English, and that they share a similar phonological environment, would suggest that velar nasal fronting could have a phonetic basis and not simply be the result of some accidental historical change. It is worth noting that I am not proposing a *purely* articulatory explanation for velar nasal fronting, as it is morphologically conditioned. However, it is likely to be one of the factors that promotes this change. In fact, another factor, segmental frequency, will be examined in Chapter 4, Section 4.2.3.

On top of the articulatory account which was mentioned earlier, I propose that there is a perceptual bias that contributes to velar nasal fronting. In the following sections, I will examine both the naturalistic and experimental confusion data to see whether the perceptual confusion between [ŋ] and [n] is asymmetrical or not, and whether the direction is the same as that of velar nasal fronting [ŋ] > [n].

3.8.3.1 Overview

The naturalistic data being examined are the same context-free consonant confusion matrix. The experimental data being examined are some of those studies mentioned in Section 3.7.1. The experimental matrices are:

- One matrix from Wang and Bilger (1973) with one of the two VC sets (VC2) and two noise conditions (Noise and Quiet).

- Three matrices from Cutler et al. (2004) with the VC syllable type across three SNR levels.

Just as with the analyses of TH-fronting, the c bias measure from the naturalistic matrix was calculated, and each of the experimental matrices. The corresponding c bias values were plotted in Figure 3.42 as boxplots, separated by studies on the x-axis, with c bias values plotted on the y-axis as both individual points and box plots. The main finding is that the c bias values are positive for the naturalistic matrix as well as all four experimental matrices. A positive c bias value means that the perceptual confusion favours [n] over [ŋ]: that is, [ŋ] is perceived as [n] more often than the reverse.

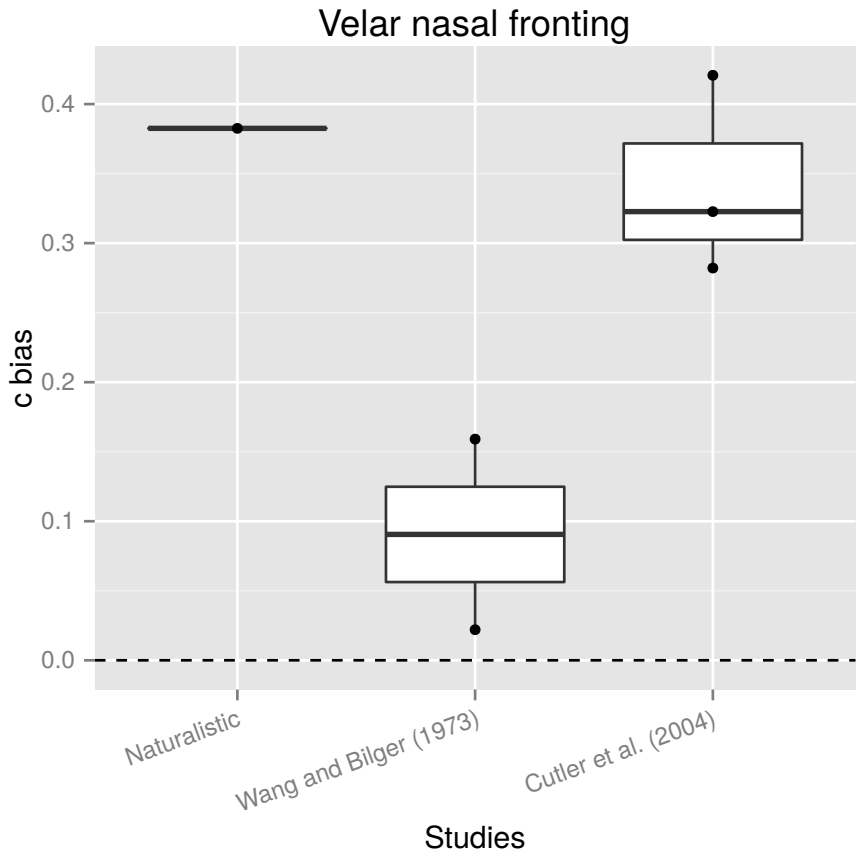


Figure 3.42: Strength of velar nasal fronting across naturalistic and experimental studies: the points represent the c bias values, aggregated with boxplots.

Without a sufficient amount of data points, all the c bias values from *both* the naturalistic and the experimental studies ($N = 6$) are analysed with a one-sample Wilcoxon signed rank test, testing the hypothesis that the true value is zero. We found that the c bias values (the mean value: 0.2649) are significantly different from zero, with $p = 0.03125$ (two-tailed), indicating that it is a significant positive c bias. This supports the view that velar nasal fronting is motivated by a perceptual bias. The next section will examine the effect of noise levels on the asymmetrical pattern; however, since there are only two experimental studies, Wang and Bilger (1973) and Cutler et al. (2004), and neither of them have an extensive coverage of SNR levels, the findings should be regarded as tentative.

3.8.3.2 Noise levels

Figure 3.42 showed that the median c bias value is relatively small with the two matrices from Wang and Bilger (1973).

Looking more closely at the breakdown of the c bias values in Figure 3.43, it is clear that the noise condition has an extremely low c bias value (albeit positive), and that the Quiet condition has a much higher c bias value. This suggests that the asymmetrical pattern can be found by manipulating the signal levels, without adding masking noise to the stimuli, and that the noise manipulation tested by Wang and Bilger (1973) (-10dB to +15dB with white noise) might be too severe for revealing this asymmetrical pattern.

Let us move on to the three matrices from Cutler et al. (2004). Figure 3.44 shows that the c bias value decreases steadily as SNR increases. Furthermore, the lowest c bias value, which is around 0.25 at +16dB, is already higher than the highest c bias value of Wang and Bilger (1973). This suggests that the asymmetry is particularly robust with multi-talker babble as the masker.

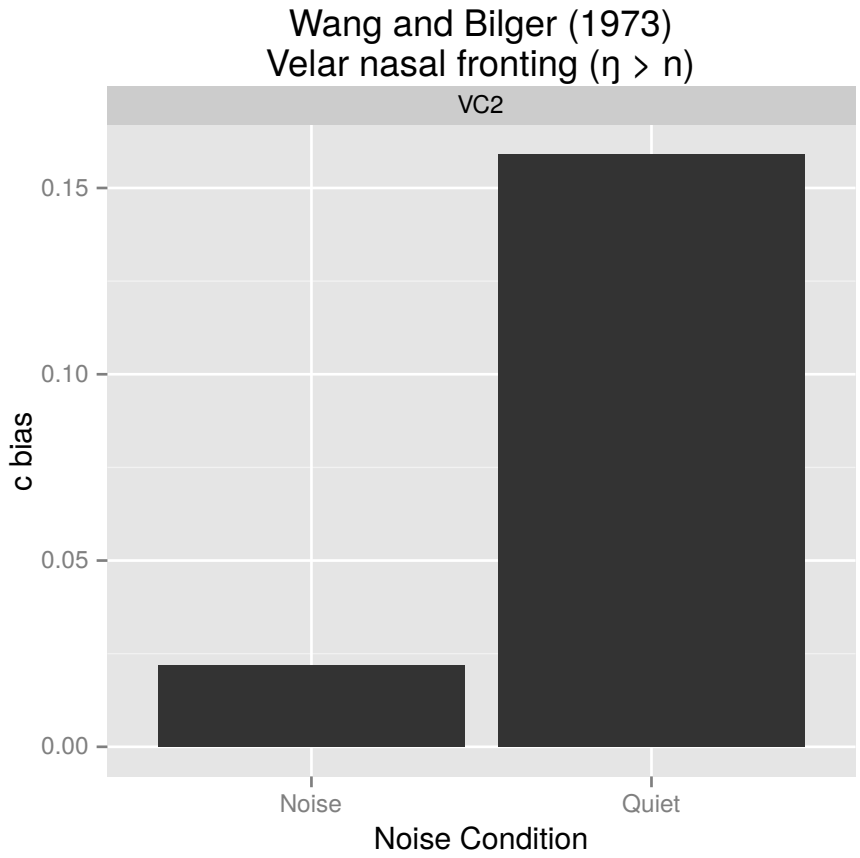


Figure 3.43: Strength of velar nasal fronting in Wang and Bilger (1973) with different noise conditions: the bars represent the c bias values.

3.8.4 Back vowel fronting

Labov (1994a, p. 116) proposed three principles of vowel chain shifts. The third principle is that back vowels move to the front. A later version of the principle was that tense vowels move to the front along peripheral paths and lax vowels move to the back along non-peripheral paths (Labov, 1994a, p. 200). Each of the two versions of the third principle makes a different prediction about the direction of vowel movement. In the following sections, I will examine the first version of the third principle, which predicts that the back vowels move uniformly to the front of the vowel space, possibly at the same height.

Labov (1994a, p. 273) used his data on naturally occurring understandings (a

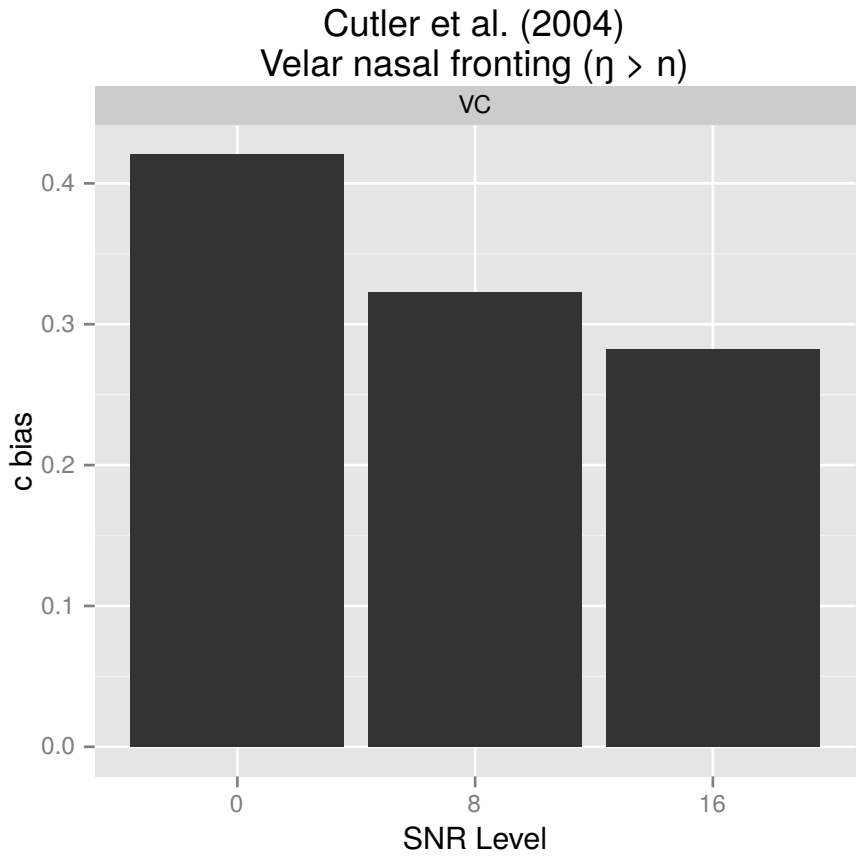


Figure 3.44: Strength of velar nasal fronting in Cutler et al. (2004) with different SNR levels: the bars represent the c bias values.

subset of the naturalistic corpus – the Labov corpus) to support the principles posited for chain shifts. He argued that the confusion data cannot be used to support the idea that confusions are the causes of sound change, as they could equally be the result of the change; therefore, the evidence is ambiguous in terms of causality. In any case, the confusion patterns can nonetheless reflect the patterns predicted by the principle of sound shifts.

Using an open-set response task, Benkí (2003) collected misperception data of CVC nonsense syllables, masked with noise at four SNR levels. He conducted a similar analysis, examining the asymmetrical patterns in vowel confusions. From the confusion patterns for the following pairs of vowels, [u, i], [e, o] and [ɑ, æ], he found that the [u] and [o] were perceived as [i] and [e] respectively more often than

the reverse: that is, there is a fronting pattern; while [æ] was perceived as [ɑ] more often than the reverse, thus reflecting a backing pattern. The author attributed this to the second version of Labov's third principle of chain shifts, which, as mentioned above, states that tense vowels move to the front, while lax vowels move to the back, and his data do fit this quite well, given [u] and [o] are back tense vowels and [æ] is commonly accepted as being a lax vowel.

Following in the footsteps of Benkí (2003), I examined seven pairs of front and back vowels of the same height. They are [ʉ > i], [u > i], [ʊ > ɪ], [o > e], [ɔ > ɛ], [ɑ > æ] and [ɑ > a]. Please note that ">" is used to encode the *assumed* directionality for a given pair of segments. If the c bias is positive, then the asymmetry is the same as the direction of the arrow; if the c bias is negative, then the asymmetry is in the opposite direction.

However, unlike Benkí (2003), both lax and tense vowels were included, which added the two lax vowel pairs [ɔ > ɛ] and [ʊ > ɪ]. Furthermore, [ɑ > a] was included, since [ɑ] and [a] are more closely matched for height than [ɑ] and [æ]. Finally [ʉ > i] was included only for the analyses of the naturalistic data and not the experimental data, because the experimental data did not examine [ʉ].

The naturalistic data being examined are the same context-free consonant confusion matrix. The experimental data being examined are the six matrices from Cutler et al. (2004) with two syllable types (CV and VC), across three SNR levels.

Let us start with the naturalistic data. Figure 3.45 summarised the c bias values of the seven pairs on the y-axis, while the x-axis is the vowel pairs ordered by vowel height (high to low). The figure shows that five of the seven pairs have a positive c bias, indicating a fronting pattern. Among these five pairs with a positive c bias, [ʉ > i] and [ʊ > ɪ] have a relatively strong bias. This pattern fits well with the fact that the back vowels [ʉ] and [ʊ] are more front than the other back vowels. One could therefore speculate that the strength of confusion asymmetries of two phones

is dependent on their acoustic distance. The two tense pairs [u > i] and [o > e] have a positive c bias, which matches the findings by Benkí (2003). More specifically, the c bias is stronger with [u > i] than with [o > e] in both the current study and Benkí (2003). In the current study, the bias with [ɑ > æ] is negative (recall that a negative c bias for [x > y] means [y] is perceived as [x] more often than the reverse). Although this pair was also found to be negative in Benkí (2003), the c bias in the current study is negligible (barely visible in the plot) compared to other pairs. Finally, [ɑ > a] also has a negative c bias and the strength is also relatively weak. Overall, the confusion pattern reflects a fronting of back vowels, and this supports Labov's third principle of chain shifts.

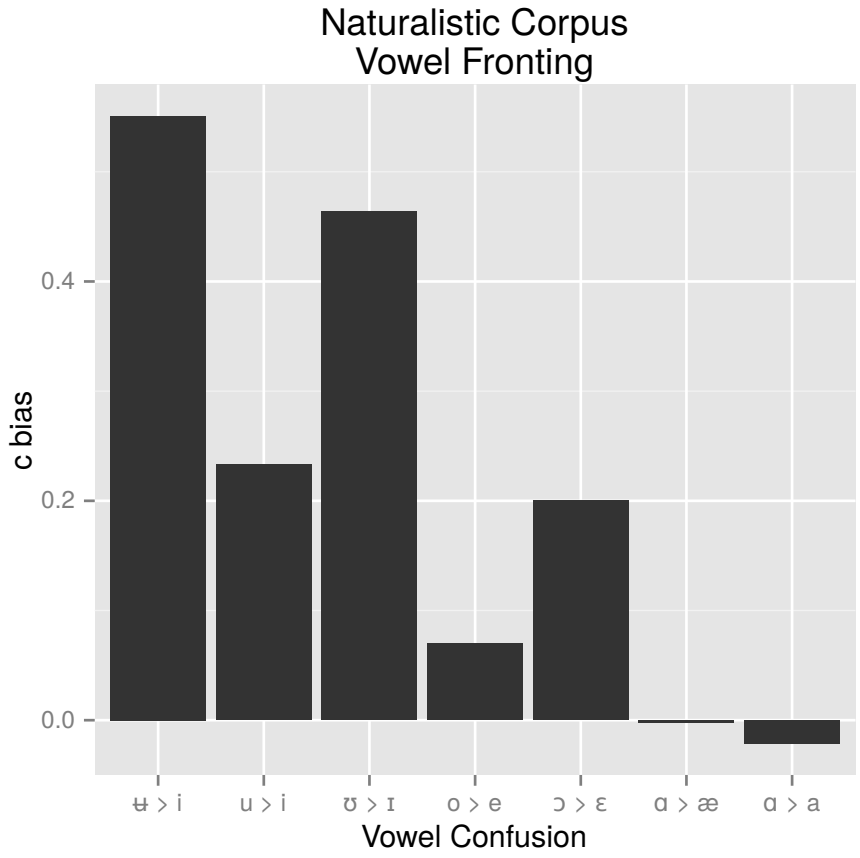


Figure 3.45: Strength of vowel fronting in naturalistic corpus: the bars represent the c bias values.

Let us move on to the experimental data. Figure 3.46 summarised the c bias

values of the six pairs (the seven pairs mentioned above but excluding [ʉ > i]). There are two subfigures, with CV on the left and VC on the right. Each subfigure shows the c bias values on the y-axis, and the SNR levels on the x-axis. At each SNR level, the c bias values of the seven vowel pairs are shown as bars ordered by vowel height (high to low). First of all, some vowel pairs had their c values set to zero because they contain no confusions in either direction (this treatment was mentioned in the method section, Section 3.8.1). Let us focus on the pairs with a non-zero bias value. We can see that the overall pattern diverges from both the naturalistic data and Benkí (2003).

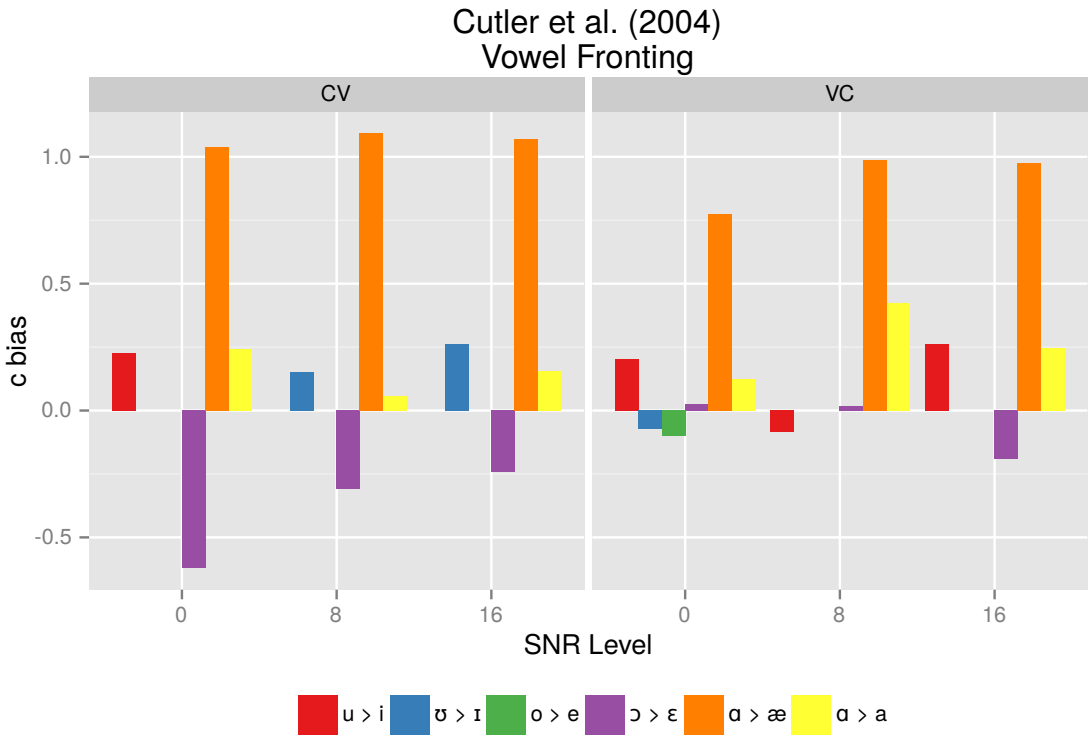


Figure 3.46: Strength of vowel fronting in Cutler et al. (2004) with different SNR levels in CV and VC syllables: the bars represent the c bias values of the seven vowel pairs (with an assumed direction).

Firstly, [ɑ > æ] has an extremely high positive c bias across syllable types and SNR levels. Similarly, [ɑ > a] also has a positive c bias across conditions, although

it is less strong than [ɑ > æ]. The robust positive c bias with [ɑ > æ] and [ɑ > a] is unexpected, given the findings in the naturalistic data and Benkí (2003). Secondly, [u > i] and [ʊ > ɪ] are relatively robust among the remaining four pairs, with most of the non-zero bias values being positive. The bias directions of these two pairs matched those found previously in naturalistic and experimental settings. Finally, the lax vowel pair [ɔ > ɛ] has a strong and consistent negative c bias, especially in CV positions. This is the opposite of what we found in the naturalistic data, and I have no immediate explanations for such a divergence.

In sum, the naturalistic data showed a strong fronting pattern of back vowels, and this matches the experimental study by Benkí (2003) using CVC syllables. On the whole, the experimental data from Cutler et al. (2004) also showed a fronting pattern but the pattern is less robust.

3.8.5 Conclusion

This section examined three asymmetrical patterns in both the naturalistic and experimental data. They are TH-fronting, velar nasal fronting and back vowel fronting.

Section 3.8.2 examined TH-fronting. First of all, putting the asymmetry aside, we saw that [θ] and [f] are perceptually most similar to each other, among all the phones, and this was supported by two out of three of the hierarchical clustering trees (Figure 3.19 and Figure 3.20) which showed that [θ] and [f] are of the same hierarchical cluster. The clustering of the phones is a structural reflection of the naturalistic confusion patterns, which often recapitulates natural classes.

Most experimental conditions across three studies showed a statistically significant positive c bias, indicating that overall TH-fronting is robust experimentally. The naturalistic bias is consistent with the experimental bias, in that they are both positive.

However, some of the experimental conditions have a negative c bias, almost all

of which are the result of extremely narrow bandpass filtering: e.g. six out of 17 c bias values from Miller and Nicely (1955). Given that in both naturalistic and most experimental conditions, the TH-fronting bias is positive, this would suggest that bandpass filtering is less ecologically valid than adding masking noise which resulted in positive c biases. The findings in Section 3.7 led to the same conclusion regarding bandpass filtering.

Furthermore, TH-fronting was found to be two to three times stronger in VC syllables than CV syllables in Cutler et al. (2004). However, we found the reverse pattern in Wang and Bilger (1973). I argued that the reverse pattern in Wang and Bilger (1973) is less reliable, given Cutler et al. (2004) is more complete, in that all possible CV and VC syllables were tested, than Wang and Bilger (1973). Furthermore, independent evidence from varieties of English and across languages support TH-fronting being stronger at VC than CV. Varieties of English (e.g. Scottish English and African American Vernacular English) have a restriction on where TH-fronting can occur, such that word initial positions are often disallowed or not preferred. Finally, cross-linguistic patterns of *coda neutralisation* would predict that featural distinctions are often lost in coda positions rather than in onset positions, perhaps due to their relative perceptual saliency (Steriade, 2001) with codas being perceptually less salient.

Finally, comparisons with Miller and Nicely (1955) and Phatak and Allen (2007) suggest that the c bias value for TH-fronting is dependent on the SNR levels, with a function that contains two maxima of c bias values. But further comparisons have to be made with additional studies (such as Phatak, Lovitt, and Allen (2008) and future studies which examine a wide range of SNR levels) to substantiate this claim.

Section 3.8.3 examined velar nasal fronting. The naturalistic matrix and all the experimental matrices have a positive velar nasal fronting bias, which is statistically significantly different from zero (no bias), thus suggesting that it is a robust bias.

The biases of velar nasal fronting are consistently stronger in Cutler et al. (2004) than in Wang and Bilger (1973), which suggests that a white noise masker (used by Wang and Bilger (1973)) might be too severe for revealing the velar nasal fronting bias. We again found that the c bias value is a function of the SNR levels, with the c bias value decreasing as the SNR level increases.

Section 3.8.4 examined back vowel fronting. The naturalistic data showed a strong fronting pattern of back vowels, and this matches the experimental study by Benkí (2003) using CVC syllables. Interestingly, Benkí (2003) found that [ɑ > æ] has a negative c bias, and both [ɑ > æ] and [ɑ > æ] were also found to have a negative c bias in the naturalistic data; however, the strength of the biases were negligible. The experimental data by Cutler et al. (2004) showed multiple divergences. Firstly, [ɑ > æ] and [ɑ > æ] both have a consistently strong positive c bias. Secondly, the lax vowel pair [ɔ > ɛ] has a consistently strong negative c bias, especially in CV positions.

Overall, the results of all three asymmetrical patterns are encouraging, showing that these patterns are robust across most experimental conditions (syllable types, SNR levels and bandpass filters) and, crucially, they are also found in naturalistic settings. This reinforces the ecological validity of experimentally induced perceptual confusions as evidence for the framework in which the listener is a source of sound change (Ohala, 1981; Ohala, 1989). It is worth noting that an alternative account based on top-down factors for asymmetrical patterns is proposed later in Chapter 4, Section 4.2.3.

3.9 Conclusion

The focus of this chapter is to examine the bottom-up phonetic and phonological factors that play a role in naturalistic misperception.

In Section 3.3, the analytical techniques that are commonly used to analyse confusion matrices were summarised. The section described methods for converting count data to perceptual distances, as well as techniques for comparing distance matrices on both global and structural levels. In particular, a methodological contribution was made on smoothing sparse matrices of misperception, namely the iterative Witten-Bell smoothing method which has an advantage being entirely data-driven.

In Section 3.4, we identified whether there were any phonetic and phonological biases in naturalistic misperception on a featural level: place, manner and voicing for the consonants, height and backness for the vowels. Firstly, we found that an account that is based on sonority or the availability of acoustic cues can explain the perceptibility (as indicated by the confusion rate) of the different voicing values (voiced and voiceless), as well as different manners of articulation (glide, liquid, nasal, fricative and stop but not affricate), but not the different places of articulation. In fact, the perceptibility of place is best accounted for with a combination of two phonological theories: the underspecification of coronal (Lahiri and Reetz, 2002) and the place markedness scale (Lombardi, 2002). We found that again sonority cannot explain the perceptibility of vowel height and backness, because their confusions are asymmetrical. The asymmetrical patterns of vowel height can be explained using Steriade's (2001) account of perceived similarity, while the asymmetrical patterns of vowel backness were analysed in more depth in a later section.

Section 3.5 and Section 3.6 quantified how much of the naturalistic segmental confusions of the consonants and of the vowels is purely phonetic/phonological. To do so, the distances between any two segments were compared based on their confusability, with the distances based on acoustic measurements (for the vowels) and feature values (for the consonants). The distances of both global and structural levels were compared using correlation tests and visualisations of the projected structures. The results of the correlation tests and visualisations show that a substantial por-

tion of the vowel confusions can be explained purely with acoustic distances of the vowels. Based on the correlation tests alone, there is a low similarity between the confusion-based distances and phonetic/phonological distances for the consonants, thereby indicating the processes involved in consonant perception are much more complex than the vowels. However, when examining the visualisations, the projected structures of the consonant confusions revealed multiple phonetic dimensions – such as sonorance, spread glottis, voicing, frication, nasality, liquid, sibilancy and duration – which suggests that some phonetic biases are involved.

Section 3.7 examined the ecological validity of specific experimental manipulations that are used in experimentally induced misperception studies by comparing experimental data of previous studies to the naturalistic corpus on both global and structural levels. In terms of SNR levels, it was found that extreme SNR levels (too high or too low) tend to be the least similar to naturalistic conditions. This reflects the fact that at extremely low SNR levels the signals are too degraded to reveal non-random confusions, while at extremely high SNR levels, the signals are not degraded enough to induce confusions. More specifically, the correlations with the matrices at different SNR levels and the naturalistic matrix showed that the relationship between SNR levels and the similarity level with the naturalistic matrix has a upside-down U-shaped function. The location of the peak correlation (i.e. the SNR with the highest correlation) varies from study to study, and it is likely to be dependent on the masking noise types. In terms of the bandwidth manipulation, it was found that the peak correlations tend to be the ones with the broadest bandwidth, thus suggesting that bandwidth manipulation is less ecologically valid than SNR manipulation. Given that there is a relationship between experimental manipulation (at least with SNR levels and bandwidth filtering) and the amount of correlation with the naturalistic matrix, the naturalistic matrix can therefore serve a benchmark corpus, with which the ecological validity of various experimental manipulations can be evaluated.

Section 3.8 examined three asymmetrical patterns, which are TH-fronting, velar nasal fronting and back vowel fronting, in both the naturalistic and experimental data. Overall, both the naturalistic and the experimental data indicate that perceptual confusions contain asymmetries that mirror these asymmetrical patterns. The strength of these asymmetries is a function of experimental conditions, such as SNR and bandwidth manipulations. Among the experimental conditions with a negative bias, almost all were due to narrow bandpass filtering. This again suggests that bandwidth manipulation is less ecologically valid than SNR manipulation.

With TH-fronting, the bias was found to be stronger in VC syllables and in CV syllables in Cutler et al. (2004). This is supported by independent evidence from the restriction of TH-fronting in some English dialects in which word-initial positions (not medial and final positions) are often disallowed or not preferred. Coda neutralisation across languages is argued to be motivated by perceptual factors (Steriade, 2001), such that coda positions are less perceptually salient than onset positions, and this is reflected in the perceptual bias of TH-fronting.

With velar nasal fronting, the naturalistic matrix and all the experimental matrices have a positive velar nasal fronting bias. It was found that using multi-talker babble noise as a masker can generate a stronger positive bias than using white noise and that the strength of the bias with multi-talker babble noise is similar to that with the naturalistic data. This reflects the fact that multi-talker babble noise is a much more realistic masker than white noise.

With back vowel fronting, a positive bias was found with most of the vowel pairs with a front vowel and a back vowel using the naturalistic data. The exceptions were the low vowels [ɑ, æ] and [ɑ, æ], which have a negligible negative bias. These matched the experimental study by Benkí (2003) which tested CVC syllables. However, the experimental data by Cutler et al. (2004) showed that [ɑ, æ] and [ɑ, æ] both have a consistently strong positive bias across SNR levels and syllable types. Together, the

three asymmetrical patterns found in misperception support Ohala's framework of sound change in which the listeners are a source (Ohala, 1981; Ohala, 1989).

To conclude, the three sets of analyses in this chapter have demonstrated that naturalistic misperception has a definite phonetic and phonological basis, even at the lowest level of confusion matrices, both featurally and segmentally on both global and structrual levels. Concretely, on a featural level, the misperception trends have a phonetic and phonological explanation. On a segment level, vowel confusions have a stronger phonetic/phonological bias than consonant confusions. By setting the naturalistic matrix as an ecological benchmark, experimental matrices from four different studies were compared with the naturalistic matrix on both global and structural levels. The comparisons revealed that the relationship between the listening conditions and the amount of similarity is not random, and that some listening conditions are more ecologically valid than others. Finally, our analyses of asymmetrical patterns confirm that certain sound change patterns are motivated by perceptual asymmetries, and crucially, these patterns are found in naturalistic settings and are not confined to specific experimental conditions.

Chapter 4

Top-down lexical factors

4.1 Introduction

This chapter aims to examine the top-down lexical factors that play a role in naturalistic misperception. Some of the previous analyses of naturalistic misperception using the sub-corpora of the our combined mega corpus have identified a few top-down lexical factors such as segmental frequency (Bird, 1998), syllable factors (Browman, 1978) and word frequency (Bond, 1999; Vitevitch, 2002; Tang and Nevins, 2014). Overall, their findings were encouraging, suggesting that top-down lexical factors do have an effect on naturalistic misperception, and that they are consistent with experimental findings. However, given that the data were collected by different people, it is possible that their findings are susceptible to certain idiosyncrasies due to reporting biases. Furthermore, the amount of naturalistic data used by these studies was small; therefore, it is possible that their findings are due to chance. These drawbacks highlight the need for a reanalysis of these findings using the combined corpus. In addition to the three top-down factors, mentioned above (segmental frequency, syllable factors, and word frequency), the effect of the conditional probability of a word in an utterance was also examined (which, in information theory, is referred

to as self-information, Shannon, 1948). This chapter will examine whether there are top-down effects from linguistic units of various sizes – segments, syllables, words, and utterances. If so, how strong are these effects? These four factors serve as four main sections in this chapter. Each of the four sections is introduced below, starting with segmental frequency.

4.1.1 Segmental frequency

The role of segmental frequencies in segmental confusions will first be examined. Segmental frequency is the frequency of the occurrence of segments found in a large sample of the language. I selected two aspects of segmental confusions that could be explained with segmental frequency.

4.1.1.1 Target and response biases

The first aspect of segmental confusions concerns the target and response biases in misperception. Target bias means that certain phones are more (or less) likely to be spoken but misperceived. That is, given there is a misperception, not all segments are equally likely to be the target. Similarly, response bias means that certain phones are more (or less) likely to be the resultant perceived phones in a misperception. That is, given there is a misperception, not all segments are equally likely to be the response. Can the biases (if any) be explained by the segmental frequency in the language?

It is important to understand that the target bias means that certain phones are more likely to surface as the target of a misperception, and it does *not* mean that certain phones are more likely to undergo misperception. In other words, target bias is referring to the probability of a phone being the target segment of a misperception, and it is *not* referring to the probability of a phone being erroneously misperceived. To further clarify what target and response biases are, let us consider the following

example. 100 phones were presented to a listener and 40 phones were misperceived. Amongst these 40 phones (the intended segments), what is the distribution of the intended segments? Can the distribution of the intended segments be predicted by the distribution of their segmental frequency in the language? These 40 phones were misperceived as another 40 phones (the perceived segments). What is the distribution of the perceived segments? Can the distribution of the perceived segments be predicted by the distribution of their segmental frequency in the language?

To examine this, the frequency of being a target in a misperception will be computed for each segment type. This will then be correlated with the frequency of each segment type found in the language. A significant correlation would mean that when a segment is misperceived, the likelihood of this segment being segment x is dependent on how frequent segment x is in the language. Similarly, for the response bias, the frequency of being a response in a misperception will be computed for each segment type, and will then be correlated with the frequency of each segment type found in the language. A significant correlation would mean that when a segment is misperceived, the likelihood of the perceived segment being segment x is dependent on how frequent segment x is in the language.

4.1.1.2 Asymmetrical confusion

The second aspect of segmental confusions concerns their asymmetrical patterns. In Chapter 3, Section 3.8, three well-known asymmetrical patterns in English were analysed, namely TH-fronting, velar nasal fronting, and back vowel fronting. The question is whether asymmetrical patterns such as these ones, and in general, can be predicted by the relative frequencies of the segments in the language. For instance, [θ] is being perceived as [f] more often than the reverse, but it is also true than [f] is a more frequent segment than [θ]. If the relative frequencies can affect the asymmetrical confusions, then confusion asymmetries are a function of both perceptual biases and

frequency biases.

4.1.1.3 Frequency measures

Furthermore, three different measures of segmental frequency will be examined for the strength of their effect on the three aspects of segmental confusions mentioned above. The three measures are token frequency (the number of times a given segment is found in the language), type frequency (the number of words that contain a given segment) and weighted type frequency (the number of words that contain a given segment, weighted by the token frequency of the words).

The unweighted type frequency measure is purely lexically-based, while the token frequency measure is not. The weighted type frequency measure is a hybrid measure, which is partially lexically-based. If we find that the type frequency measure generally predicts segmental confusions better than the other two non-lexically-based measures, then we could argue that listeners are sensitive to lexical items in segmental misperception.

4.1.2 Syllable factors

Moving away from segments into syllables, we could examine whether certain factors on the syllable level have a top-down effect on segmental misperception. Three factors are tested – syllable constituency, syllable position and stress.

Syllable constituency is the position of the segment in a syllable, namely onset, nucleus and coda. Syllable position is the position of the syllable that contains the segment in a polysyllabic word. Three positions can be generalised, namely word initial, word medial, and word final. Stress is whether the syllable is stressed or unstressed.

Focusing on whether these factors have an effect on whether a segment is more likely to be misperceived, the following questions could be asked: Do segmental

errors occur evenly across the three syllable constituents? Do we expect segments in certain syllable positions to be misperceived more often than others? Are segments in unstressed syllables more often misperceived than those in stressed syllables? Finally, do we expect the effect of syllable constituency and stress to be different between monosyllabic and polysyllabic words?

4.1.3 Word frequency

Let us move on to a larger linguistic unit. The relationship between the frequency of the intended word and that of the perceived word will be examined. First, is there a relationship between the frequency of the intended word and that of the perceived word, $Freq._{Perceived} = f(Freq._{Intended})$? Second, is the frequency of the perceived word more frequent than or similar to that of the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$?

This will allow us to find out whether listeners are sensitive to frequency (or its correlates) on the segmental level as well as the word level in misperception. Furthermore, it will shed light on the mechanisms/strategies that listeners use when retrieving lexical items in speech perception.

4.1.4 Self-information

The last top-down factor concerns the amount of self-information a word has and its effect on whether a word is more likely to be misperceived. By self-information, we are referring to Shannon information, which is a function of the average unpredictability in a random variable (Shannon, 1948).

Two kinds of self-information were tested. One is based on the unconditional probability of a word, which is basically the token frequency of a word in a sample of the language. The other is based on the conditional probability of a word, given its previous words. The self-information of a word is the negative log of the probability

of a word; therefore, the more probable a word is, the less self-information it has.

Our question is whether the amount of self-information of a word can be used to predict how likely it is that it will be misperceived in an utterance. If the conditional self-information is shown to be a good predictor after taking into account the unconditional self-information, then it would show that listeners are sensitive, not only to the token frequency of a word, not also to the frequency of a word given its context.

Furthermore, the direction of the effect of self-information on the likelihood of word errors can inform us of possible causes of misperception. On the one hand, it is well-known that high frequency words have a lower processing cost than low frequency words (Brysbaert and New, 2009; New et al., 2007; Keuleers, Brysbaert, and New, 2010; Ernestus and Cutler, 2014). Therefore one possible cause of misperception is that words with high self-information (therefore low frequency/less probable) are more likely to be misperceived because of processing difficulties. On the other hand, words with low self-information (therefore high frequency/more probable) are prone to phonetic reduction (Wright, 1979; Aylett and Turk, 2004; Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001; Coetzee and Kawahara, 2013). Therefore, words with low self-information are more likely to be misperceived because of the amount of phonetic information. Since the two explanations make different predictions, our analyses can reveal which of the two is more plausible.

4.1.5 Summary

This chapter is broken down into five sections. The first four sections contain the analyses of the four top-down factors described previously. First, Section 4.2 will examine the effect of segmental frequency on two different aspects of segmental confusions. Second, Section 4.3 will evaluate the effect of three syllable factors on the likelihood of a segment error. Third, Section 4.4 will examine the frequency

relationship between the intended and perceived words. Fourth, Section 4.5 will evaluate the effect of self-information on the likelihood of a word error in an utterance. Finally, Section 4.6 will conclude the findings and contributions made in this chapter.

4.2 Segmental frequency

This section examines the role of segmental frequencies in segmental confusions. Segmental frequency is the frequency of the occurrence of segments observed in a representative sample of the language. We will focus on two aspects of segmental confusions that could be the result of the segmental frequencies in the language. In addition, three different frequency measures will be tested.

The first and the simplest measure is token frequency, which is the number of occurrences of a given phone (Kučera and Francis, 1967; Nusbaum, Pisoni, and Davis, 1984). The second measure is type frequency, which is the number of lexical items containing a given phone (Kučera and Francis, 1967; Nusbaum, Pisoni, and Davis, 1984). The third measure is like the second measure but weighted by the token frequency of each of the lexical items containing a given phone (Nusbaum, Pisoni, and Davis, 1984). They are summarised below.

- Token: The number of occurrences of a given phone
- Type: The number of lexical items containing a given phone
- Type (Weighted): The sum of the log-transformed frequencies of the lexical items containing a given phone

The role of frequency has played a central role in linguistics. Perhaps the most prominent research in linguistics using frequency was done by Bybee on its role in morpho-phonology and historical analogical changes (Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001). Generally speaking, token frequency refers to the

number of occurrences of a unit in the language, while type frequency refers to the number of occurrences of a specific pattern. To introduce these concepts more clearly, it is worth using examples from morphology. In morphology, the unit of token and type frequencies is a word. Token frequency is the frequency of a word form, e.g. *broke*. Say *broke* occurred 60 times in a corpus of one million words. Type frequency is the frequency of occurrences of a specific pattern. Say that there are three word forms that have the irregular past tense pattern – *broke, spoke* and *wrote*; the type frequency would then be three. However, it is possible that amongst the word forms of the irregular past tense, their token frequencies differ hugely and therefore each contributes a different amount of weight. Rather than saying that each of the three word forms contributes equally, we would weigh each of them by their respective token frequencies. The type frequency of the irregular past tense is therefore the sum of the log-transformed token frequency of the three word forms. This measure is the weighted type frequency.

Let us return to the level of segments. While there is no doubt that these three measures are highly correlated, they do make different predictions about the nature of segmental frequencies in perceptual confusions. If the two type frequency measures were able to capture more variance than the token frequency measure in perceptual confusions, then one could argue that the listeners are sensitive to lexical information (i.e., the segmental frequencies are computed from the words the listeners know, not from a large sample of segments.). Amongst the two type frequency measures, the literature has conflicted views of whether a weighted measure can better reflect our linguistic knowledge. When calculating neighbourhood density, Bailey and Hahn (2001) proposed a metric that weighs the lexical neighbours of a target word by their respective token frequency. That is, some neighbours contribute more than others. They demonstrated that a weighted measure can capture more variance in behavioural data (non-word acceptability ratings). A later study by Albright

(2007) replicated the analyses in Bailey and Hahn (2001), but the author could not find a significant improvement in the amount of variance explained. In fact, a number of previous studies claimed that pattern strength in the lexicon is determined by type and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). In sum, the two type frequency measures are expected to outperform the token frequency measure. Furthermore, the weighted type frequency measure is expected to perform worse than the unweighted type frequency measure. If it were to outperform the unweighted measure, the difference should be negligible.

Having described the three frequency measures and their respective predicted performance, we will now briefly describe the two aspects of segmental confusions that are being examined for the existence of any segmental frequency bias. A more detailed description of each of the two aspects can be found in their respective introduction sections.

The first aspect concerns whether frequency can capture the target and response biases. Are certain phones more (or less) likely to be spoken but misperceived (the target)? Are certain phones more (or less) likely to be involved in the resultant perceived phones (the response) in a misperception? If so, whether these patterns can be captured by frequency. In other words, does the frequency distribution of the phones in the language have a similar frequency distribution of the phones being the target and the response of a misperception? From the perspective of a listener, given a segment will be misperceived by the listener, this segment is more likely to be a segment that is frequently spoken in the language, than a segment that is less frequently spoken. Similarly, the responses that the listener gives as the perceived segments (though incorrectly) are biased by the frequency of the segments in the language; that is, the listener would perceive a specific segment more often because this segment is frequent in the language.

The second aspect concerns the asymmetrical confusions. Perceptual confusions are often asymmetrical, i.e. a segment x is perceived as a segment y more often than reverse. We test whether the direction and strength of the asymmetrical confusions across all pairs of segments can be explained by the relative frequency of the two segments in the language. Say that there is an asymmetry between [f] and [θ] in the direction of [θ] > [f]. It is possible that this is due to the fact that [f] is more frequent than [θ] in the language; therefore, listeners are biased to perceive [θ] as [f] more often than the reverse.

In sum, Section 4.2.1 outlines the data that are examined. Section 4.2.2 examines whether frequency can capture the target and response biases. Section 4.2.3 examines whether the strength and direction of the asymmetrical confusions can be predicted by the relative segmental frequencies. Each of the latter two sections contains its own introduction, method, analysis and conclusion sections. Finally, Section 4.2.4 concludes the findings of these two sets of the analyses.

4.2.1 Data extraction

The naturalistic data used in this section are the context-free segmental confusions, as described in Chapter 3, Section 3.2.

Given the three frequency measures, three sets (token, type and weighted type) of actual segmental frequencies were extracted from a control written English corpus as described in Chapter 2, Section 2.3. First, a frequency list was compiled from the corpus. Second, in order to remove words that were erroneously introduced into the corpus due to typos, words that occurred in fewer than three pieces of subtitle texts (i.e. three episodes/films) were removed.[1] Finally, given that we transcribed tapping across word boundaries, words ending in /t/ or /d/ followed by a vowel could have more than one pronunciation (e.g. *it* has two pronunciations: [ɪt] and [ɪɾ]). The

[1] I thank Dr. Emmanuel Keuleers for suggesting this filter.

type frequency measure was computed over the dominant pronunciation, whereas the weighted type frequency measure was computed over the dominant pronunciation but weighted with the combined token frequency of all the pronunciations. The token frequency measure was computed over all the pronunciations and their corresponding token frequencies.

26 consonants were considered – [p, t, k, b, d, g, ʃ, ʒ, tʃ, dʒ, θ, ð, s, z, f, v, h, m, n, ŋ, ɹ, l, p^h, t^h, k^h, ɾ]. The reason for excluding the glides [j, w] is because in the corpus transcription, they are used as both consonants in onset positions and as offglides of the vowels; therefore, to avoid ambiguity, they were excluded from the consonant set.

14 vowels were considered – [i, ɪ, e, ɛ, æ, a, ɑ, ɔ, o, u, ɜ, ʌ, ʊ, ə], excluding [ʉ] and [ɒ]. The reason for excluding these two vowels is to focus on the General American accent, since the written corpus from which I extract the frequency norms was transcribed with a General American accent, so the frequency norms cannot be found for these two vowels.

4.2.2 Target and response biases

The first aspect of segmental confusions concerns the target and response biases in misperception. Concretely, the questions are whether certain phones are more (or less) likely to be spoken but misperceived (the target), or involved in the resultant perceived phones (the response) in a misperception, and if so, whether this pattern can be captured by frequency. Say that [t] is the most frequently spoken but misperceived segment (i.e. this segment is an intended segment, but it was perceived as something else). Similarly, say that [t] is the most frequently perceived segment for a misperceived intended segment (i.e. a given segment was misperceived as this segment). An obvious explanation would be their high frequencies are simply because [t] is one of the most frequent segments in the language.

It is important to understand that the target bias means that certain phones are more likely to surface as the target of a misperception, and it does *not* mean that certain phones are more likely to undergo misperception. To avoid ambiguity, in this section, we will refer to the segmental frequencies in the language as the *Actual* (segmental) frequencies, the frequencies of an intended segment in a segmental misperception as the *Target* (segmental) frequencies, and the frequencies of a perceived segment in a segmental misperception as the *Response* (segmental) frequencies.

If the actual frequencies correlate with the target frequencies, then it would suggest (given that there is a perceptual error) that the probability of a certain phone being the intended segment of this error is a function of the probability of this phone in the language. In other words, the more frequently an intended segment is produced, the more likely it will be the target of a misperception. This is to say, there is a target bias due to frequency.

Similarly if actual frequencies correlate with response frequencies, then it would suggest that the probability of a given segment being chosen (incorrectly) as the perceived segment is determined by how frequent the perceived segment is in the language. Given an intended segment will be misperceived, the listener will choose a segment as the response based on how frequent it is. This is to say, there is a response bias due to frequency.

In addition to the question of whether there is a target bias and a response bias due to actual frequency, the next question is how much of the variance of these biases can be captured with frequency. The findings from this is crucial, because the variance that cannot be explained by frequency is therefore potentially captured with other non-frequency factors. In terms of the target bias, one non-frequency account is a phonetic account which predicts that a phone that is phonetically less robust (the amount/strength of the phonetic cues) is more likely to be a target of misperception. In terms of the response bias, a non-frequency account is also a phonetic account

which predicts that the choice of the (incorrectly) perceived segment is dependent on its perceived similarity to the intended segment. In sum, this analysis can indirectly highlight the strength of non-frequency factors that are involved in the target and response biases, and the most obvious factor is the phonetic properties of the phones in terms of robustness and their mutual perceived similarity.

These questions were previously examined for naturalistic misperception by Bird (1998) (using 300 instances of naturalistic misperception, which is a sub-corpus of our mega corpus).

Focusing on substitutions, Bird (1998) conducted a correlation analysis separately for consonants and vowels. The author correlated the actual frequencies with the target frequencies, and with the response frequencies. While all the correlations were statistically significant, the correlation values with the vowels were higher (R = 0.89 – 0.93) than those with the consonants (R = 0.80 – 0.84). These high and significant correlations suggest that an extremely high proportion of the variance is explained by the actual frequencies alone. The fact that the correlations were not perfect (i.e. R was not 1) suggests other factors are at work (though playing a very minor role), causing certain phones to be involved in misperceptions more often or less often that the actual frequencies would predict.

Bird's (1998) frequency analyses opened up a range of questions. Firstly, given Bird's (1998) data were based on 300 instances, which is a relatively small sample compared to our mega corpus (around 5,000 instances), can the correlation results be replicated? Secondly, the author focused on substitution. Can we expect to find similar correlations with insertion and deletion?

To conclude, a number of questions can be raised regarding this aspect of segmental confusions. The key question is whether there is a frequency bias for a phone being the target segment and the response segment of a misperception. The second question is, given there is a bias, how strong is this bias? How much of the vari-

ance can be explained with the segmental frequencies alone? The third question is which of three frequency measures (token, type and weighted type) can capture the most variance. The fourth question is whether the findings of the previous questions would differ between consonants and vowels, and between substitutions, insertions and deletions.

4.2.2.1 Method

Given we are interested only in the segments involved in misperceptions, the correctly perceived segments were ignored. That is, the diagonal cells of the confusion matrix were ignored.

A non-parametric correlation, Spearman, was used to compare the two sets of frequencies, since the frequency values are not normally distributed.

4.2.2.2 Analyses

4.2.2.2.1 Consonants The correlation results of the consonants are summarised in Table 4.1. The table contains the correlation values with the level of statistical significance indicated by the number of asterisks. The table categorises the correlation values by the target frequency (substitution and deletion) and response frequency (substitution and insertion) across the table horizontally, as well as by the three frequency measures vertically. The correlation value in bold in each column is the best correlation amongst the three frequency measures.

From the table, we see that all the correlation values are statistically significant and at a strong to very strong level. The lowest value is 0.7820, and the highest value is 0.9670. This clearly indicates that the actual segmental frequency is a strong factor for the target bias and response bias of misperception for consonants.

For both target and response frequencies (substitution, insertion and deletion), type frequency yielded better correlation than the weighted type frequency. This was

	Target		Response	
Frequency Measure	Substitution	Deletion	Substitution	Insertion
Token	0.8417^{***}	$\mathbf{0.9393}^{***}$	0.8273^{***}	$\mathbf{0.9670}^{***}$
Type	$\mathbf{0.9183}^{***}$	0.7824^{***}	$\mathbf{0.9008}^{***}$	0.7936^{***}
Type (Weighted)	0.9042^{***}	0.7820^{***}	0.8943^{***}	0.7841^{***}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.1: Segmental frequency correlations (Spearman, two-tailed) of consonants between target and response frequencies with actual frequencies of three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

expected, since the weighted type frequency is weighted with the token frequency of the relevant lexical items, and previous studies claimed that pattern strength in the lexicon is determined by type frequency and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). However, it is not always the case that both the type frequency measures (weighted and unweighted) yield better correlations than the token frequency measure. This is only the case for substitution (target and response), while for insertion (response) and deletion (target) token frequency outperforms both measures of type frequency. One possible explanation is that for substitution two lexical items must be involved in the misperception, while for insertion and deletion it is possible (but not necessary) that only one lexical item is involved (i.e. a whole word insertion and a whole word deletion). In this way, insertion and deletion are less sensitive to lexical information than substitution, and yield poorer correlations with the two type frequency measures which are lexically based.

All the correlations are visualised as scatterplots fitted with a linear regression line with confidence intervals. They are Figures 4.1, 4.2, 4.3 and 4.4. Overall, the relative strength of the correlation values is well reflected in the plots, particularly with insertion and deletion.

Although the correlation values are strong, they are not perfect, just as Bird's

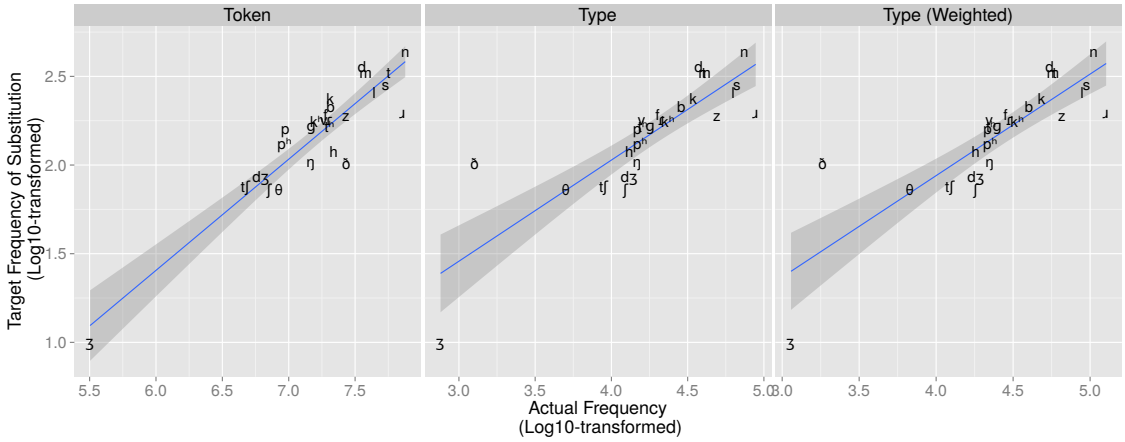


Figure 4.1: The relationship between the target frequencies of substitution and three measures of actual segmental frequencies: consonants

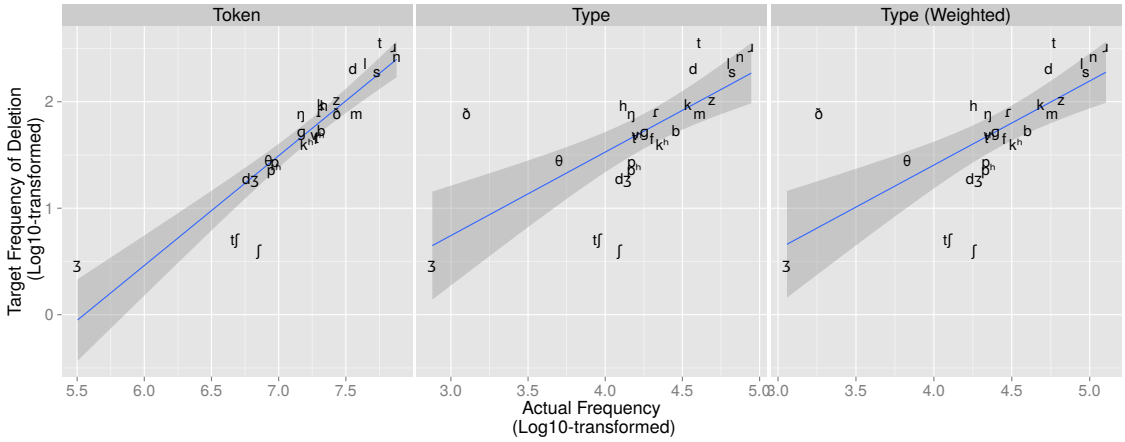


Figure 4.2: The relationship between the target frequencies of deletion and three measures of actual segmental frequencies: consonants

(1998) findings. A closer look at the segments that cannot be explained by the best actual frequency measure (the measure that yields the highest correlation value) could potentially reveal any non-frequency factors. The segments that fall outside the confidence intervals of the linear regression lines (in Figures 4.1, 4.2, 4.3 and 4.4) are summarised in Table 4.2.

First of all, it is worth noting that the diverged segments in the target and response biases of substitution can be analysed together. Similarly, the diverged segments in the target bias of deletion and the response bias of insertion can be analysed

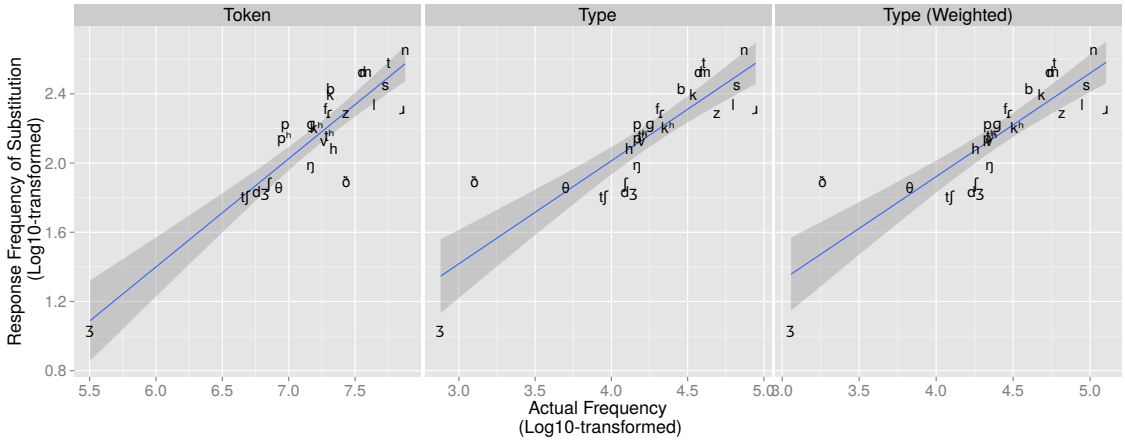


Figure 4.3: The relationship between the response frequencies of substitution and three measures of actual segmental frequencies: consonants

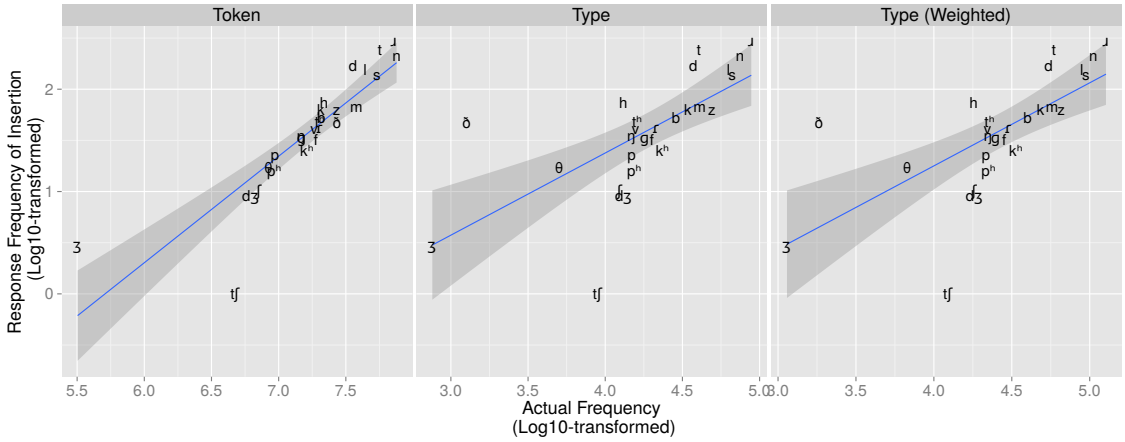


Figure 4.4: The relationship between the response frequencies of insertion and three measures of actual segmental frequencies: consonants

	Target		Response	
	Substitution	Deletion	Substitution	Insertion
More often	[t, d, n, m, ð]	[t, d, ʒ, l, ŋ]	[t, d, n, m, ð, p, b, f]	[t, d, ʒ, l, ɹ, ŋ]
Less often	[z, dʒ, tʃ, ʃ, ʒ, ɹ, ŋ]	[tʃ, ʃ, m]	[z, dʒ, tʃ, ʃ, ʒ, ɹ, l, ŋ]	[dʒ, tʃ, ʃ]

Table 4.2: Consonant segments diverged from actual frequency: the row "More often" denotes the segments that are the target/response of a misperception more often than expected by the best actual frequency measure; the row "Less often" denotes the segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

together. This is because the response patterns can be the result of hypercorrection (Ohala and Shriberg, 1990).

In Table 4.2, the row "More often" contains segments that are the target/response of a misperception more often than expected by the best actual frequency measure (i.e. these segments are above the linear regression). The row "Less often" contains segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

The types of diverged segments are examined in the following order.

1. The segments that are the target and response of a substitution more often than expected
2. The segments that the target of a deletion and the response bias of an insertion more often than expected
3. The segments that are the target and response of a substitution less often than expected.
4. The segments that the target of a deletion and the response bias of an insertion less often than expected

Let us start with the consonant segments that are the target/response of a substitution **more** often than expected. The diverged target segments are [t, d, n, m, ð], and the diverged response segments are [t, d, n, m, ð, p, b, f]. First, [t, d] can be explained by the fact that they are perceptually weak segments (stops are the least sonorous manner) and often undergo lenition intervocalically (Kirchner, 2001). A closer look at the raw confusion matrix in Figure 3.7 in Chapter 3 reveals that [t] and [d] are most confusable with each other, with [t] being perceived as [d] 1.85% of the time, and [d] being perceived as [t] 2.9% of the time. This suggests that the [t] and [d] are diverged from the expected actual frequency due to voicing confusion. Voicing confusion can be viewed as the result of lenition and the hypercorrection of lenition, if the voicing confusions occur intervocalically (further analyses are needed to examine the environment of these voicing confusions).

Second, [n] can be explained by the fact that it occurs **more** often in unstressed environments, e.g. in the word "and" and in prefixes such as "un" and "in". Unstressed environments should be more susceptible to misperception than stressed environments because stressed environments are perceptually more prominent (longer duration, higher intensity). Furthermore, "and", "un" and "in" are highly frequent in the lexicon and they therefore are more likely to have a shorter duration and undergo processes of phonetic reduction (Wright, 1979). Regarding the divergence of [m], a closer look at the raw confusion matrix in Figure 3.7 in Chapter 3 reveals [n] and [m] are most confusable with each other, with [n] being perceived as [m] 2.96% of the time, and [m] being perceived as [n] 5.7% of the time. Given the high confusion between [n] and [m], the divergence of [m] can also be explained. This divergence of [n, m] was also found by Bird (1998).

Third, [ð] can be explained by the fact that it is mainly found in high frequency function words such as "the", "that", "this" "their" etc. Since, high frequency words tend to be phonetically weakened (Wright, 1979), [ð] is misperceived more often than expected by its actual frequency.

Finally, [p, b, f] are the response of a misperception more often than expected by their actual frequencies, for which I have no immediate explanation.

Let us move on to with the consonant segments that are the target of deletion and the response of insertion **more** often than expected. The diverged target segments are [t, d, ʒ, l, ŋ], and the diverged response segments are [t, d, ʒ, l, ɹ, ŋ]. First, [t, d] can be explained by the fact they are often deleted, especially in word-final positions of mono-morphemic words (Guy, 1991; Coetzee and Kawahara, 2013). Second, [l, ɹ] can be explained by the fact that they are often the second/third consonant of a onset cluster (e.g. [pl], [fɹ], [spɹ] etc.). It is well known that these positions are prone to deletion (Harris, 1994), as predicted by their sonority slopes (the deletion should result in a maximal sonority rise) (Ohala, 1999). Finally, I have no explanation for

why [ʒ, ŋ] are inserted or deleted more often than expected.

To briefly conclude, the diverged consonant segments that are the target/response of a substitution/insertion/deletion more often than expected by their actual frequencies can be accounted for using the fact that their phonetic properties are particularly susceptible to misperception.

Next, the consonant segments that are the target/response of a substitution *less* often than expected are examined. The diverged target segments are [z, dʒ, tʃ, ʃ, ʒ, ɹ, ŋ], and the diverged response segments are [z, dʒ, tʃ, ʃ, ʒ, ɹ, l, ŋ]. All the fricatives and affricates [z, dʒ, tʃ, ʃ, ʒ] can be explained by the fact that they are perceptually robust and their acoustic cues lie within the consonants themselves; that is, they are relatively independent of their environment (Wright, 2004). This pattern with the fricatives is consistent with Bird's (1998) findings that the fricatives [s, z] are the target/response of a substitution less often than expected by their actual frequency. Two of the remaining diverged segments are [ɹ, l]. One explanation is that they are liquids which have high acoustic energy, and are high on the sonority scale; therefore, they are particularly salient, and less prone to errors. The last diverged segment is [ŋ] for which I have no explanation since the other nasals have the reverse pattern, [m, n] are the target/response of a substitution *more* often than expected.

Let us move on to with the consonant segments that are the target of deletion and the response of insertion *less* often than expected. The diverged target segments are [tʃ, ʃ, m], and the diverged response segment are [dʒ, tʃ, ʃ]. Similar to the diverged segments with the substitutions, all the fricatives and affricates [dʒ, tʃ, ʃ] can be explained with the fact that they are perceptually robust (Wright, 2004). The remaining diverged segment is [m], for which I have no explanation.

To conclude, most of the diverged consonant segments that are the target/response of a substitution/insertion/deletion more/less often than expected can be explained phonetically. Those that are the target/response more often than expected are pho-

netically weak, while those that are the target/response less often than expected are phonetically strong. Therefore, the target/response patterns that cannot be captured with a frequency account can be captured with a phonetic account.

4.2.2.2.2 Vowels Let us move on to the vowels. The correlation results are summarised in Table 4.3, with the same format as Table 4.1.

	Target		Response	
Frequency Measure	Substitution	Deletion	Substitution	Insertion
Token	**0.8637*****	**0.8185*****	0.8471***	**0.6960****
Type	0.8593***	0.6336*	**0.8845*****	0.5352*
Type (Weighted)	0.8330***	0.6029*	0.8691***	0.5264^{+}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.3: Segmental frequency correlations (Spearman, two-tailed) of vowels between target and response frequencies with actual frequencies of three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

The overall patterns in Table 4.3 are essentially the same as those with the consonants. All but one of the correlations reached statistical significance ($\alpha = 0.05$). The exception is the correlation between the weighted type frequency and insertion.

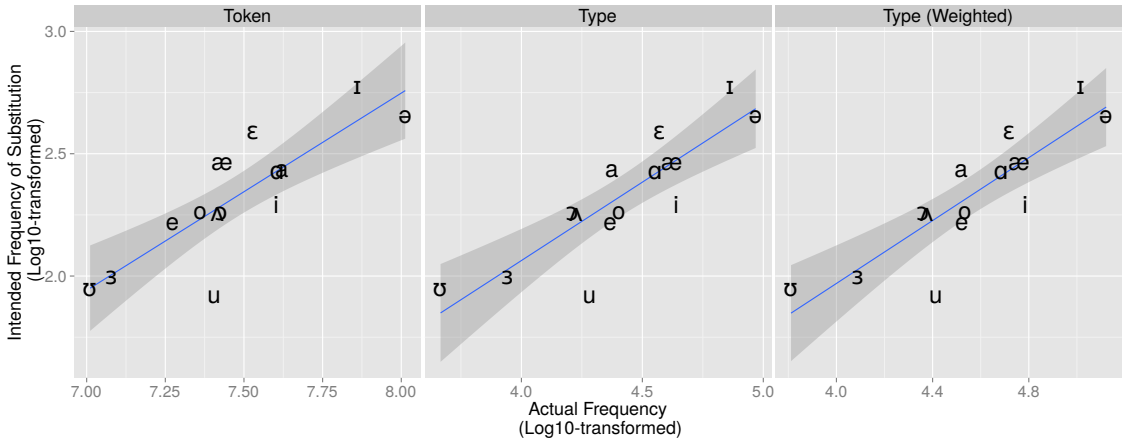


Figure 4.5: The relationship between the target frequencies of substitution and the three measures of actual segmental frequencies: vowels

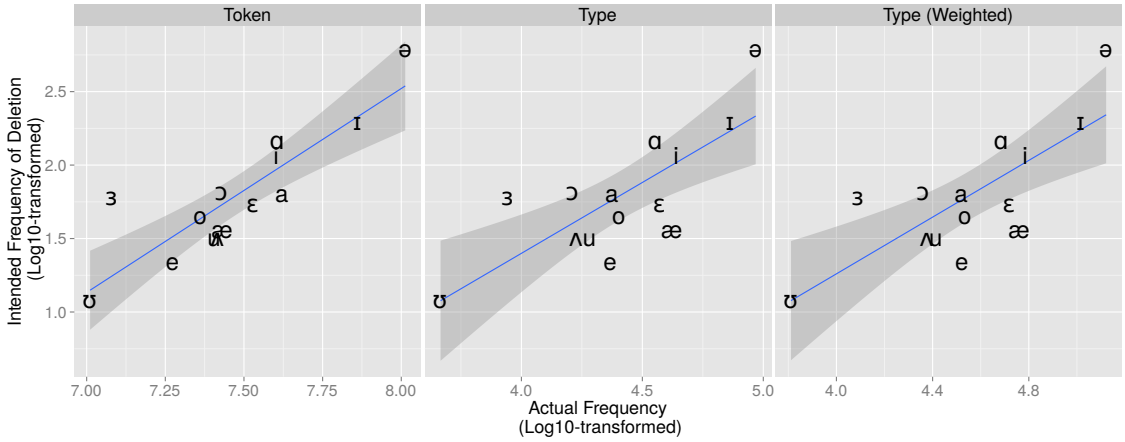


Figure 4.6: The relationship between the target frequencies of deletion and three measures of actual segmental frequencies: vowels

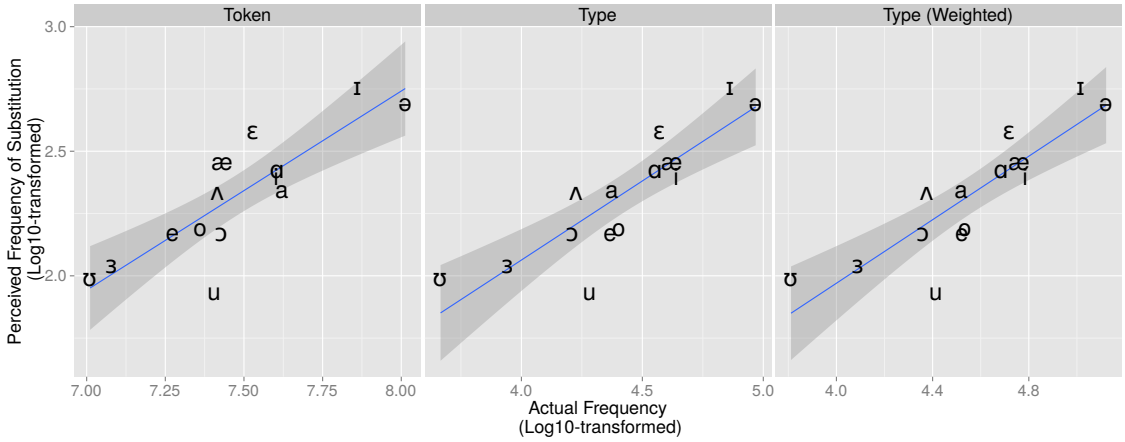


Figure 4.7: The relationship between the response frequencies of substitution and three measures of actual segmental frequencies: vowels

Again, all the correlations are visualised as scatterplots, each fitted with a linear regression line with confidence intervals. They are Figures 4.5, 4.6, 4.7 and 4.8. Overall, the relative strength of the correlation values is well reflected in the plots, particularly with insertion and deletion. Visually, all of the correlations do not appear to be skewed by extreme outliers.

Given the correlation values are not perfect, it is worth examining the diverged segments. The segments that fall outside the confidence intervals of the linear regression lines (in Figures 4.5, 4.6, 4.7 and 4.8) are summarised in Table 4.4. In Table

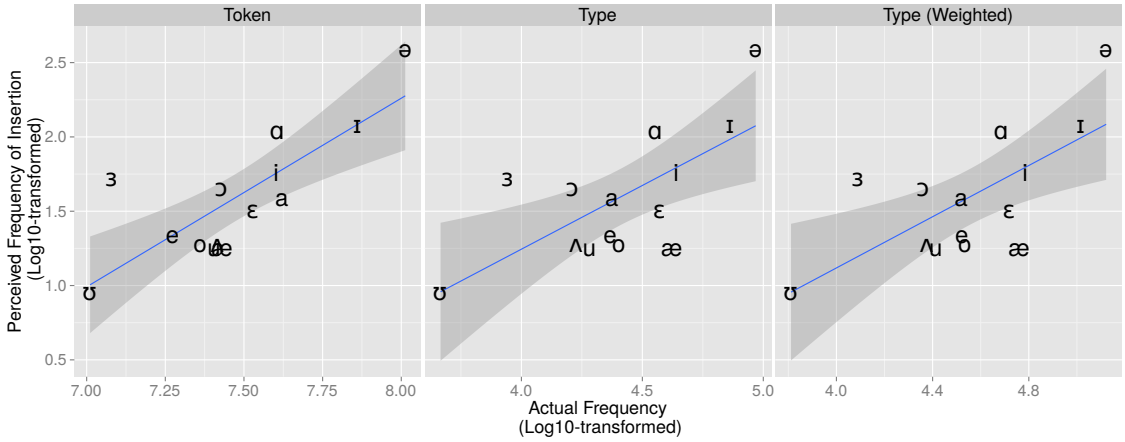


Figure 4.8: The relationship between the response frequencies of insertion and three measures of actual segmental frequencies: vowels

4.4, the row "More often" contains segments that are the target/response of a misperception more often than expected by the best actual frequency measure (i.e. these segments are above the linear regression). The row "Less often" contains segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

	Target		Response	
	Substitution	Deletion	Substitution	Insertion
More often	[æ, ɛ]	[ɜ, ɑ]	[ɪ, ɛ, ʌ]	[ɜ, ɑ]
Less often	[u, i]	[u, a, ʌ, æ]	[u, e, o]	[u, o, ʌ, æ]

Table 4.4: Vowel segments diverged from actual frequency: the row "More often" denotes the segments that are the target/response of a misperception more often than expected by the best actual frequency measure; the row "Less often" denotes the segments that are the target/response of a misperception less often than expected by the best actual frequency measure.

Let us start with the vowel segments that are the target/response of a substitution **more** often than expected. The diverged target segments are [æ, ɛ], and the diverged response segments are [ɪ, ɛ, ʌ]. All the diverged segments are in fact lax vowels, which are phonetically short, with a lower intensity than tense vowels. Therefore, these lax vowels are the target/response of a substitution more often than expected because

they are phonetically weak.

Interestingly, the vowel segments that are the target/response of a substitution **less** often than expected are all tense vowels. The diverged target segments are [u, i], and the diverged response segments are [u, e, o]. Using the same argument as the lax vowels, the divergence of these tense vowels can be explained; tense vowels are longer with a higher intensity than lax vowels. Therefore, they are the target/response of a substitution less often than expected, because they are phonetically strong. Furthermore, these tense vowels in most of the vowel sets transcribed in the naturalistic corpus are followed by an offglide [j] or [w], which makes them more distinctive and less prone to misperception.

The vowel segments that are the target of deletion and the response of insertion more often than expected are [ɜ, ɑ]. The divergence of [ɜ] can be explained with its high confusion with [ə] and [ɑ]. [ɜ] is mostly often perceived as [ɑ] 2.45% of the time, followed by [ə] 2.31% of the time (see Figure 3.8 in Chapter 3). [ɜ] being confused as [ə] is perhaps due to their close acoustic distance. However, I have no explanation for why [ɜ] is perceived most often as [ɑ].

The vowel segments that are the target of deletion and the response of insertion less often than expected are [u, o, a, ʌ, æ]. I have no explanation for these diverged segments, since they are a mixture of tense and lax vowels, front and back vowels, and close and open vowels. To conclude, just as the diverged consonant segments, most of the diverged vowel segments can be explained using a phonetic account.

Finally, the results of the consonants and the vowels are compared. For the vowels, the correlation value ranges from 0.5264 to 0.8845, while tor the consonants, the correlation value ranges from 0.7820 to 0.9670. While the correlation values of the vowels were all relatively high, they are lower than those of the consonants. This indicates that the amount of frequency bias is higher for the consonants than for the vowels. The opposite is true with the phonetic bias, as found in Section 3.5 and

Section 3.6 in Chapter 3. The amount of phonetic bias was stronger for the vowels than for the consonants. Furthermore, the analysis of the diverged consonants and vowels suggests that what cannot be captured with a frequency bias can be captured with a phonetic bias. Together, one could speculate that the amount of frequency bias complements the amount of phonetic bias in segmental confusions; that is to say, they have an inverse relationship.

Regarding the three frequency measures, again unweighted type frequency outperforms weighted type frequency. The two measures of type frequency outperformed token frequency for substitution, but only for the response frequencies. Just as with consonants, we found that token frequency outperforms both measures of type frequency for insertion and deletion. The earlier discussion of the three frequency measures for the consonants also applies to the vowels.

4.2.2.3 Conclusion

This section examined whether the target and response frequencies in segmental misperception can be explained using the actual frequencies in the language. This question was examined for substitution (target and response), insertion (response) and deletion (target) errors of consonants and vowels.

Section 4.2.2.2.1 examined the substitution, insertion and deletion errors of consonants. The strength of the correlations was at a strong to very strong level (ρ = 0.7820 – 0.9670). Section 4.2.2.2.2 examined the substitution, insertion and deletion errors of vowels. Again, the strength of the correlations was strong (ρ = 0.5264 – 0.8845). These strong correlations indicate that the actual segmental frequencies in the language are a strong factor for the probability of a certain phone being a target or a response of a misperception. To recap, we are *not* referring to the probability of a phone being misperceived, and we are referring to the probability of a phone being a target or a response of a misperception, given there is a misperception.

Concretely, the segment [n] is a more frequent segment than [ʒ]. Given a phone x will be misperceived, this phone x is more likely to be [n] rather than [ʒ]; therefore, the target is biased by the actual frequency. Similarly, given a phone x will be misperceived, the perceived phone y is more likely to be [n] rather than [ʒ].

Most of the segments that diverged from their actual frequencies can be explained using a phonetic account. With the consonants, the segments that were the target/response of a substitution/insertion/deletion more often than expected by their actual frequencies were 1) phonetically weak – [t, d], 2) susceptible to cluster reduction – [l, ɹ]) and 3) susceptible to phonetic weakening due to lexical frequencies – [n] in "and", and [ð] in "the". Similarly, the consonant segments that were the target/response of a substitution/insertion/deletion less often than expected by their actual frequencies were mostly fricatives and affricates which are phonetically strong. With the vowels, there was a clear tense-lax difference with substitutions. Lax vowels were the target/response of a substitution more often than expected by their actual frequencies, because lax vowels are phonetically weak. Tense vowels were the target/response of a substitution less often than expected by their actual frequencies, because tense vowels are phonetically strong.

Furthermore, we found that the correlation values are lower for vowels than for consonants, which indicates that consonants are more sensitive to this frequency bias than vowels. Given that the opposite is true with phonetic bias, as found in Chapter 3 that the diverged segments can mostly be explained using a phonetic account, I speculated that the amount of frequency bias has an inverse relationship with the amount of phonetic bias in segmental confusions. Further analyses are needed to substantiate this speculation by regressing (e.g. with a regression model) the confusion patterns with *both* the frequency bias and the phonetic bias, because it is possible that the phonetic bias also correlates with the frequency bias.

Again, it was found that two measures of type frequency outperformed token

frequency for substitution. This advantage of type frequency will also be found in Section 4.4.2. Surprisingly, token frequency outperformed both measures of type frequency for insertion and deletion. Given that type frequencies are lexically-based, one explanation is that this difference between substitution and insertion/deletion is due to the fact that insertion and deletion are less sensitive to lexical information than substitution, because substitution errors have to involve two lexical items, while insertion and deletion errors could involve only one (i.e. whole word deletions/insertions).

4.2.3 Asymmetrical confusion

The second aspect of segmental confusions concerns their asymmetrical patterns. Recall that three asymmetrical patterns (namely TH-fronting, velar nasal fronting, and back vowel fronting) in naturalistic and experimental misperception were analysed in Chapter 3, Section 3.8. Indeed, all three patterns were confirmed, with [θ] being perceived as [f], [ŋ] as [n] and back vowels as front vowels, more often than the reverse. Finally, we used them as evidence for a perceptual-based account of sound change. However, it is possible that their asymmetries are affected by their relative segmental frequencies. For instance, say that [f] is more frequent in the language than [θ]; [f] could then be chosen as the perceived segment for the intended segment [θ] more often than the reverse, because there is a response bias due to frequency differences. This asymmetrical pattern of [θ] > [f] can therefore be explained without the need of invoking accounts of perceptual biases.

It is worth noting that this bias is similar to the response bias mentioned earlier in Section 4.2.2. Nonetheless, they differ in terms of whether the correctly perceived segments are considered. The bias in Section 4.2.2 concerns only the segments that are involved in a segmental misperception and not the correctly perceived segments, while the current bias concerns both because asymmetricality depends on the propor-

tions of correctly and incorrectly perceived segments (see Chapter 3, Section 3.8.1 for the method for calculating asymmetries).

Benkí (2003) conducted an analysis of whether the asymmetrical patterns in segmental confusions can be captured under a frequency/lexical account. Using experimentally induced misperception of nonsense CVC syllables, the author computed the strength and direction of the asymmetries using the criterion measure (henceforth c bias) from choice theory of eleven pairs of segments. These eleven pairs of segments were three onset pairs – [t, p], [k, p] and [ɹ, l], three vowel pairs – [æ, ɑː], [u, i] and [o, e], and five coda pairs – [t, p], [k, p], [k, t], [g, d] and [m, n]. The relative frequency measures were computed by subtracting the frequency of one of the two segments in a given pair from the frequency of the other segment in the same pair. Four different frequency measures were tested separately. They are a) the number of occurrences per 100 phonemes, b) the number of lexical items containing the phoneme, c) the number of occurrences per million words, and d) the sum of the log-transformed frequencies of the lexical items containing the phoneme. In fact, c) is virtually the same as our token frequency measure, and b) and d) are the same as our two measures of type frequency. The author found that on the whole all of the frequency measures captured a sizable portion of the variance (R^2 from 0.2 to 0.3) of the c bias values; however, none of them were statistically significant at $\alpha = 0.05$. Of the four frequency measures, the number of lexical items containing the phoneme (type frequency) captured most variance, $R^2 = 0.290$, and with the smallest p-value, $p = 0.088$, which is near-significant.

Benkí's (2003) findings are encouraging. The high level of variance explained across multiple relative frequencies indicates that the relative frequency of the two segments can predict the strength and direction of their confusions. Although the p-values did not reach significance, it is likely that this is due to the small number of pairs tested (11 pairs). Therefore, by testing more segmental pairs, we could then

have a more complete picture of whether the relative frequency is a useful factor for predicting asymmetries. In the current analyses, all segmental pairs are tested separately for consonants and vowels. Furthermore, just as Benkí (2003), multiple frequency measures are examined.

To conclude, a number of questions can be raised. Firstly, can the strength and direction of the asymmetrical pattern for each pair of phones be captured by the relative segmental frequencies in the language? Secondly, how much variance can be captured? Thirdly, which of the three frequency measures can capture the most variance? Finally, would the findings of the previous questions differ between consonants and vowels?

4.2.3.1 Method

Regarding the consonant pairs, with 26 consonants, 325 consonant pairs are possible. For the vowel pairs, with 14 vowels, 91 vowel pairs are possible. The strength and direction of the asymmetries were estimated using the criterion measure (c bias) as described in Chapter 3, Section 3.8.1. The c bias values were computed for all 325 consonant pairs and all 91 vowel pairs. Just as in Chapter 3, Section 3.8.1, we excluded pairs that have no confusion in either direction, because their resultant c bias values are dependent purely on the smoothing process. The order of the two segments in a given pair can affect the sign of the c bias; therefore, it is worth establishing a notation system for later reference. For a given pair [Segment 1 > Segment 2], a positive c bias value means that Segment 1 is perceived as Segment 2 more often than the reverse, a negative c bias value means the Segment 2 is perceived as Segment 1 more often than the reverse, and a zero c bias value means that there is no asymmetrical confusion. The first segment in a given pair is referred to as Segment 1 and the second segment as Segment 2.

The relative frequency of each segmental pair was calculated by taking a ratio of

the frequency of Segment 2 and the frequency of Segment 1. To remove the skewness of frequency values, the ratios were then log-transformed for all three measures. This is summarised as the following metric: $Log10(Frequency_{Seg2}/Frequency_{Seg1})$. A positive log-ratio means that Segment 2 is more frequent than Segment 1, a negative log-ratio means that Segment 1 is more frequent than Segment 2, and a zero log-ratio value means that the two segments are equally frequent. This log-ratio has a further advantage of having zero as the centre of the scale just as the c bias value. Therefore, if relative frequencies can predict asymmetries, then a positive correlation is expected between the log-ratios and the c bias values.

A non-parametric correlation, Spearman, was used to compare the two sets of frequencies, since the data are not normally distributed; therefore, a non-parametric correlation is more appropriate.

4.2.3.2 Analyses

Table 4.5 summarises the correlation analyses for consonants and vowels between the c bias values (which reflect the confusion asymmetries) and the log-ratios (which reflect the frequency asymmetries). The table shows the correlation values (Spearman, two-tailed) as well the level of statistical significance.

Frequency Measure	Consonants	Vowels
Token	**0.8068*****	**0.8478*****
Type	0.7080***	0.7851***
Type (Weighted)	0.7109***	0.7847***
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$		

Table 4.5: Correlations (Spearman, two-tailed) between confusion asymmetries and frequency asymmetries of consonants and vowels with three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

All the correlation values are highly significant at a strong to very strong level (ρ = 0.71 – 0.85). Token frequency yields higher correlation values than the two

measures of type frequency. The unweighted frequency measure does not outperform the weighted one consistently, only with vowel asymmetries, and not with consonant asymmetries. Finally, we see that the correlations with the vowels are stronger than those with the consonants. These findings are surprising, considering in the previous analyses of segmental frequency in Section 4.2.2 we found the exact opposite patterns – a) consonants are more affected by frequency than vowels and b) the two measures of type frequency outperform token frequency. Regarding how the vowels are more affected by frequency than consonants in terms of confusion asymmetries, I have no immediate explanation. Regarding the sudden advantage of token frequency in predicting confusion asymmetries, one explanation lies in how asymmetries are defined. Recall in Chapter 3, Section 3.8.1, we described the criterion measure (c bias) which is used to reflect the confusion asymmetries. The c bias measure relies on the proportion (not count) of confusions in each direction, and the frequencies of the correctly perceived segments (the diagonal cells in a confusion matrix) are required to compute the proportions. In the naturalistic corpus, the frequency of the correctly perceived segments should highly correlate with their frequency in the language, because the naturalistic corpus is a sample of the language. This is indeed the case, as indicated by the correlation (Spearman, two-tailed) between the frequency of the correctly perceived segments and their frequency in the language. With the consonants, the correlation values are 0.9835, 0.8427 and 0.8345, with token, type and weighted type frequency respectively, and they were all highly significant. With the vowels, the correlation values are 0.9648, 0.8373 and 0.8109, with token, type and weighted type frequency respectively and they were all highly significant. Given that the correctly perceived segments are extracted from all the segments in all words that are correctly perceived in the corpus, these extracted frequencies are, in fact, token frequency, and not type frequency. Therefore, the advantage of token frequency in predicting the confusion asymmetries can be explained.

To examine these relationships further, we will visualise the correlations as scatterplots, fitted with linear regression lines. The scatterplots of the consonants are shown in Figure 4.9 and those of the vowels are shown in Figure 4.10. Focusing on the consonants, Figure 4.9 shows that the regression lines with the two measures of type frequency are poorly fitted, compared to that with token frequency. A closer inspection of the segmental pairs reveals that the poor fits are due to [ð] having a low type frequency, as [ð] is found in all the pairs that are outliers (visually). Let us move on to the vowels. Figure 4.10 shows that the regression line has a tighter fit with token frequency than with the two measures of type frequency. However, unlike the consonants, visually the differences cannot be attributed to specific segments.

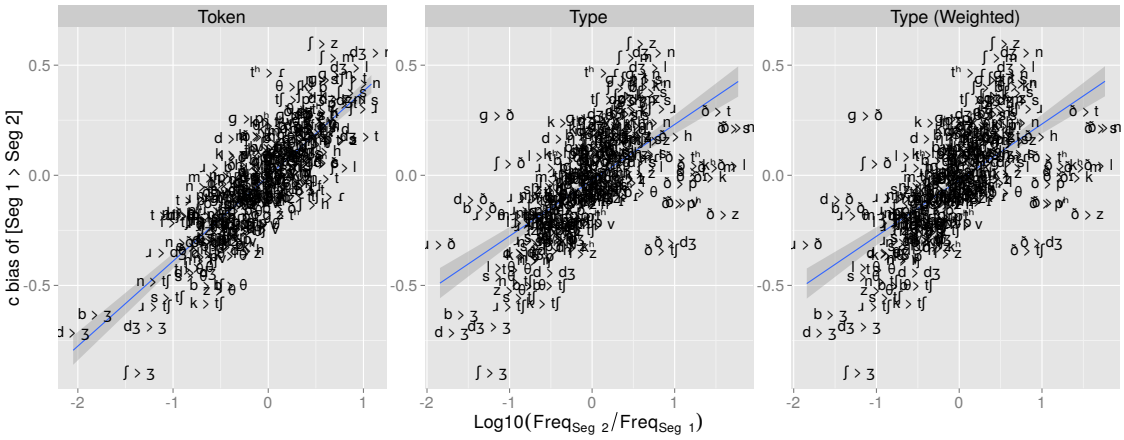


Figure 4.9: The relationship between confusion asymmetries and frequency asymmetries of consonants

Finally, based on our results in this section, we should reconsider our conclusion based on the analyses in Chapter 3, Section 3.8, where we analysed the three asymmetrical patterns, namely TH-fronting, velar nasal fronting, and back vowel fronting. Our results in this section suggest that confusion asymmetries are affected by the relative segmental frequencies found in the language, by examining all the possible asymmetries (i.e. all combinations of two segments). In fact, TH-fronting and velar nasal fronting can both be explained using a frequency account, since [n] is more frequent than [ŋ], and [f] is more frequent than [θ].

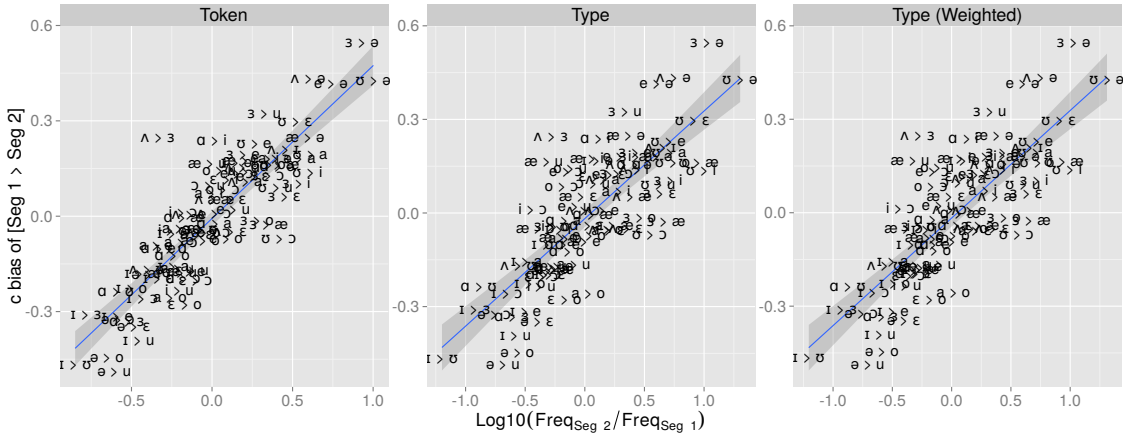
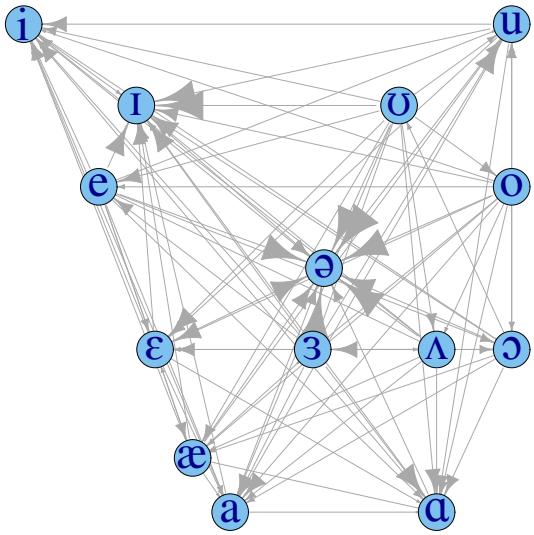


Figure 4.10: The relationship between confusion asymmetries and frequency asymmetries of vowels

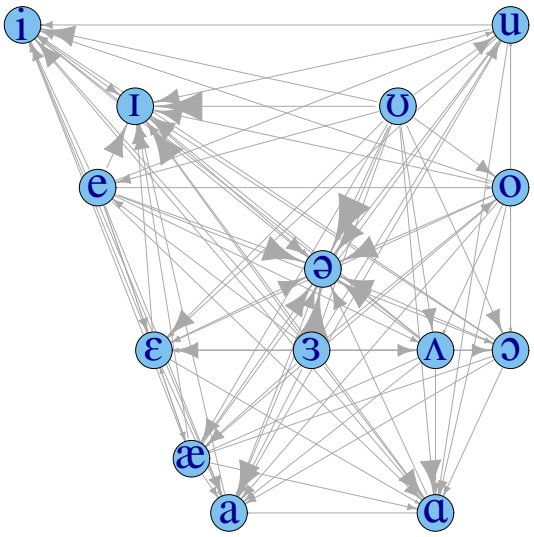
By visualising the direction and strength of the asymmetries with a vowel chart, we can better evaluate the back vowel fronting pattern. The asymmetrical patterns for all the vowel combinations are shown in Figure 4.11. Figure 4.11 contains three sub-figures. Figure 4.11a summarises the confusion asymmetries. Figure 4.11b summarises the token frequency asymmetries. Figure 4.11c summarises the type frequency asymmetries (there are no visual differences between the weighted and unweighted type frequency measures). In each figure, each vowel is connected with all other vowels with a straight line, the direction of the arrow head reflects the direction of the asymmetry, and the size of the arrow head reflects the strength of the asymmetry. Visually, all three figures are extremely similar in terms of the direction of the arrows.

A back vowel fronting pattern can indeed be found in the confusion asymmetries (Figure 4.11a) as all the back vowels (except [ɑ]) have fewer incoming arrows than the front vowels. However, the same pattern can be found also with the token frequency asymmetries and type frequency asymmetries in Figure 4.11b and Figure 4.11c respectively. This suggests that the back vowel fronting pattern is affected by segmental frequencies.

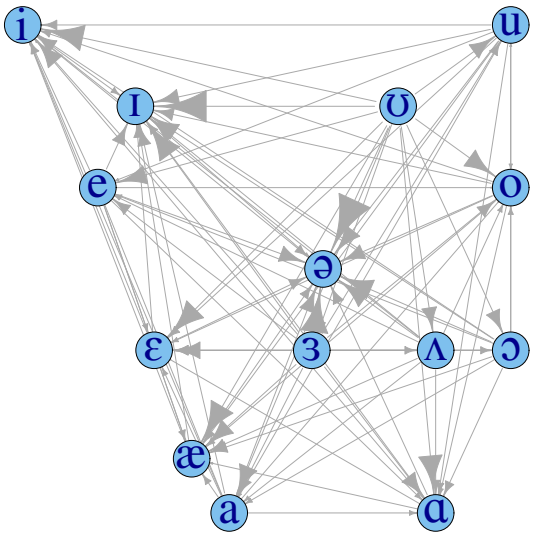
Although unrelated to the fronting pattern, it is worth noting that there is a



(a) Confusion asymmetries



(b) Token frequency asymmetries



(c) Type frequency asymmetries

Figure 4.11: Visualisation of vowel asymmetries: a) confusion asymmetries, b) token frequency asymmetries, c) type frequency asymmetries (both weighted and unweighted).

centring pattern with most vowels moving into [ə] in both confusion asymmetries and frequency asymmetries. This centring pattern in frequency asymmetries could in fact explain an earlier observation in Chapter 3, Section 3.4.3.1, that there is a confusion bias of open/close vowels being perceived as mid vowels, which is essentially

centring.

4.2.3.3 Conclusion

This section examined the effect of frequency on the asymmetrical patterns in segmental confusions. In Chapter 3, Section 3.8, we analysed three asymmetrical patterns, namely TH-fronting, velar nasal fronting, and back vowel fronting, in naturalistic and experimental misperception. As an extension to the previous analyses, we correlated the frequency asymmetries (the difference in frequency between two segments) and the confusion asymmetries of all possible pairs of segments for both vowels and consonants. The correlations were highly significant at a strong to very strong level ($\rho = 0.71 - 0.85$). This is a surprising finding, suggesting that confusion asymmetries can be affected by frequency asymmetries.

Indeed, the three asymmetrical patterns, TH-fronting, velar nasal fronting, and back vowel fronting, were explicable using frequency, since [n] is more frequent than [ŋ], and [f] is more frequent than [θ], and front vowels are generally more frequent than back vowels. These results suggest that we should reconsider our conclusion in Chapter 3, Section 3.8, such that confusion symmetries are a function of not only perceptual biases (Ohala, 1981; Ohala, 1989), but also frequency biases.

Furthermore, we found that vowels are more affected by frequency than consonants and that token frequency outperformed type frequency. These are the exact opposite patterns from those found in Section 4.2.2. I have no explanation for why vowels are more affected by frequency than consonants. However, regarding the advantage of token frequency, I proposed that it is due to how confusion asymmetries are defined. The criterion measure (c bias) which is used to reflect the confusion asymmetries relies on the frequencies of the correctly perceived segments (the diagonal cells in a confusion matrix). The correctly perceived segments are extracted from all the segments in all words that are correctly perceived in the corpus. Therefore,

they are token frequency, and not type frequency. The advantage of token frequency should not be interpreted as a linguistic advantage, but as a methodological bias.

4.2.4 Conclusion

This section conducted separate sets of analyses to examine the role of segmental frequencies in segmental confusions. We focused on two aspects of segmental confusions that could be the result of the segmental frequencies in the language, namely a) the effect of frequency on target and response biases, and b) the relationship between frequency asymmetries and confusion asymmetries.

Section 4.2.2 found that the target and response biases can be captured nearly perfectly with segmental frequencies as indicated by the strong to very strong level of correlations. This pattern is true for substitutions, insertions and deletions. We found that the effect of frequency is stronger for consonants than for vowels.

Section 4.2.3 found that confusion asymmetries can be affected by frequency asymmetries, as indicated by the strong level of correlations. The theoretical implication of this is that confusion asymmetries are modulated by top-down factors such as segmental frequencies. In other words, the result suggests that confusion symmetries is a function of both perceptual biases (Ohala, 1981; Ohala, 1989), and frequency biases. Further work is needed to examine whether frequency asymmetries play a role in experimental misperception such as Miller and Nicely (1955).

In terms of the comparison between the different frequency measures, overall we found type frequency outperformed token frequency in substitution (but not insertion and deletion), which supports the claims that pattern strength in the lexicon is determined by type, and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004).

Together all these findings allow us to make the following conclusions. Given that a segment will be misperceived, the target and the response are determined

by the actual frequencies; that is, more frequent segments are more likely to be misheard (target) and incorrectly chosen as the perceived segment (response). This holds for substitution, insertion and deletion errors. Crucially this frequency bias operates independently on the intended segments and the perceived segments. Since frequency can partly determine the intended and perceived segment, it can also bias the overall confusion patterns, namely the asymmetricality of confusions.

To conclude, these findings confirm the fact that listeners are sensitive to frequency information on a segmental level in misperception, and as such there is a top-down effect from the lexicon.

4.3 Syllable factors

The previous section examined the role of segmental frequency in segmental misperception. Let us move away from the segment level factors. This section will examine whether factors on the syllable level play a role in segmental misperception.

We will focus on the rate of segmental errors, ignoring the nature of the perceived segments (what the intended segments are misperceived as). Three syllable-based factors are examined for their effect on the error rates. The first factor is the *syllable constituency* – the position of the segment in a syllable (namely onset, nucleus and coda). Do segmental errors occur evenly across the three constituents? The second factor is the *syllable position* – the position of the syllable (that contains the segment) in a word. For a polysyllabic word, three positions can be generalised, namely word initial, word medial, and word final. For a monosyllabic word, this three way categorisation cannot be applied. Do segmental errors occur more often in word-final syllables than word initial syllables? The third factor is *stress* – whether the segment is in a stressed or unstressed syllable. Do segmental errors occur more in unstressed than stressed syllables? These factors (apart from syllable position which

is only relevant in the polysyllabic words) are examined separately for monosyllabic and polysyllabic words.

4.3.1 Syllable constituency

The three constituents are onset, nucleus and coda. Using phonetic arguments and previous experiment findings, predictions can be made as to which constituents are more likely to be misperceived.

Firstly, the difference between the nucleus, and the onset/coda is apparent, as the nucleus contains vowels and the onset/coda contains consonants. Vowels are often longer, and acoustically more intense (cf. sonority) than consonants. Recall that in Section 3.4.1 of Chapter 3 we analysed the overall error rate of vowels and consonants, and we found that consonants are more erroneous than vowels with the rates 19.96% and 17.67% respectively. Therefore, we would naturally expect that the nucleus is less erroneous than the onset/coda.

Secondly, the perceptual difference between onset and coda is less clear. On the one hand, onsets are argued to have a higher degree of cue redundancy, e.g. there is greater redundancy of cues in the CV transition than VC transitions, which is especially true in stop consonants which have VOT and always have release bursts (Wright, 2004). On the other hand, codas are predictable by their preceding nucleus. For instance, the length of the vowel can cue the voicing of the final consonants (pre-fortis clipping) and vowels have been shown to lengthen before fricatives (Peterson and Lehiste, 1960). Vowel nasalisation is mainly caused by nasal codas, so the presence of vowel nasalisation serves as a stronger cue for nasal codas than for nasal onsets. Furthermore, there are studies that suggest the cues of codas are spanned over a greater duration than those of onsets. For instance, formant two and formant three have greater movement in codas than in onsets (Broad and Fertig, 1970); and the transition durations tend to be longer in VC than CV positions (Lehiste and

Peterson, 1961). Besides the acoustic information, the phonotactic information of English makes codas more predictable, because the range of codas is more restricted than that of onsets (Kessler and Treiman, 1997).

Thirdly, while we would naturally expect that the nucleus is less erroneous than onset/coda because of the difference between vowels and consonants, it is possible that the nucleus is just as erroneous as the coda, with the onset being the most erroneous – Onset > Nucleus/Coda (">" means more erroneous than). This prediction is supported by the following facts. Firstly, the phonotactic analyses of Kessler and Treiman (1997) found that consonants have a different distribution within the rime than outside the rime. That is, the co-occurrence constraints lie with the nucleus and the coda more than with the onset and the nucleus in a CVC syllable. Secondly, the phonetic arguments given in the previous paragraph do not only highlight the perceptual salience of codas, but also the fact that the nucleus and the coda overlap more in terms of their acoustic cues than the nucleus and the onset.

Besides using phonetic arguments to form our predictions, we could review previous confusion experiments which tested the error rates of these constituents. The classic confusion study by Wang and Bilger (1973) tested the confusiability of consonants in both CV and VC syllables. As summarised in Chapter 3, Section 3.7.1.3, Wang and Bilger (1973) tested two sets of CV and VC syllables composed of 24 consonants and three vowels. The first set of CV and VC (CV-1 and VC-1) contains the same set of consonants [p], [t], [k], [b], [d], [g], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ]. The second set of CV and VC (CV-2 and VC-2) contains different consonants with [p], [t], [tʃ], [dʒ], [l], [ɹ], [f], [s], [v], [m], [n], [h], [hʷ], [w], [j] for CV-2, and [p], [t], [g], [ŋ], [m], [n], [f], [θ], [s], [ʃ], [v], [ð], [z], [ʒ], [tʃ], [dʒ] for VC-2. On the one hand, CV-2 and VC-2 are in a sense more realistic, as they contain consonants that can only be in either CV or VC, namely [h] and [ŋ]. On the other hand, CV-1 and VC-1 are balanced, which allows for a more direct comparison of onsets and

codas. Codas were more erroneous than onsets in the second syllable set (CV-2 and VC-2); however, this difference is inconsistent with the first syllable set, with codas being more erroneous only at more difficult signal to noise ratios and with the vowel [ɑː, and uː], but not [iː] (Wang and Bilger, 1973, pp. 1251–1252). Despite the inconsistency with CV-1 and CV-2, codas were, overall, more erroneous than onsets. In another confusion experiment, Cutler et al. (2004) tested all possible standard American English CV and VC sequences. However, unlike Wang and Bilger (1973), onsets were more erroneous than codas (with an average 5% difference in error rate).

Besides the confusion experiments of CV, VC syllables, Redford and Diehl (1999) tested 147 CVC syllables (7 consonants × 3 vowels × 7 consonants), and found that codas are more erroneous than onsets. They conducted a further acoustic analysis of the stimuli and found that the perceptual advantage of onsets is partly due to their longer duration and higher amplitude. That is, onsets are produced with greater acoustic distinctiveness than codas. In another CVC confusion experiment conducted by Benkí (2003), it was found that codas were more erroneous than both onsets and nuclei, and that onsets and nuclei are similarly erroneous (Benkí, 2003, pp. 137–140).

The conflicting findings between Cutler et al. (2004) and the other studies (Redford and Diehl, 1999; Benkí, 2003) were examined by Cutler et al. (2004). The author subsetted their data to best match the consonants and vowels used in Redford and Diehl (1999) and Benkí (2003). After the subsetting, they still found onsets being more erroneous than codas. The conflicting findings are therefore likely due to the different experimental conditions, such as the signal to noise ratio, the number of consonants and vowels tested, the number of speakers and listeners tested, etc. Given this mismatch between Cutler et al. (2004) and the other experimental studies, the confusion patterns in the naturalistic corpus could, in fact, be used to settle the debate. In any case, there is converging evidence from multiple confusion studies

(Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003) with the exception of Cutler et al. (2004) that codas are more erroneous than onsets. In addition, Benkí (2003) found that onsets and nuclei are similarly erroneous.

In sum, considering both phonetic arguments and the relative error rates found in previous experimental confusion data, it is unclear what the general pattern is. In fact, a range of predictions can be made regarding the relative error rates of onset, nucleus and coda. They are summarised below (">" means more erroneous than).

- *[Onset, Coda] > Nucleus*
- *Onset > Coda > Nucleus*
- *Coda > Onset > Nucleus*
- *Coda > [Onset, Nucleus]*
- *Onset > [Nucleus, Coda]*

Furthermore, it is unclear whether these predictions would hold for both monosyllabic and polysyllabic words, given that all the above arguments were based on data on single syllables. The perceptual/phonetic arguments were based mostly on experimental data that tested only single syllables (Peterson and Lehiste, 1960; Wright, 2004). The same is true for the phonotactic analysis by Kessler and Treiman (1997) which was also based on single syllables (monomorphemic CVC words). Finally, the experimental confusion data are also restricted to single syllables (CV, VC or CVC).

To conclude, amongst these predictions, the most likely one is the one based on experimental confusion studies, *Coda > [Onset, Nucleus]*, because we are also analysing confusion data (though naturalistic, not experimental). Since the experimental data were based on single syllables, the most conservative prediction is that the error rates in monosyllabic words have the trend – *Coda > [Onset, Nucleus]*, with coda being more erroneous than onset/nucleus. We lack specific predictions for polysyllabic words.

4.3.2 Syllable position

In order to form a prediction regarding the effect of syllable position on error rates, we first consider the acoustic realisation of the segments in word-initial, medial and final positions.

In an acoustic analysis, Lindblom (1968) tested the effect of syllable position on the duration of segments. The author found that segments are longer in final syllables than in medial syllables, which in turn are longer than the segments in initial syllables. This lengthening effect holds for both unstressed and stressed syllables and can be attributed to a final lengthening effect.

The final lengthening effect (Klatt, 1975) is the effect of lengthening the final word of a phrase. Traditionally, experimental studies have examined only the final syllable of the final word, but in fact there is evidence in American English that the lengthening begins before the final syllable (Turk and Shattuck-Hufnagel, 2007). In a production study of American English by Turk and Shattuck-Hufnagel (2007), they found that, besides the final syllable (especially the rime), other regions are lengthened as well. These are the rime of the main stressed syllable (when it is not word final), and the regions between the main stressed syllable and the final syllable (though only sporadically). In sum, generally there is a progressive lengthening effect across the final word of a phrase with an increase of lengthening towards the final syllable.

Using the patterns found in the final lengthening effect, assuming that most words can appear as the final word of a phrase, word final syllables can therefore on average be longer than word medial syllables, which can be in turn longer than word initial syllables. As such, this can be used to form a prediction that segments in word final syllables are less erroneous than those in word medial syllables and in turn are less erroneous than those in word initial syllables, because the longer the duration of a segment, the more perceptually salient it is, which means it is less erroneous. This

can be summarised as: *Word Initial* > *Word Medial* > *Word Final* (">" means more erroneous than).

Alternatively we could extend the phonotactic account of predictability mentioned in Section 4.3.1. Concretely, the number of possible segments in a given segment position in a word decreases as the position moves from left to right. That is, the number of lexical candidates decreases with every additional segment perceived. This is essentially the idea of uniqueness points in word recognition (Luce, 1986a). This could mean that segments in word-final syllables are more predictable than those in earlier syllables, because there are fewer potential lexical candidates. These predictable segments will therefore have lower error rates. In sum, this predictability account makes the same prediction as the duration account, with the same trend *Word Initial* > *Word Medial* > *Word Final*.

4.3.3 Stress

Acoustically speaking, stressed syllables are more perceptually salient than unstressed syllables, such that they have longer duration, higher intensity and they can carry extreme intonation (Browman, 1980). Indeed, a stressed syllable has been argued to be an "island of reliability" (Pisoni, 1981). That is, it contains reliable phonetic information. Furthermore, there are models of word segmentation that rely on the stressed syllable as a segmentation cue (Cutler and Butterfield, 1992; Cutler and Norris, 1988), which implicitly highlights the importance of stress and syllables in perception.

Given the robustness of a stressed syllable, we would expect that segments in a stressed syllable are less likely to be misperceived than those in an unstressed syllable.

4.3.4 Method

All data used in the current section are the naturalistic segmental confusions, which are context-free, as described in Chapter 3, Section 3.2. Three syllable factors were computed for each of the intended segments, namely, syllable constituency (onset, nucleus, coda), syllable position (initial, medial and final), and stress (stressed and unstressed). In addition to these factors, each of the intended segments was tagged as monosyllabic or polysyllabic.

The *glmer* function from *lme4* (Bates et al., 2014) in *R* (R Core Team, 2013) was used to construct logistic mixed-effects models, with the *bobyqa* optimizer. The predictee and predictors are listed below.

Predictee: *Segment Error* (Incorrect vs. Correct)

Predictors of fixed effects: *Syllable Constituency*, *Syllable Position* and *Stress*. All the predictors are categorical.

Variables of random effects: *Intended Words*, *Utterances*, and *Corpora*.

In terms of the fixed effects, the categorical predictors need to be contrast coded. *Stress* is coded as [Unstressed vs. Stressed]. *Syllable Position* is reversed helmert coded, with [Final vs. the mean of Medial and Initial], and [Medial vs. Initial]. The reason for coding syllable position as such is to better capture the progressive effect of syllable position. Finally, *Syllable Constituency* is coded in two different ways, one for monosyllabic words and one for polysyllabic words. For the monosyllabic words, it is coded as [Onset vs. Coda] and [Nucleus vs. Coda]. For the polysyllabic words, it is coded as [Nucleus vs. Onset] and [Coda vs. Onset]. The reason for doing so will become apparent in the beginning of the analyses section.

In terms of the random effects, three variables were included, *Intended Words*, which is the intended word that contains the segment, *Utterances*, which is the

unique number given to each utterance (which is each an instance of misperception), and *Corpora*, which is the seven subcorpora used to construct the combined corpus: Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). These random effects would allow us to control for the variability of segment errors in specific words, utterances, and corpora.

Multiple, separate, mixed-effects models were created to test the fixed effects. They are described during the analyses.

4.3.5 Analyses

In this section, we will begin by visualising the error rates and describe any differences between error rates due to the three factors. After identifying these differences, the statistical models will then be constructed to examine whether these differences are significant. Finally, we will discuss the significant differences and whether they confirm our initial predictions of the three factors.

Before creating the mixed-effects models, we will first visualise the error rates in all possible combinations of the three syllable factors. This is done separately for monosyllabic and polysyllabic words. The visualisation is shown in Figure 4.12. The figure is divided into eight sets of bar charts. In each bar chart, there are three bars, representing the error rates of the three syllable constituents – onset, nucleus and coda – respectively. In each row of the figure, there are three bar charts, representing the error rates of the three syllable positions – initial, medial and final. The bar charts in the first two rows represent the error rates of unstressed polysyllabic words and unstressed monosyllabic words. Finally, the bar charts of the last two rows represent the error rates of the stressed polysyllabic words and stressed monosyllabic words. It is worth noting that the assignment of the syllable position for monosyllabic words as word initial is arbitrary. Interestingly, it has been suggested that monosyllables can be treated as initial syllables in terms of their phonological behaviour (Becker,

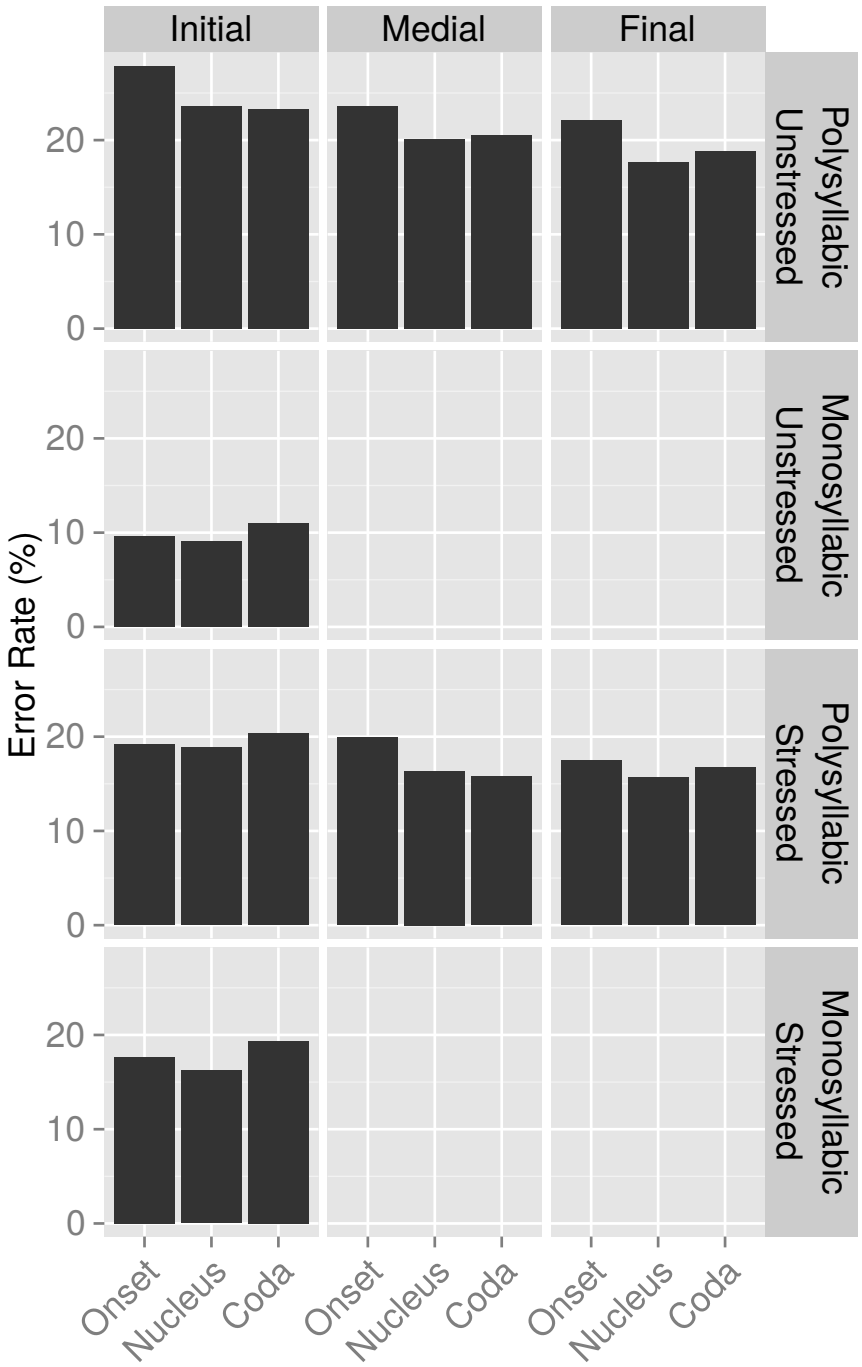


Figure 4.12: Segmental error rates by syllable constituency, syllable position, and stress: error rate is defined as the number of segmental errors in position x divided by the number of segments in position x.

Nevins, and Levine, 2012).

Starting with the unstressed polysyllabic words (the first row), we can see that there is a constituency effect and a syllable position effect. Onset has a higher error rate than nucleus and coda, regardless of the syllable position. Initial syllables have a higher error rate than medial syllables, which in turn have a higher rate than final syllables. Both of these effects can also be found with the stressed polysyllabic words (the third row), but the size of the effect seems to be weaker. The exception is that the constituency effect is absent in the initial stressed syllables in polysyllabic words. Overall, there is a stress effect in the polysyllabic words. Segments in unstressed syllables have higher error rates than those in stressed syllables.

Moving on to the monosyllabic words, there is a definite stress effect. Segments in unstressed syllables have *lower* error rates than those in stressed syllables. The direction of this effect is unexpected, and will be discussed later. Furthermore, there is a subtle constituency effect, with coda being more erroneous than onset and nucleus. This effect holds for both stressed and unstressed monosyllabic words.

4.3.5.1 Polysyllabic words

The polysyllabic words are analysed in a mixed-effects logistic model with syllable constituency, syllable position and stress as fixed effects, and intended word, utterance, and corpora as random intercepts. The model has the formula:

$$\begin{aligned} &Segment\ Error \sim\ Syllable\ Constituency + Syllable\ Position + Stress + \\ &(1|Intended\ Word) + (1|Utterances) + (1|Corpora) \end{aligned}$$

Given the syllable constituency has onset being more erroneous than nucleus and coda (Onset $>$ [Nucleus, Coda]), the following contrast coding is used to test this trend – [Nucleus vs. Onset] and [Coda vs. Onset]. If both of the contrasts are significant, then this would confirm the trend Onset $>$ [Nucleus, Coda].

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-2.0202	0.1379	-14.647	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Onset]	-0.2449	0.0341	-7.182	$6.86 \times 10^{-13***}$
Syllable Constituency [Coda vs. Onset]	-0.1288	0.0448	-2.879	0.004^{**}
Syllable Position [Medial vs. Initial]	-0.1256	0.0289	-4.353	$1.35 \times 10^{-5***}$
Syllable Position [Final vs. (Medial, Initial)]	-0.0855	0.0144	-5.956	$2.58 \times 10^{-9***}$
Stress [Unstressed vs. Stressed]	0.4152	0.0426	9.747	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	2.2454
Utterances (Intercept)	1.9136
Corpora (Intercept)	0.1066

Data size	N
Observations	37,145
Intended Word	3,367
Utterances	3,621
Corpora	7

Table 4.6: Logistic mixed-effects model: predicting segment errors in stressed and unstressed polysyllabic words with syllable factors – syllable constituency, syllable position and stress.

Table 4.6 summarises the mixed-effects model. All three syllable factors – syllable constituency, syllable position and stress – are significant. Stress has the expected effect, such that a segment in an unstressed syllable is more likely to be misheard than that in a stressed syllable, as indicated by the positive estimate. Both contrasts of syllable position, [Medial vs. Initial] and [Final vs. (Medial, Initial)], have a negative estimate, which means a segment in a medial syllable is more likely to be misheard than that in an initial syllable, and a segment in a final syllable is more likely to be misheard than that in a medial or initial syllable. Both contrasts of syllable constituency, [Nucleus vs. Onset] and [Coda vs. Onset], have a negative estimate which means nucleus segments and coda segments are less likely to be misheard than onset segments.

Firstly, the mixed-effects model (as summarised in Table 4.6) confirms stress as a significant factor of segmental errors. It is in the predicted direction, Unstressed

> Stressed, with unstressed syllables being more erroneous than stressed syllables. This finding supports the idea of a stressed syllable being an "island of reliability" (Pisoni, 1981).

Secondly, syllable position is also a significant factor, and again the effect is as predicted, *Word Initial* > *Word Medial* > *Word Final*, with a decreasing amount of errors from the first syllable to the last syllable of a word. This finding can be explained by two accounts. The first account is the final lengthening effect (Lindblom, 1968; Turk and Shattuck-Hufnagel, 2007) (the amount of lengthening increases towards the end of a polysyllabic word). The second account is the predictability effect (cf. the uniqueness point Luce, 1986a) (such that the predictability of each segment increase towards the end of a word).

Thirdly, syllable constituency is also a significant factor. The effect matches one of the predictions, Onset > [Nucleus, Coda], such that onset is a more erroneous constituent than nucleus and coda. This prediction is based on the fact that there is a considerable amount of overlapping of phonetic cues between coda and nucleus, and that the rime is perceptually more salient than the onset. The fact that it does not match the prediction *Coda* > *[Onset, Nucleus]* is not too surprising; although it is based on confusion experiments, the stimuli tested were always monosyllables, and therefore the prediction should not necessarily hold for polysyllables. As we will see later, the prediction *Coda* > *[Onset, Nucleus]* is indeed more appropriate for monosyllabic words.

Finally, while both contrasts of syllable constituency are significant, the contrast [Coda vs. Onset] has a relatively high p-value of 0.004, which is perhaps due to the divergences with the initially stressed polysyllabic word condition as previously mentioned. Given the divergence, another mixed-effects model is constructed to examine whether the syllable constituency factor holds for unstressed syllables. The model has the formula:

$$Segment\ Error \sim Syllable\ Constituency + Syllable\ Position + (1|Intended\ Word) + (1|Utterances) + (1|Corpora)$$

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-2.2335	0.0911	-24.518	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Onset]	-0.0924	0.0533	-1.732	0.0834^{+}
Syllable Constituency [Coda vs. Onset]	0.0056	0.0807	0.070	$0.9443^{n.s.}$
Syllable Position [Medial vs. Initial]	-0.1351	0.0586	-2.303	0.0213^{*}
Syllable Position [Final vs. (Medial, Initial)]	0.0119	0.0358	0.331	$0.7409^{n.s.}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	2.0192
Utterances (Intercept)	2.4475
Corpora (Intercept)	0.0119

Data size	N
Observations	18,389
Intended Word	3,364
Utterances	3,619
Corpora	7

Table 4.7: Logistic mixed-effects model: predicting segment errors in stressed polysyllabic words with syllable factors – syllable constituency and syllable position.

Table 4.7 summarises the findings of the above model. Indeed both factors, syllable constituency and syllable position, are attenuated in stressed syllables. One of the two contrasts of syllable constituency [Coda vs. Onset] is insignificant; the other contrast [Nucleus vs. Onset] is only near-significant. Furthermore, one of the two contrasts of syllable position [Final vs. (Medial, Initial)] is insignificant. One explanation for such an attenuation is that there is a ceiling effect, such that the strong perceptual salience of stressed syllables overshadows the relative perceptual difference across syllable constituents and across syllable positions.

4.3.5.2 Monosyllabic words

The monosyllabic words were analysed in a mixed-effects logistic model with syllable constituency and stress as fixed effects, and intended word, utterance, and corpora

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-1.6967	0.1744	-9.729	$< 2 \times 10^{-16***}$
Syllable Constituency [Nucleus vs. Coda]	-0.2649	0.0361	-7.338	$2.17 \times 10^{-13***}$
Syllable Constituency [Onset vs. Coda]	-0.1594	0.0406	-3.924	$8.71 \times 10^{-5***}$
Stress [Unstressed vs. Stressed]	-1.0064	0.0969	-10.388	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Random effects	Variance
Intended Word (Intercept)	1.656
Utterances (Intercept)	2.236
Corpora (Intercept)	0.187

Data size	N
Observations	50,039
Intended Word	2,119
Utterances	4,286
Corpora	7

Table 4.8: Logistic mixed-effects model: predicting segment errors in stressed and unstressed monosyllabic words with syllable factors – syllable constituency and stress.

as random intercepts. The model has the formula:

$$\textit{Segment Error} \sim \textit{Syllable Constituency} + \textit{Syllable Position} + (1|\textit{Intended Word}) + (1|\textit{Utterances}) + (1|\textit{Corpora})$$

Table 4.8 summarises the mixed-effects model. Both syllable factors, syllable constituency and stress, are significant. Stress has an unexpected effect, such that a segment in an stressed syllable is *more* likely to be misheard than that in an unstressed syllable, as indicated by the negative estimate. Both contrasts of syllable constituency, [Nucleus vs. Coda] and [Onset vs. Coda], have a negative estimate, which means a nucleus segment and an onset segment are less likely to be misheard than a coda segment.

Firstly, the unexpected effect of stress has three potential explanations – a reporting bias in the naturalistic misperception corpora, the differing definitions of stress for monosyllables and polysyllables, and a lexical frequency effect. Recall that unstressed monosyllabic words are essentially function words in our corpus (as defined in Chapter 2, Section 2.2.3.2). Stressed monosyllabic words are therefore content

words, and they carry more information than unstressed monosyllabic words. This would mean misperceiving stressed monosyllabic words would disrupt communication more than misperceiving unstressed monosyllabic words, and therefore they are noticed and reported more frequently by the reporters of the naturalistic corpora (Browman, 1980). This is evident by the fact that in the naturalistic corpus, 543 out of 4,861 instances are misperceptions of a single word, and 535 of which are content words. Regarding the differing definitions of stress, a stressed syllable in a polysyllabic word is stressed relative to the other syllables in the same word. A stressed syllable in a monosyllabic word, however, is stressed relative to other words in the utterance (Browman, 1980). Finally, function words are of higher frequency than content words, and it is possible that high frequency words are less prone to errors; therefore, unstressed monosyllables were less erroneous than stressed monosyllables. This frequency effect is in fact confirmed in Section 4.5.

Secondly, the syllable constituency has the following effect, Coda > [Onset, Nucleus]. This is consistent with the findings from previous confusion experiments which show that coda is more erroneous than onset (Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003) and that onset and nucleus are similarly less erroneous than coda (Benkí, 2003). Recall that this effect is different from the one with the polysyllabic words (Onset > [Nucleus, Coda]) but this is expected given that the previous confusion experiments were all based on the results of monosyllabic stimuli.

Interestingly, the variance of *Corpora* is the smallest of all the random effects in both mixed-effects models (Figure 4.7 and Figure 4.8). This indicates that there is a high level of consistency across corpora.

4.3.6 Conclusion

In this section, three syllable factors were examined for their effects on segmental errors in monosyllabic and polysyllabic words. The three factors were syllable con-

stituency (onset, nucleus and coda), syllable position (word initial, medial and final), and stress (stressed and unstressed).

The effect of syllable constituency in monosyllabic words is that coda segments are more likely to be misperceived than onset segments and nucleus segments – Coda > [Onset, Nucleus]. This is consistent with findings from previous confusion experiments on CV, VC and CVC nonsense syllables. This pattern can be partially explained in terms of acoustic differences between onset and coda (Redford and Diehl, 1999), in which onset is found to be acoustically more distinctive than coda. Regarding onset and nucleus having similar error rates, although Benkí's (2003) confusion data showed this pattern, no explanation was given.

Interestingly, the effect of syllable constituency is different with polysyllabic words, such that the onset is more likely to be misperceived than the rime – Onset > [Nucleus, Coda]. I argued that the mismatch is expected since the arguments and data used to support the trend with monosyllabic words were almost all based on monosyllables.

One partial explanation for this mismatch is that the true effect of syllable constituency for both monosyllabic and polysyllabic words is Coda > [Onset, Nucleus] and the mismatch is due to some of the nuclei and codas of the polysyllabic words having additional cues, which lower their error rate to a level that is even lower than the rate of their corresponding onset. Firstly, in polysyllabic words, the coda consonant in initial and medial syllables could be followed by a sonorant onset (while the coda consonant in final syllables could not), therefore getting additional transitional cues from the sonorant on the right, and these additional transitional cues can lower the error rate. Secondly, using a predictability account, the nucleus and the coda consonants in all syllable positions are predictable using their preceding segments to reduce the number of possible lexical candidates (cf. uniqueness point). This increase in predictability should be greater for the coda than the nucleus because

the coda comes after the nucleus in a syllable. Since the "true" (under this account) effect is that the coda is more erroneous than the nucleus, this greater increase of predictability for the coda would lower its high error rate.

The effect of syllable position was also confirmed, and it has the trend, Word Initial > Word Medial > Word Final. Word initial syllables are more erroneous than medial syllables, which in turn are more erroneous than final syllables. This can be explained with the final lengthening effect (Lindblom, 1968; Turk and Shattuck-Hufnagel, 2007) and/or the predictability effect (Luce, 1986a).

Both syllable constituency and syllable position effects are stronger when the syllables are unstressed rather than stressed. I argued that this attenuation in stressed syllables is due to a ceiling effect caused by the high perceptual salience of stressed syllables overshadowing both the syllable constituency and syllable position effects.

Finally, the effect of stress is that unstressed syllables are more erroneous than stressed syllables, which supports the idea that a stressed syllable is an "island of reliability" (Pisoni, 1981). The pattern, however, showed that this is only true for polysyllabic words, and not monosyllabic words. I argued that this difference is due to a reporting bias in the naturalistic corpus, differing definitions of stress between monosyllabic and polysyllabic words and/or a lexical frequency effect. Stressed monosyllabic words (basically content words) are more noticeable (and therefore reported more often) when misperceived than unstressed monosyllabic words (function words). A stressed syllable in a polysyllabic word is stressed relative to other unstressed syllables in the word, but a stressed monosyllabic word is stressed relative to the other words/syllables in the utterance. Unstressed monosyllabic words are mostly function words which are of high frequency words, and high frequency words are less likely to be misperceived.

To conclude, all three syllable factors had a definite effect on whether a segment will be misperceived. This highlights the fact that factors on the syllable level have

a top-down effect in naturalistic misperception.

4.4 Word frequency

This section examines the relationship between the frequency of the intended word and that of the perceived word. In addition, the relationship between the frequency of the intended segment and that of the perceived segment is also examined in order to eliminate the possibility that the frequency relationship between words can be reduced to the frequency relationship between segments. In other words, this is to see if the relationship of word frequency is the additive result of a lower level of frequency effect.

First, Section 4.4.1 examines the word frequency effect. Second, Section 4.4.2 examines the segmental frequency effect.

4.4.1 Word frequency

Let us start with the frequency relationship between words. Two questions are examined. First, is there a relationship between the frequency of the intended word and that of the perceived word, $Freq._{Perceived} = f(Freq._{Intended})$? Second, is the perceived word just as frequent as or more frequent than the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$ ("$>$" means more than, and "$\approx$" means similar to)?

If there is a relationship (e.g. a correlation), then we need to account for the fact that listeners can somehow estimate the frequency of the word that they were expecting, even though the actual intended word was not perceived. This estimation of the frequency of the intended word can be explained using the graceful degradation account and its extension based on the correlation between lexical frequency and duration (Vitevitch, 2002).

Graceful degradation is the ability of a processing system to not break down in a catastrophic way when the input is incomplete, but to output a representation that best matches the input (McClelland, Rumelhart, and Hinton, 1986). In the context of misperception, the perceptual system uses the information in the degraded signal to retrieve a lexical item. One of the remaining cues of the intended word could be its duration. It is based on the idea that high frequency words tend to be produced more quickly than low frequency words which tend to be produced more slowly (Wright, 1979). So although listeners cannot retrieve the intended word, the listener can still retrieve the duration of the intended word which can be used to derive the lexical frequency (Vitevitch, 2002; Tang and Nevins, 2014).

Furthermore, should we find that the perceived word has a higher frequency than the intended word, then the finding would support an account of ease of lexical retrieval. High frequency words have a lower processing cost than low frequency words and can be retrieved more quickly from the lexicon. When the intended word cannot be retrieved, listeners can either a) do their best to estimate the intended word (using its duration) and select words that can be retrieved more easily, or b) simply select words that are generally easy to retrieve, i.e. high frequency words.

4.4.1.1 $Freq._{Perceived} = f(Freq._{Intended})$

The first question is whether the frequency of the perceived word and that of the intended word have a relationship. This was addressed in previous studies using naturalistic and experimental data. Tang and Nevins (2014) conducted a similar analysis with an earlier version of the combined naturalistic corpus, which was smaller. 2,171 pairs of intended and perceived words were extracted after removing those with zero frequency and duplicates, and there was a positive correlation which is significant at a moderate level ($R = 0.33$, $df = 2{,}169$, $p < 0.002$).

In experimental studies of misperception of English words, Pollack, Rubenstein,

and Decker (1960) analysed word frequency of the intended words and the perceived words and they did not find a significant correlation. The lack of correlation could be due to the fact that the experiment tested only 144 words (which is a small sample), and the fact that the 144 words were repeatedly tested could prime the choice of the perceived words (i.e. there is a higher chance of selecting a test word as a perceived (though incorrect) word (Felty et al., 2013)). In another study, Felty et al. (2013) conducted a large word confusion experiment. 1,428 words which were randomly sampled from the English lexicon were presented in isolation with noise added to the signal. The authors found that there was a positive correlation which is significant at a moderate level ($R = 0.154$, $df = 21{,}842$, $p < 0.0001$) between the intended and perceived words. The fact that this result contradicts that of Pollack, Rubenstein, and Decker (1960) suggests that there is a subtle frequency relationship which can only be found with a larger sample.

Could the positive correlation simply be the results of confounds? Two potential confounds are identified and discussed below. The first one concerns word pairs of different lengths. Firstly, the number of syllables is usually preserved in word confusions (which constitutes 74% of the word confusion errors in Felty et al. (2013)). Therefore, long words are misperceived as long words, and short words as short words. Secondly, longer words are less frequent than shorter words. Together this means that by considering word pairs that contain words of different lengths together, a positive correlation will naturally emerge, even though there could be no (or even negative) correlation with words of the same length. For instance, the monosyllabic word pairs have a zero correlation, and the polysyllabic word pairs also have a zero correlation; but since monosyllabic words are more frequent than polysyllabic words, there will be a positive correlation when considering monosyllabic and polysyllabic word pairs together.

This would in fact explain the lack of correlation in Pollack, Rubenstein, and

Decker (1960) which only tested monosyllabic words. However, Felty et al. (2013) also tested the correlation with only monosyllabic word pairs, and a weak but significant correlation was still found ($R = 0.108$, $df = 6{,}546$, $p < 0.0001$). The fact that the correlation value dropped after controlling for the number of syllables showed that this confound is valid but nevertheless cannot fully explain all the variance. In fact, this was not controlled for in Tang and Nevins (2014), which could contribute to the significant correlation.

Another potential confound is to do with the number of identical word pairs. It is possible that a specific word is confused more often with another word; for instance, in the naturalistic data, the Labov corpus contained five instances of *copy* being perceived as *coffee*. The inclusion of these duplicate word pairs could skew the correlation if we treat the duplicates as independent data points. Tang and Nevins (2014) controlled for this by removing any duplicates, and it is not clear whether this was controlled for in Pollack, Rubenstein, and Decker (1960) and Felty et al. (2013).

In sum, the findings from both naturalistic and experimental data suggest that the frequency relationship between the intended and perceived words is stronger in naturalistic settings than in experimental settings. Furthermore, with the experimental data, the strength of the relationship appears to be dependent on experimental procedures, such as the number of stimuli and whether the stimuli were presented repeatedly. However, there are potential confounds that were not controlled for in some or all of the studies mentioned above, casting doubt on the validity of the findings.

4.4.1.2 $Freq._{Perceived} > or \approx Freq._{Intended}$

The second question concerns the nature of the frequency relationship between the intended and perceived words. This was addressed by previous studies using naturalistic and experimental data.

In naturalistic misperception, Bond (1999, p. 103) randomly sampled 75 pairs of word confusions from the author's own corpus (the Bond corpus) that have relatively simple errors and do not contain proper names. It was found that of the 75 pairs, the perceived word was more frequent than the intended word in 36 pairs, and the reverse is true in the remaining 39 pairs. This difference is not statistically significant under a chi-squared test ($\chi^2 = 0.12$, df = 1, p-value = 0.729).

In another study of naturalistic misperception, Cutler and Butterfield (1992) conducted frequency analyses of word confusions that are involved in juncture misperception, using data from the Bond corpus as well as the author's own unpublished corpus. Juncture misperception is when a word boundary is inserted or deleted, which results in one word being perceived as multiple words and multiple words being perceived as one. Starting with 246 instances of juncture misperception, 165 instances were left after removing those containing proper names or only grammatical words. The authors found that the perceived word was more frequent than the intended word in 81 pairs, and the reverse is true in the remaining 84 pairs; this difference is not significant ($\chi^2 = 0.0545$, df = 1, p-value = 0.8153).

Vitevitch (2002) re-examined this question using the Bond corpus, with a different kind of statistical analysis. The author excluded word pairs that contain complex errors. These word pairs are those with extensive mismatches as well as those which are due to juncture errors (one word is perceived as two words, and vice-versa). Furthermore, certain word pairs were excluded if the lexical variables that the author investigated (one of which is word frequency) were not available. 88 word pairs were left for analyses. An ANOVA (which is identical to an unpaired t-test) was performed using word frequency as the dependent variable and no significant differences were found. The lack of a difference would therefore imply that the intended and perceived words are similarly frequent, and that the perceived words are *not* more frequent than the intended word.

Finally, Tang and Nevins (2014) (previously mentioned) performed a similar frequency analysis. Out of the 2,171 word pairs, the number of pairs with $Freq._{Perceived} > Freq._{Intended}$ is 1,072. In the other direction, the number of pairs with $Freq._{Intended} > Freq._{Perceived}$ is 1,099. A chi-squared test yielded $\chi^2 = 0.3358$, df = 1, p-value = 0.5623, which is statistically insignificant.

Let us move on to experimental misperception. Felty et al. (2013) (previously mentioned) also analysed whether the frequency of the perceived word was significantly different from the frequency of the intended word in word confusions. It was found that the perceived words have a higher frequency than their intended words. To assess the statistical significance, instead of using the chi-squared test or t-test, a Monte Carlo simulation was done – for each word pair, the perceived word is randomly replaced with a word that has the same number of segmental differences from the intended word as the perceived word. 10,000 simulations of the word pairs were performed, and the frequency of the intended word and the fake perceived word was computed across all pairs for each simulation. They found that all 10,000 simulations have a mean difference (the perceived frequency minus the intended frequency) that is lower than the original mean difference, which suggests that the mean difference with the actual word pairs is significant.

Could these findings be explained by confounds? Three confounds are identified and discussed below. The first confound concerns duplicated pairs of word confusions, as they could skew the difference in either direction with the perceived/intended word being more frequent, and could average out any potential differences. This was controlled for in Vitevitch (2002) and Tang and Nevins (2014), but it is not clear if this was controlled for in Bond (1999, p. 103), Cutler and Butterfield (1992) and Felty et al. (2013).

The second potential confound concerns using word confusions that are involved in juncture misperception. Given that a juncture misperception involves perceiving

one word as multiple words and vice-versa, it is not clear from Cutler and Butterfield (1992) which one of the multiple words was chosen as the word for the frequency analysis. For instance, *how big is it?* was perceived as *how bigoted*. Do we take the frequency of *big*, *is* or *it*, to compare with that of *bigoted*? Given that the authors filtered out instances containing only grammatical words, it is likely that they chose the frequency of the content word, but what if there is more than one content word?

The third confound concerns using word confusions in which the intended word and perceived word are of different length (syllables or segments). Longer words tend to be less frequent than shorter words; therefore, whichever word (intended/perceived) is longer in a given pair of words, the frequency will be lower for that word. Say that on average the perceived words have fewer syllables than the intended words, then naturally the frequency of the perceived word will be higher than the frequency of the intended word. In fact, this could explain the findings by Felty et al. (2013). They found that the perceived words largely have the same number of segments and syllables as the intended words, but there is a tendency for the perceived word to be shorter (fewer segments and syllables). The fact that they found that the perceived word are more frequent can be explained by this confound. This was controlled for in Vitevitch (2002) in terms of the number of syllables, but not in Tang and Nevins (2014). It is not clear if it was controlled for in Bond (1999, p. 103) and Cutler and Butterfield (1992).

In sum, these findings from multiple studies of naturalistic misperception suggest that the frequency of the perceived word is not significantly different from that of the intended word in word confusions, i.e. they are similarly frequent. This finding is robust across the size of the sample (N = 75 – 2,171) as well as statistical methods (Chi-squared or ANOVA). Although on the surface the experimental findings contradict with naturalistic findings, the significant difference in the experimental data is perhaps confounded by the difference in word length. Again there are potential

confounds that were not controlled for in some or all of the studies mentioned above, casting doubt on the validity of the findings.

Given the potential confounds, the current analysis will re-examine the two questions while controlling for the confounds mentioned above as well as using both large and small samples of naturalistic data. To recap, the first question is whether there is a relationship between the intended and perceived word, $Freq._{Perceived} = f(Freq._{Intended})$, and the second question is whether the perceived word is more frequent than or similarly frequent to the intended word, $Freq._{Perceived} > or \approx Freq._{Intended}$. Crucially, it is possible that there is not a relationship, $Freq._{Perceived} \neq f(Freq._{Intended})$, and yet the frequency of the perceived word is still higher than the frequency of the intended word. This is the case when the perceived words are generally highly frequent, regardless of the frequency of the intended words. In the next section, I will outline the method that was used by this analysis, including the steps for evaluating the potential confounds.

4.4.1.3 Method

Using the segmental confusion data described in Chapter 3, Section 3.2, 8,259 pairs of word confusions were extracted.

Recall that the word pairs that are involved in juncture misperception can introduce complications when choosing a word pair (e.g. *big is it* > *bigoted?*). Therefore, we removed these many-to-one and one-to-many word confusion pairs, which left us with 4,268 pairs of one-to-one word confusions.

As mentioned in Section Chapter 2, 2.1.2.1, proper names are known to behave differently from non-proper names during lexical retrieval (Valentine, Brennen, and Brédart, 1996); therefore, all word confusions that involved proper names were removed. This left us with 3,244 pairs.

The token frequencies of the words found in these 3,244 pairs were extracted

from a control written English corpus as described in Chapter 2, Section 2.3. After removing the pairs containing zero frequency (i.e. not found in the corpus), 3,135 pairs remained. The token frequency was log10-transformed.

To examine the robustness of the findings, we performed our analysis repeatedly on multiple subsets of the data. The factors that are used to subset the 3,135 pairs are described below. The first factor is the choice of corpora. This is to examine if the findings are affected by certain subcorpora, since they differ in terms of collection locations, collectors' biases and sample sizes. The combined corpus, as well as the subcorpora, were considered individually (the subcorpora are described in Chapter 2, Section 2.1). The Bond corpus was divided into two, the adult misperception corpus and the children misperception corpus. The Nevins corpus was also divided into two, the data that were collected in 2009 and those collected in 2010. Therefore, together there are eight subsets (one combined corpus, and seven subcorpora), which are the Combined corpus, Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). The second factor is whether or not to remove duplicated word pairs. This created two subsets, those with duplicates and those without duplicates. The third factor is the removal of the word pairs with a different number of syllables. Two subsets were created, those with and without these pairs. The fourth factor concerns the number of syllables of the word pairs with matching number of syllables. Three subsets were created, the monosyllabic word pairs, the polysyllabic word pairs and those with both monosyllabic and polysyllabic word pairs.

All possible subsets of the word pairs based on these factors were tested. Should the findings be found consistently across all/most subsets then we can be doubly sure that the findings are not skewed by some or all of these factors.

Correlation analyses were performed for the question of whether the frequency of the intended word correlates with the frequency of the perceived word. A non-

parametric correlation, Spearman (two-tailed), was used to compare the two sets of frequencies, since the two sets of frequency values are not normally distributed.

Paired t-tests were performed for the question of whether the frequency of the perceived word is higher than or similar to the frequency of the intended word for a given substitution. Since the difference between two frequency values is not normally distributed, the p-values are calculated via 10,000 permutations.

4.4.1.4 Analyses

4.4.1.4.1 $Freq._{Perceived} = f(Freq._{Intended})$ Table 4.9 summarises the correlation analyses with all eight subsets of corpora, with and without duplicates. The size of the samples is shown in the columns with the header *N*. The correlation values are under the headers ρ with the level of statistical significance denoted as superscripts.

Overall, we see that all the correlation values are positive, ranging from 0.52 to 0.76, and they are highly significant ($p < 0.001$). This clearly indicates that there is a strong and significant relationship between the frequency of the intended word and that of the perceived word in word confusions. The correlation is robust across subsets of corpora, even when the sample size is small (which is the case with the Bond (Children) corpus, N = 55 or 56). It is also robust with and without duplicates, which is seen by the correlation values only dropping slightly after removing duplicates.

By visualising the correlations, we can get a better idea of the relationship and whether they are skewed by outliers. The correlations of the seven subcorpora with duplicates (excluding the combined corpus) are shown as scatterplots, each fitted with a regression line in Figure 4.13. From the figure, a strong relationship can be seen across all subcorpora, and they do not appear to be skewed/biased by extreme outliers. This indicates that the correlation values are valid.

However, the positive correlation could be due to the fact that we included both

	With Duplicates		Without Duplicates	
Corpus	N	ρ	N	ρ
Combined Corpus	3,135	0.6173^{***}	2,861	0.5767^{***}
Browman (1978)	129	0.5251^{***}	129	0.5251^{***}
Bird (1998)	259	0.7580^{***}	254	0.7466^{***}
Labov (2010)	592	0.5806^{***}	546	0.5798^{***}
Bond (Adult) (1999)	448	0.5380^{***}	440	0.5274^{***}
Bond (Children) (1999)	56	0.6011^{***}	55	0.5949^{***}
Nevins (2009)	811	0.6689^{***}	765	0.6364^{***}
Nevins (2010)	840	0.5703^{***}	815	0.5615^{***}
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.9: Correlations between the frequency of the intended word and the perceived word in word confusions, across corpora, with and without duplicates: the N columns contain the sample size and the ρ columns contain the correlation values; the superscript symbols denote the level of statistical significance.

polysyllabic word pairs and monosyllabic word pairs. Furthermore, it could be an artefact of some word pairs having a different number of syllables between the intended and the perceived words. For these reasons, we excluded the word pairs that have a different number of syllables. We then subdivided the remaining word pairs by whether they are monosyllabic or polysyllabic. We repeat this set of analyses across corpora, with and without duplicates. The correlation results, with and without duplicates, are summarised in Table 4.10 and Table 4.11 respectively. Each table shows the correlation values varied across corpora (vertically) and across subsets of syllable size (horizontally). From the left, the column *Mono. + Poly.* contains the correlations with both monosyllablic and polysyllablic word pairs, and the two on the right, *Mono.* and *Poly.*, contain the correlations with monosyllabic and polysyllabic word pairs respectively.

First of all, we examine the effect of removing pairs with a different number of syllables between the intended and the perceived words. By comparing the third column of Table 4.9 and the third column of Table 4.10, we see that the correlation values increased (though only slightly) after removing these pairs. This increase

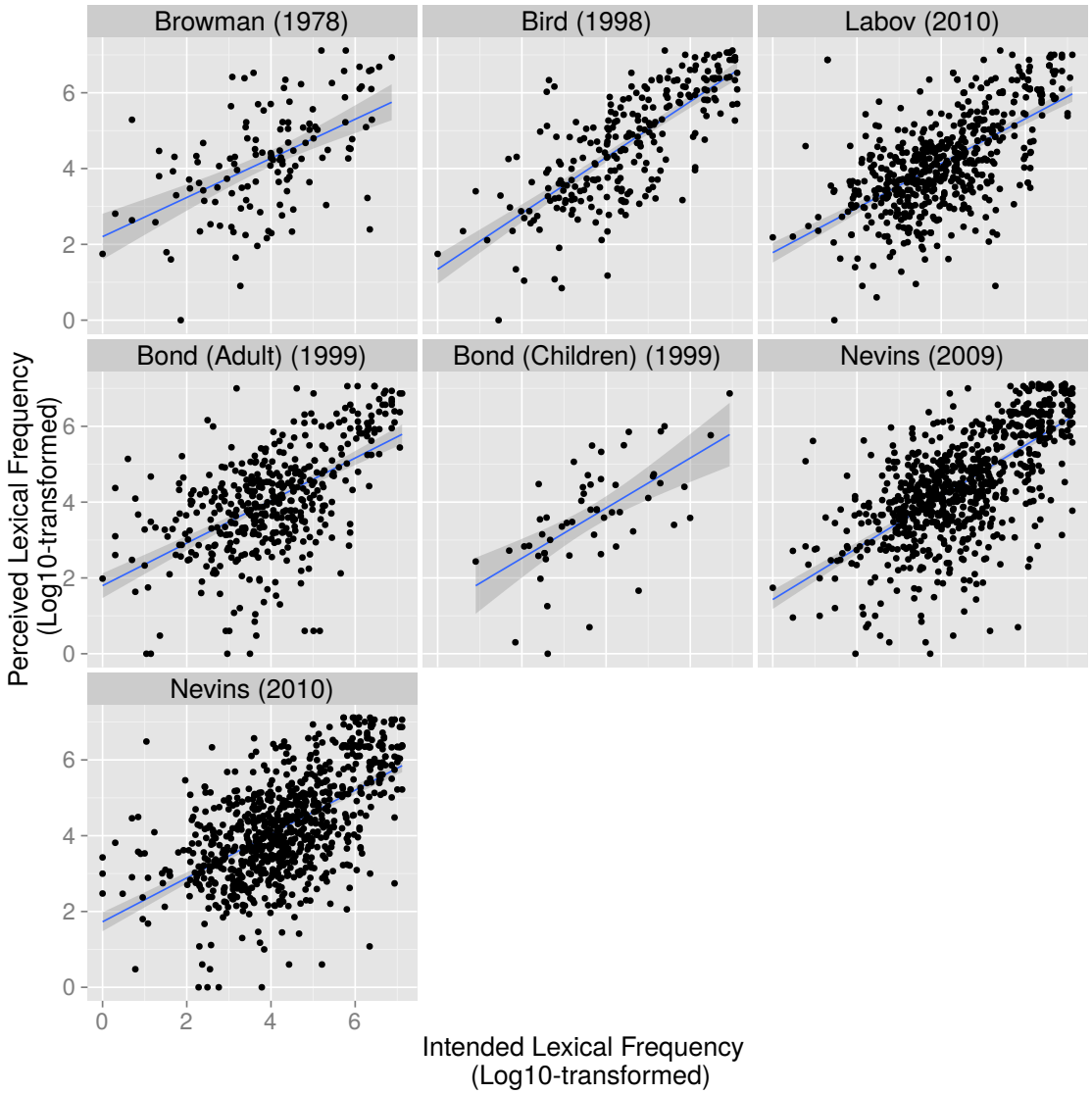


Figure 4.13: The relationship between the frequency of the intended word and the frequency of the perceived word in word confusions with duplicates, divided by corpora

is expected because these pairs have a different number of syllables, and therefore will have a larger difference in frequency. This increase is also true after removing duplicates; this can be seen by comparing the fifth column of Table 4.9 and the third column of Table 4.11.

Second, a comparison between Table 4.10 and Table 4.11, which differ in terms of whether the duplicates were removed, shows again that removing duplicates makes nearly no difference to the findings, as it only slightly lowered the correlation values.

		Mono. + Poly.		Mono.		Poly.
Corpus	N	ρ	N	ρ	N	ρ
Combined Corpus	2,668	0.6465^{***}	1,867	0.6318^{***}	801	0.3318^{***}
Browman (1978)	103	0.5593^{***}	60	0.4533^{***}	43	0.3934^{**}
Bird (1998)	223	0.7587^{***}	164	0.7082^{***}	59	0.5171^{***}
Labov (2010)	516	0.6094^{***}	366	0.6177^{***}	150	0.3094^{***}
Bond (Adult) (1999)	376	0.5962^{***}	252	0.5917^{***}	124	0.3134^{***}
Bond (Children) (1999)	51	0.5720^{***}	31	0.5471^{**}	20	$0.3729^{n.s.}$
Nevins (2009)	702	0.6973^{***}	509	0.6771^{***}	193	0.3691^{***}
Nevins (2010)	697	0.5978^{***}	485	0.5590^{***}	212	0.2646^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.10: Correlations between the frequency of the intended word and the perceived word in word confusions with duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the ρ columns contain the correlation values; the superscript symbols denote the level of statistical significance.

Therefore, we will not examine Table 4.11 any further.

Third, focusing on Table 4.10, we see that all but one correlation were significant. The insignificant correlation is the subset with the polysyllabic word pairs and Bond (Children) (1999) corpus which is likely due to its small sample size (N = 20). The correlations with only the monosyllabic word pairs are similarly stronger (ρ = 0.45 – 0.7) than those with both monosyllabic and polysyllabic words (ρ = 0.55 – 0.75). This agrees with Felty et al.'s (2013) findings which showed that there is a significant correlation with monosyllabic word pairs, and that the positive correlation is not merely an artefact of mixing both monosyllabic and polysyllabic word pairs. While there is a modest correlation with polysyllabic word pairs, their correlation values (ρ = 0.3 – 0.51) were nearly half as low as those with the monosyllabic word pairs (ρ = 0.55 – 0.75). This can be clearly seen in a visualisation of the correlations in Figure 4.14. The figure is divided into seven scatterplots (one for each corpus). Each scatterplot has two sets of points, one for monosyllables and the other for polysyllables, each fitted with a regression line. The difference between monosyllables and polysyllables is apparent, since the slope of the lines with the polysyllables is

	Mono. + Poly.		Mono.		Poly.	
Corpus	N	ρ	N	ρ	N	ρ
Combined Corpus	2,409	0.6020^{***}	1,634	0.5759^{***}	775	0.3180^{***}
Browman (1978)	103	0.5593^{***}	60	0.4533^{***}	43	0.3934^{**}
Bird (1998)	218	0.7460^{***}	159	0.6902^{***}	59	0.5171^{***}
Labov (2010)	477	0.6026^{***}	337	0.6166^{***}	140	0.2532^{**}
Bond (Adult) (1999)	368	0.5863^{***}	244	0.5811^{***}	124	0.3134^{***}
Bond (Children) (1999)	50	0.5650^{***}	30	0.5257^{**}	20	$0.3729^{n.s.}$
Nevins (2009)	656	0.6647^{***}	468	0.6393^{***}	188	0.3610^{***}
Nevins (2010)	673	0.5882^{***}	465	0.5440^{***}	208	0.2731^{***}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.11: Correlations between the frequency of the intended word and the perceived word in word confusions without duplicates, subsetted by corpora and monosyllabicity: the columns N contains sample size and the columns ρ contains the correlation values for each subset; the superscript symbols denote the level of statistical significance.

consistently flatter than that of the lines with the monosyllables. One explanation for this difference between monosyllabic and polysyllabic words is that the word length of the intended and perceived words was only partially controlled by matching the number of syllables, but not the number of segments. Assuming that on average each monosyllable is longer/shorter than another monosyllable by x number of segments, a polysyllabic word pair is likely to have a difference in word length of x times the number of syllables. Therefore, the difference in the number of segments between monosyllabic word pairs is likely to be smaller than that between polysyllabic word pairs. In any case, the overall relationship remains robust after dividing the pairs into monosyllables and polysyllables.

In sum, we found that there is a strong and significant correlation between the frequency of the intended word and the frequency of the perceived word in word confusions. This correlation is robust regardless of the choice of the corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables.

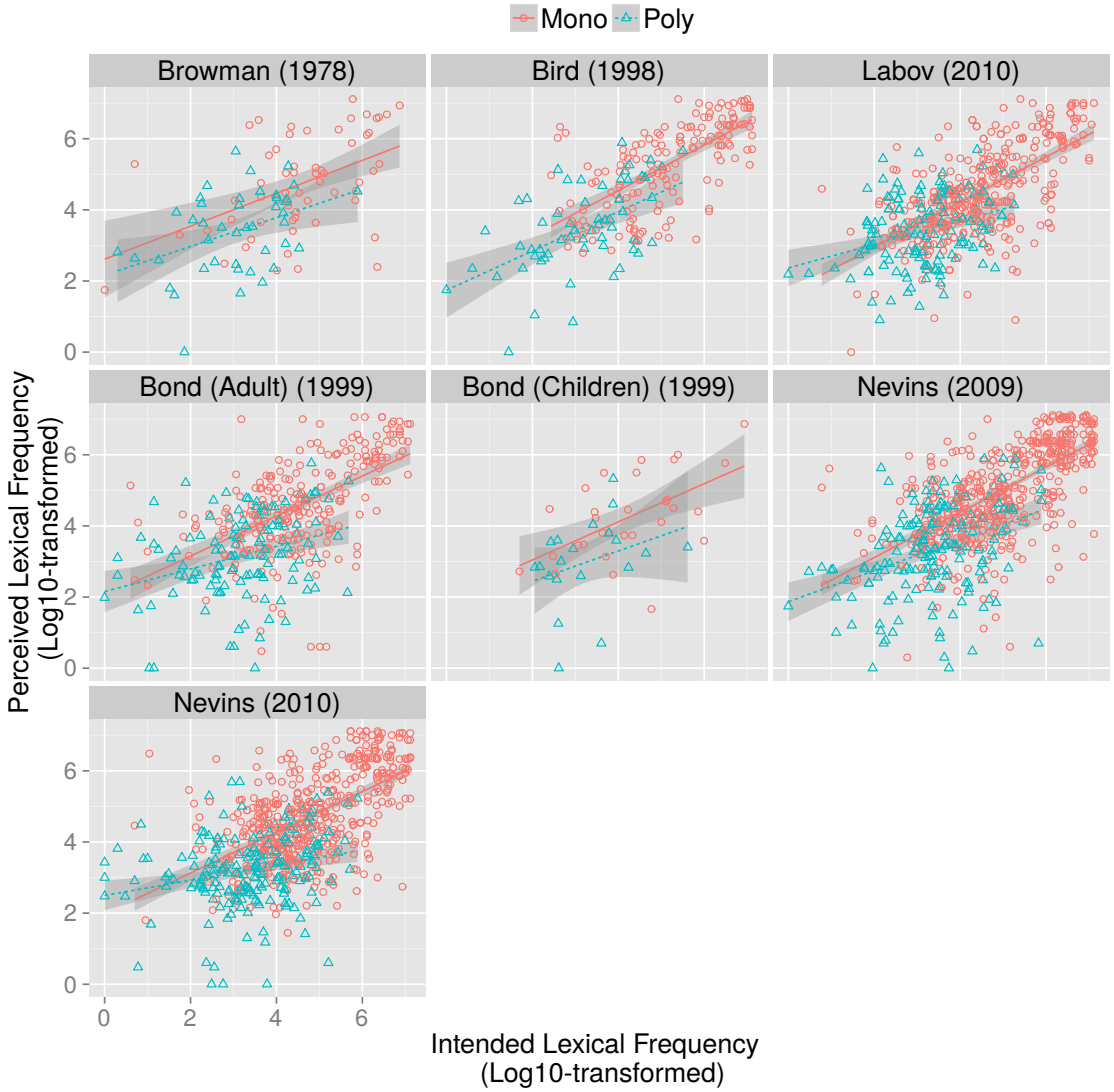


Figure 4.14: The relationship between the frequency of the intended word and the frequency of the perceived word in word confusions, with duplicates, divided by corpora and monosyllabicity

4.4.1.4.2 $Freq._{Perceived} > or \approx Freq._{Intended}$ This section examines the question of whether the frequency of the perceived word is higher than or similarly to that of the intended word in word confusions. Paired t-tests were performed for the same subsets of the word pairs as shown in the previous section.

Table 4.12 summarises the results of the paired t-tests with all eight subsets of corpora, with and without duplicates. The size of the samples is shown in the

columns with the header *N*. The t-values are under the headers *t* with the level of statistical significance denoted as superscripts as well as shown in full between the brackets. A positive t-value indicates that the frequency of the perceived word is higher than that of the intended word, and a negative t-value indicates the reverse. The bold t-values are the statistically significant ones.

Focusing on the third column (with duplicates), there is a tendency for the perceived word to be more frequent in a word confusion ($t = 0.9894$) in the combined corpus; however, it is not significant ($p = 0.1594$). Could it be that a subset of the data has a tendency in the opposite direction, thus averaging out the difference? To tackle this question, we examine each of the subcorpora. In fact, it is true that the difference is being averaged out, since the t-values of the seven subcorpora have inconsistent signs. The t-values of four subcorpora were positive, and they are Browman (1987), Bird (1998), Labov (2010) and Bond (Adult) (1999); all except the Bond corpus were significant. The t-values are negative for the remaining three, and they are Bond (Children) (1999), Nevins (2009) and Nevins (2010); only Nevins (2010) was significant. It is clear that the insignificance in regard to the combined corpus is due to the Nevins (2010) corpus and perhaps the Nevins (2009) corpus averaging out the positive t-values from the other corpora; since they are both relatively large, they therefore have a bigger effect on the combined corpus. These patterns remain the same without duplicates, as shown in the fifth column.

Furthermore, just as the correlation analyses, we analyse the effect of removing pairs that have a different number of syllables and the effect of dividing the pairs into monosyllables and polysyllables. The t-test results (subsetted by monosyllabicity and corpora), with and without duplicates, are summarised in Table 4.13 and Table 4.14 respectively.

First of all, we examine the effect of removing pairs with a different number of syllables between the intended and the perceived words. By comparing the third

Corpus	With Duplicates		Without Duplicates	
	N	t	N	t
Combined Corpus	3,135	$0.9894^{n.s.}$ (p=0.1594)	2,861	$0.6172^{n.s.}$ (p=0.2689)
Browman (1978)	129	$\mathbf{2.5299}^{**}$ (p=0.0071)	129	$\mathbf{2.5299}^{**}$ (p=0.0071)
Bird (1998)	259	$\mathbf{2.2367}^{*}$ (p=0.0124)	254	$\mathbf{2.225}^{*}$ (p=0.0139)
Labov (2010)	592	$\mathbf{2.8826}^{**}$ (p=0.0022)	546	$\mathbf{2.5977}^{**}$ (p=0.0047)
Bond (Adult) (1999)	448	$0.5081^{n.s.}$ (p=0.3023)	440	$0.4512^{n.s.}$ (p=0.3217)
Bond (Children) (1999)	56	$-0.35^{n.s.}$ (p=0.3630)	55	$-0.3628^{n.s.}$ (p=0.3594)
Nevins (2009)	811	$-0.8262^{n.s.}$ (p=0.2072)	765	$-0.8325^{n.s.}$ (p=0.2060)
Nevins (2010)	840	$\mathbf{-2.0081}^{*}$ (p=0.0215)	815	$\mathbf{-2.1356}^{*}$ (p=0.0164)
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.12: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, with and without duplicates and subsetted by corpora: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

column of Table 4.12 and the third column of Table 4.13, we see that the t-values decreased after removing these pairs for most of the corpora, except Bird (1998) and Nevins (2009). This decrease suggests that amongst the word pairs that have a different number of syllables there were more word pairs which have fewer syllables in the perceived word (than the intended word) than those which have more syllables in the perceived word. Nonetheless, the sign of all the t-values and the corpora, which have significant t-values, remained the same. This decrease is also true after removing duplicates, and it is big enough to change the sign of t-test with the combined corpus from positive to negative; this can be seen by comparing the fifth column of Table

Corpus	Mono. + Poly.		Mono.		Poly.	
	N	t	N	t	N	t
Combined Corpus	2668	$0.2723^{n.s.}$ (p=0.3956)	1867	$0.4297^{n.s.}$ (p=0.3319)	801	$-0.1192^{n.s.}$ (p=0.4521)
Browman (1978)	103	$\mathbf{1.7318}^{*}$ (p=0.0429)	60	$0.8733^{n.s.}$ (p=0.1966)	43	$\mathbf{1.7102}^{*}$ (p=0.0486)
Bird (1998)	223	$\mathbf{2.458}^{**}$ (p=0.0076)	164	1.5466^{+} (p=0.0585)	59	$\mathbf{2.0835}^{*}$ (p=0.0201)
Labov (2010)	516	$\mathbf{2.1364}^{*}$ (p=0.0159)	366	$0.9591^{n.s.}$ (p=0.1695)	150	$\mathbf{2.3743}^{**}$ (p=0.0092)
Bond (Adult) (1999)	376	$0.1104^{n.s.}$ (p=0.4573)	252	$0.4062^{n.s.}$ (p=0.3413)	124	$-0.3415^{n.s.}$ (p=0.3676)
Bond (Children) (1999)	51	$-0.5403^{n.s.}$ (p=0.2971)	31	$0.1427^{n.s.}$ (p=0.4446)	20	$-0.9683^{n.s.}$ (p=0.1726)
Nevins (2009)	702	$-0.5292^{n.s.}$ (p=0.2980)	509	$-0.0148^{n.s.}$ (p=0.4947)	193	$-0.8767^{n.s.}$ (p=0.1921)
Nevins (2010)	697	$\mathbf{-2.5034}^{**}$ (p=0.0071)	485	-1.4346^{+} (p=0.075)	212	$\mathbf{-2.2463}^{*}$ (p=0.0137)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.13: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, with duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

4.12 and the third column of Table 4.14.

Second, a comparison between Table 4.13 and Table 4.14, which differ in terms of whether the duplicates were removed, shows again that removing duplicates makes nearly no difference to the findings. The most striking effect is that the t-value of the subset with the monosyllabic word pairs and the combined corpus is reduced from 0.4297 to near zero. Given the small difference, we will not examine Table 4.14 any further.

Third, focusing on Table 4.13, most of the t-values with the monosyllabic word pairs (the fifth column) and those with the polysyllabic word pairs (the seventh col-

Corpus	Mono. + Poly.		Mono.		Poly.	
	N	t	N	t	N	t
Combined Corpus	2,409	$-0.1227^{n.s.}$ (p=0.4532)	1,634	$7 \times 10^{-4\ n.s.}$ (p=0.4998)	775	$-0.2028^{n.s.}$ (p=0.4186)
Browman (1978)	103	$\mathbf{1.7318}^{*}$ (p=0.0411)	60	$0.8733^{n.s.}$ (p=0.1966)	43	$\mathbf{1.7102}^{*}$ (p=0.0486)
Bird (1998)	218	$\mathbf{2.4459}^{**}$ (p=0.0075)	159	1.5308^{+} (p=0.0625)	59	$\mathbf{2.0835}^{*}$ (p=0.0201)
Labov (2010)	477	$\mathbf{1.9388}^{*}$ (p=0.02605)	337	$0.827^{n.s.}$ (p=0.2096)	140	$\mathbf{2.1899}^{*}$ (p=0.0141)
Bond (Adult) (1999)	368	$0.0467^{n.s.}$ (p=0.4815)	244	$0.3259^{n.s.}$ (p=0.3762)	124	-0.3415 (p=0.3666)
Bond (Children) (1999)	50	$-0.5537^{n.s.}$ (p=0.293)	30	$0.1247^{n.s.}$ (p=0.4493)	20	$-0.9683^{n.s.}$ (p=0.1724)
Nevins (2009)	656	$-0.5343^{n.s.}$ (p=0.2964)	468	$-0.0702^{n.s.}$ (p=0.4711)	188	$-0.8047^{n.s.}$ (p=0.2144)
Nevins (2010)	673	$\mathbf{-2.6324}^{**}$ (p=0.0044)	465	-1.5905^{+} (p=0.0556)	208	$\mathbf{-2.2519}^{*}$ (p=0.0114)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$

Table 4.14: Paired t-tests (one-tailed) on the frequency of the intended and perceived words, without duplicates, subsetted by corpora and monosyllabicity: the N columns contain the sample size and the t columns contain the t-values; the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold values are the significant t-values.

umn) have similar signs to those with both the monosyllabic and polysyllabic word pairs (the third column). However, the significant t-values with both the monosyllabic and polysyllabic word pairs are significant only with the polysyllabic word pairs, and not the monosyllabic word pairs. This suggests that the overall pattern is dependent mostly on polysyllabic word pairs.

In sum, we found that the difference between the frequency of the intended word and that of the perceived word is inconsistent across corpora. The significant negative t-values with the Nevins (2010) corpus were unexpected, given that previous findings found either a null difference or a positive difference (with the perceived word being more frequent). It is not clear why Nevins (2010) and Nevins (2009) have a negative

difference, while the other corpora have a positive difference. This inconsistency suggests that the difference (positive or negative) is not robust; therefore, it is not the case that listeners would choose a higher frequency word in word confusions, but instead they would choose a similarly frequent word.

4.4.1.5 Conclusion

This section examined the two aspects of the frequency relationship between the intended and perceived words.

Firstly, we found that there is a strong and significant relationship between the frequency of the intended word and that of the perceived word – $Freq._{Perceived} = f(Freq._{Intended})$, as found in the correlation analyses. This is consistent with the large-scaled experimental confusion study by Felty et al. (2013), which also found a statistically significant correlation. Interestingly, the strength of the correlation is nearly twice as strong in our naturalistic data than the Felty et al.'s (2013) experimental data. Since Felty et al.'s (2013) study tested single words which were presented in isolation, and the naturalistic data were based on words in sentences, one might expect the frequency relationship to be stronger in Felty et al.'s (2013) study than ours. One possible explanation is that the experiment was unnatural, since in our everyday life we do not listen to words in isolation. This unnaturalness of the experiment might therefore attenuate the strength of frequency effect.

Furthermore, the polysyllabic words had weaker (but significant) correlations than the monosyllabic words, for which I have no concrete explanation. One explanation is that the difference in the number of segments between monosyllabic word pairs is likely to be smaller than that between polysyllabic word pairs, because the word length of the intended and perceived words was not matched by the number of segments.

In any case, having controlled for potential confounds, such as the choice of the

corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables, this correlation remains robust.

Secondly, we found that overall the frequency of the perceived word and the frequency of the intended word are not significantly different, which suggests that they are similar – $Freq._{Perceived} \approx Freq._{Intended}$. By subsetting the naturalistic corpus, in some of the subcorpora the perceived word was found to be significantly higher than the intended word, while in the rest of the subcorpora the direction was either reversed or insignificant. This contradicts Felty et al.'s (2013) findings that the perceived word is more frequent; however, Felty et al.'s (2013) finding could be the result of a word length confound. They found that the perceived word was on average shorter, with fewer segments and syllables, than the intended word. Given the inverse relationship between length and frequency, this naturally means that the perceived word will be more frequent. A conservative conclusion is that listeners do not simply retrieve a more frequent word.

To conclude, the findings in this section suggest that when the signal is degraded listeners would estimate the intended word with the remaining cues in the signal, such as the duration of the word. Given the relationship between frequency and duration (Wright, 1979), the resultant perceived word is therefore of a similar frequency to the intended word. Listeners do not simply retrieve an easier/more frequent word, which suggests that we should reject an ease of retrieval account.

To eliminate the possibility of this word frequency effect being the result of a segmental frequency effect, the same analyses are conducted for segments and are presented below.

4.4.2 Segmental frequency

Two questions are examined. Does the segmental frequency of the intended segment have a relationship with the segmental frequency of the perceived segment,

i.e. $Freq._{Perceived} = f(Freq._{Intended})$? Is the frequency of the perceived segment similar to or higher than that of the intended segment, i.e. $Freq._{Perceived} > or \approx Freq._{Intended}$?

Besides these two key questions, the strength of these patterns is examined between consonants and vowels, and between the three frequency measures (token frequency, type frequency and weighted type frequency) as described in Section 4.2. Three common frequency measures are examined to rule out the possibility that the strength of the relationship is strongly dependent on the chosen measure. Should we find that the segmental frequency relationship is weak (even with the best frequency measure) compared to the word frequency relationship, then it would suggest that the word frequency effect is not a by-product of the segmental frequency effect.

4.4.2.1 Method

Given that we are interested in the frequency relationship between the intended and the perceived segments in a substitution, only substitution errors are considered; and the correctly perceived segments, as well as insertion and deletion errors, were ignored. This left us with 3,329 substitution errors with the vowels, and 4,789 substitution errors with the consonants. The segmental frequency of the consonants and the vowels in the language was computed for the intended segments and perceived segments of these substitution errors.

Correlation analyses were performed for the question of whether the frequency of the intended segment is correlated with the frequency of the perceived segment. A non-parametric correlation, Spearman (two-tailed), was used to compare the two sets of frequencies, since the two sets of frequency values are not normally distributed.

Paired t-tests were performed for the question of whether the frequency of the perceived segment is higher than or similar to the frequency of the intended segment for a given substitution. Since the difference between two frequency values is not

normally distributed, the p-values were calculated via permutations. This is done with the following steps with 10,000 permutations (N = 10,000).

1. The t-value from the observed data is first calculated.
2. The data is then shuffled and a corresponding t-value is calculated.
3. The last step is repeated N times.
4. The p-value is the proportion of the absolute t-values from the shuffled data that are greater than the t-value from the observed data.

4.4.2.2 Analyses

4.4.2.2.1 $Freq._{Perceived} = f(Freq._{Intended})$ Table 4.15 summarises the correlation analyses for consonants and vowels, testing the relationship between the frequency of the intended segments and that of the perceived segments. The table shows the correlation values with their respective levels of statistical significance (as indicated by the superscripts).

	Unfiltered		Filtered	
Frequency Measure	Consonants	Vowels	Consonants	Vowels
Token	0.1631^{***}	$0.0026^{n.s.}$	0.1546^{***}	$0.0026^{n.s.}$
Type	$\mathbf{0.2032}^{***}$	0.1025^{***}	$\mathbf{0.2029}^{***}$	0.1025^{***}
Type (Weighted)	0.1868^{***}	$\mathbf{0.1109}^{***}$	0.1831^{***}	$\mathbf{0.1109}^{***}$
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$				

Table 4.15: Segmental frequency correlations (Spearman, two-tailed) of consonants between the intended and perceived segments with three frequency measures: the superscript symbols denote the level of statistical significance; the bold value in each column is the best correlation amongst the three frequency measures.

First, we focus on the consonant correlations. In the second column, we can see that the correlation ranges from 0.16 to 0.20 across the three frequency measures, all of which are highly significant. Both measures of type frequency yield better

correlations than token frequency. Furthermore, we find the unweighted type frequency yields a better correlation than the weighted measure. Both of these findings are expected, since previous studies claimed that pattern strength in the lexicon is determined by type frequency and not token frequency (Bybee, 1995; Albright and Hayes, 2003; Hay, Pierrehumbert, and Beckman, 2004). While the correlations are significant, their strength is modest (0.1 to 0.3).

By visualising the correlations, we can have a better idea about the nature of the relationship. These three correlations are visualised as scatterplots, each fitted with a linear regression line with confidence intervals in Figure 4.15. It is immediately clear that all three plots have clear outliers at low frequency values. These outliers could inflate the correlation values. This is resolved by filtering these outliers by excluding any values that have a frequency value above or below 3 standard deviations from the mean frequency value. These filtered correlations are visualised in Figure 4.16. The gradients of the line of best fit appears to be unaffected by the filtering step. This is confirmed by their respective correlation values and levels of statistical significance as shown in the fourth column of Table 4.15. The filtered correlation values are only marginally smaller than the unfiltered ones. Together, the modest correlations (with or without extreme values) and scatterplots suggest that there is a weak relationship between the frequency of the intended segment and that of the perceived segment in substitution errors of consonants.

Moving on to the vowel substitutions, we visualise the relationship, as shown in Figure 4.17. It is clear that the slopes are flatter than those of the consonant substitutions. The slope of the token frequency is almost entirely flat, suggesting a zero correlation (i.e. no relationship). These observations are indeed confirmed in the correlations in the third column of Table 4.15. It is worth noting that the filtering step did not filter any values for vowels; therefore, the fifth column in the table is identical to the third column. Again we found that the two measures of

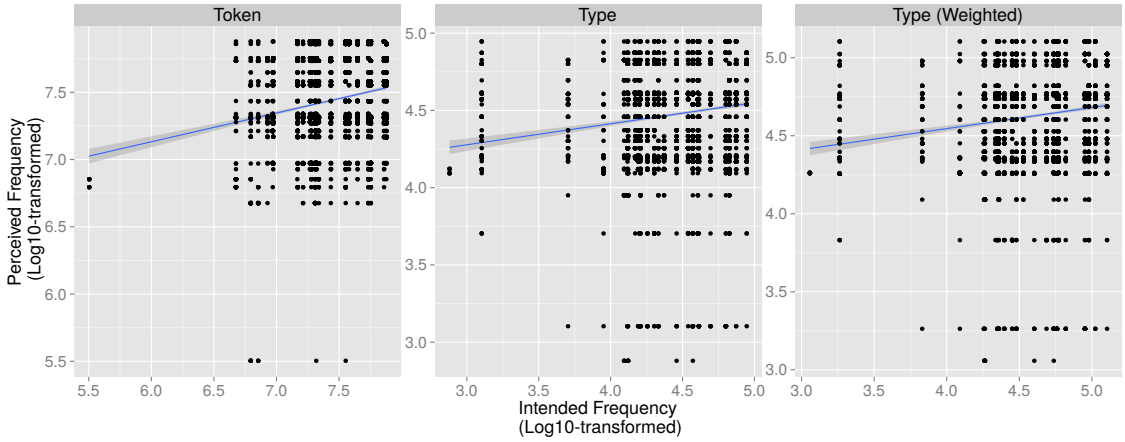


Figure 4.15: The relationship between the intended frequencies and the perceived frequencies of consonant substitutions

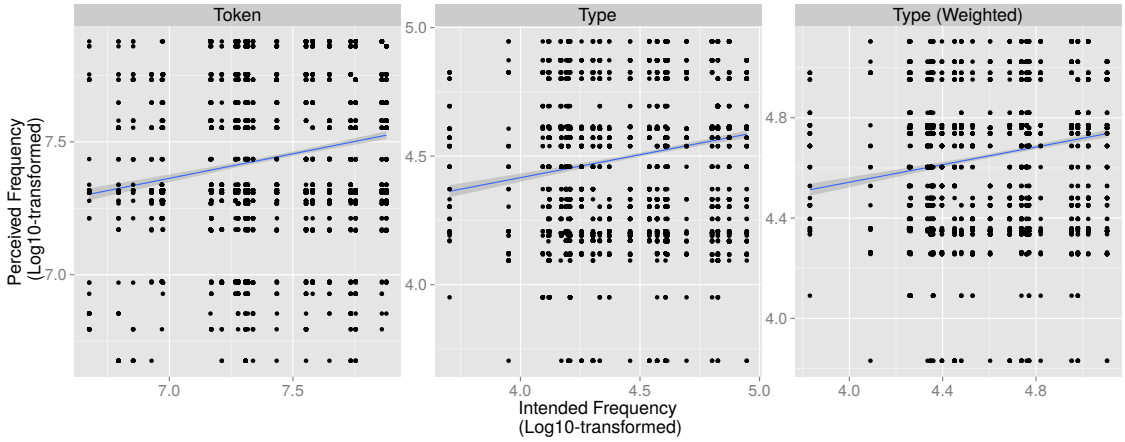


Figure 4.16: The relationship between the intended frequencies and the perceived frequencies of consonant substitutions, without extreme values

type frequency outperform token frequency in terms of the correlation values and the p-values. Crucially, the correlation with token frequency is extremely weak (ρ = 0.0026) and insignificant. Interestingly, unlike the consonant substitutions, the weighted type frequency yields a higher correlation than the unweighted measure. In any case, both the modest correlations (with the two measures type frequency) and the scatterplots suggest that there is a weak relationship between the frequency of the intended segments and that of the perceived segments for the vowels.

In sum, both consonant and vowel substitutions are governed partly by the sim-

ilarity of segmental frequency. This frequency bias is stronger for the consonants than for the vowels. The bias is sensitive to lexical information, as suggested by how the type frequency measures outperform the token frequency measure. Overall, the strength of the frequency bias is weak.

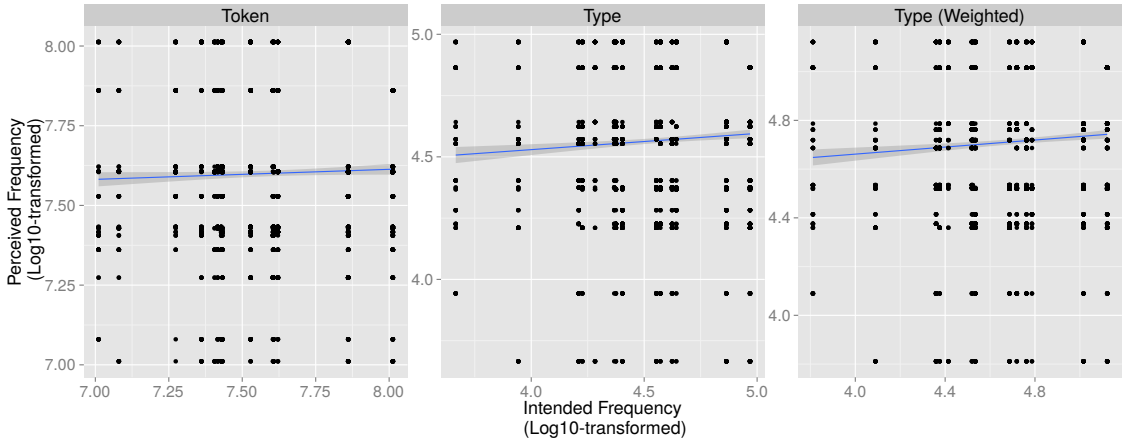


Figure 4.17: The relationship between the intended frequencies and the perceived frequencies of vowel substitutions

4.4.2.2.2 $Freq._{Perceived} > or \approx Freq._{Intended}$ The last section found that the frequency of the intended segments and that of the perceived segments are correlated. However, this does not necessarily mean that the frequency of the perceived segment tends to be higher than that of the intended segment. In this section, this question is examined.

Table 4.16 summarised the results of the paired t-tests (one-tailed), testing whether the frequency of the perceived segment is significantly different from that of the intended segment for each pair of substitutions.

The table shows that all the t-values are positive, which suggests that the frequency of the perceived segments is higher than that of the intended segments; therefore, they are not similarly frequent. However, most of the p-values were greater than 0.1, meaning that this difference is small. In terms of significance levels, only the consonants with the two type frequency measures are significant ($p < 0.05$). We

Frequency Measure	Consonants	Vowels
Token	$0.889^{n.s.}$ (p=0.1875)	$0.6061^{n.s.}$ (p=0.2714)
Type	$\mathbf{1.7687}^{*}$ (p=0.0395)	$0.6407^{n.s.}$ (p=0.2597)
Type (Weighted)	1.7625^{*} (p=0.0394)	$\mathbf{0.6563}^{n.s.}$ (p=0.2548)
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$, $^{n.s.}p > 0.1$		

Table 4.16: Paired t-tests (one-tailed) on the intended and perceived segments of consonants and vowels with three frequency measures: the level of statistical significance is denoted by the superscript symbols and the p-values are shown between the brackets; the bold value in each column is the highest t-value amongst the three frequency measures.

found that a) the two measures of type frequency yield greater t-values (i.e. a bigger difference) than the token frequency measure, b) the unweighted frequency measure yields a greater t-value than the weighted measure for the consonants and the reverse is true for the vowels, and c) the difference (across all three frequency measures) is larger for the consonants than for the vowels. These three findings were also found in the correlation analyses earlier.

In sum, there is a weak tendency for the perceived segments to be of a higher frequency than the intended segments. This tendency is statistically significant for the consonants, but not for the vowels.

4.4.2.3 Conclusion

This section examined the frequency of the intended segments and the frequency of the perceived segments that are involved in substitution errors. Concretely, we examined whether the segmental frequency of the intended segment has a relationship with the segmental frequency of the perceived segment, and whether the frequency of the perceived segment is similar to or higher than that of intended segment.

In Section 4.4.2.2.1, the analysis revealed that there is a weak relationship be-

tween the frequency of the intended segment and that of the perceived segment in substitution errors, as indicated by the modest correlation values and the corresponding scatterplots. The relationship is stronger for consonant substitutions than for vowel substitutions. Section 4.4.2.2.2 found that the perceived segment tends to be more frequent than the intended segment in a given substitution, but this tendency is only significant for the consonants.

Together, the findings suggest that the relative segmental frequencies play a minor role in consonant substitutions, and an even more minor (practically non-existent) one in vowel substitutions. The overall segmental frequency effect is weak *regardless* of the frequency measure. Comparing these findings with those in Section 4.4.1 on word frequency, it is clear that the word frequency effect cannot be reduced to a segmental frequency effect. In other words, the graceful degradation account plays a significant role in word misperception and not in segmental misperception.

4.5 Self-information

Existing models in psychoacoustic research have been developed to predict the overall speech intelligibility under the effect of noise; for instance, the Articulation Index (AI) (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Rhebergen, Versfeld, and Dreschler, 2006) and Speech Intelligibility Index (SII) (ANSI, 1997). Crucially these models predict whether an utterance in a given degraded signal will be erroneously perceived.

These models can be complemented by models that predict *which* part of the utterance will be misperceived. In fact, the analyses in Section 4.3 did precisely this by using syllable factors to predict which *segment* in the word will be misperceived. Now, we will extend the size of the unit from a segment to a word by predicting which *word* in the utterance will be misperceived.

As we have examined in Section 4.4, word frequency plays a definite role in misperception in terms of lexical retrieval; the choice of the (incorrectly) perceived word is a function of the frequency of the intended word in word confusions. This section will examine a different effect of frequency on misperception, which is the predictability of word errors. In other words, if a multi-word utterance contains at least one word error, which words are most likely to be misheard?

To do so, we will examine whether the self-information of a word can predict whether a word will be misheard. By self-information, we are referring to Shannon information, which is a function of the average unpredictability in a random variable (Shannon, 1948). The Shannon information of a word, $I(word)$, is the negative log of the probability of a word, $-log(P(word))$. The probability of a word, $P(word)$, is the unconditional probability of a word, which is the number of times a word appears in a sample of the language (such as our written control corpus) divided by the total number of words in the sample. Therefore, $I(word)$ is perfectly correlated with the log of word frequency. In other words, a low frequency word has more information than a high frequency word.

In natural language, we speak in context and not with isolated words. But how do we quantify the probability of a word given its context? Since it is difficult to model the real context (e.g. world knowledge), an estimate would be to model the local context using a language model. A language model is a probability model that assigns probabilities to sentences (and indeed the words in the sentences). Using a language model, it would be possible to estimate the conditional probability of a word given the previous words, which is then converted to Shannon information, $I(word|context) = -log(p(word|context))$. That is to say, in an multi-word utterance, the self-information of a word is dependent on its preceding words. It is important to note that while the unconditional probability and the conditional probability of a word usually correlate with each other, they are not identical. In

terms of the terminology, it is worth noting that the self-information based on the conditional probability of a word given its previous words is also called *surprisal* (Hale, 2001; Levy, 2008).

What mechanisms govern the relationship between self-information and word errors? In fact, self-information plays a role in both perception and production. Each raises a different prediction on how self-information can predict word errors.

In perception, it is well-known that the processing cost of a word is a function of word frequency. That is, a high frequency word is processed faster than a low frequency word, as repeatedly demonstrated in lexical decision tasks (Brysbaert and New, 2009; New et al., 2007; Keuleers, Brysbaert, and New, 2010; Ernestus and Cutler, 2014). This is also true for the conditional probability of a word; for instance, the reading time of naturalistic texts is shorter for words that have a higher conditional probability (Smith and Levy, 2008). In other words, the processing cost is a function of the self-information of both words, $I(word)$, and words given their context, $I(word|context)$. Given this relationship between self-information and processing cost, we can expect that a word with high self-information is more likely to be misperceived because of its high processing cost.

In production, the phonetic realisation of a word is a function of its self-information. Words with high self-information tends to be spoken over a longer period (Wright, 1979; Aylett and Turk, 2004). The same applies to phonemes with high self-information, which are produced more slowly and with more articulatory detail (van Son and van Santen, 2005). More frequent (therefore less informative) words tend to undergo morphophonological reduction and alternation (Bybee, 1995; Bybee and Hopper, 2001; Bybee, 2001; Coetzee and Kawahara, 2013). Given this relationship between self-information and phonetic realisation, we would expect a word with low self-information more likely to be misperceived because of its weak phonetic cues.

Having discussed the two predictions of how self-information can predict word errors, we will describe the steps for computing the language model as well as the statistical models in the next section.

4.5.1 Method

4.5.1.1 Data selection

The combined naturalsitic corpus from Chapter 2 is used for this analysis.

All Mondegreens (misperception of music lyrics) (253 instances) and non-English misperceptions (69 instances) were excluded. The filtered corpus contains 4,861 instances of misperception. In addition, we applied two more filters. First, we removed instances that do not contain any word errors. 42 of these instances were found and removed. These instances are mostly from the Labov corpus and they consist largely of reference errors, such as the pronoun 'her' being referring to two different female entities. Second, we remove instances that contain only errors (e.g. one-word utterances). 948 instances were found and removed. The remaining instances are multi-word intended utterances with at least one word error. This left us with 3,871 instances.

4.5.1.2 Probability estimation

For each intended sentence, the conditional and unconditional probabilities of each word was estimated. The unconditional probability of a word is simply its token frequency divided by the total number of words in the control written corpus. The conditional probability of each word was computed over a language model as described below.

The probabilities were estimated by a trigram language model trained on our 353.4 million word corpus as described in Chapter 2, Section 2.3. The model was estimated using MIT Language Modeling Toolkit (v. 0.4.1) (Hsu, 2009; Hsu and

Glass, 2008), with modified Kneser-Ney smoothing (Kneser and Ney, 1995). In a trigram model, the probability of a word given its context is modelled with the probability of a word given the two previous words. If there are more than two words before a given word, then the chain rule is applied. The modified Kneser-Ney smoothing is a standard smoothing technique for trigram models (Chen and Goodman, 1999). KenLM (Heafield, 2011) was used to make queries with the model.

4.5.1.3 Statistical model

The *glmer* function from *lme4* (Bates et al., 2014) in *R* (R Core Team, 2013) was used to construct logistic mixed-effects models, with the *bobyqa* optimizer. The predictee and predictors are listed below.

Predictee: *Word Error* (Incorrect vs. Correct)

Predictors of fixed effects: $I(word)$, $I(word|context)$, *Word Length* and *Proper Name* (Proper vs. Non-Proper). All the predictors are continuous, except *Proper Name* which is categorical.

Variables of random effects: *Utterances*, *Utterance Length* and *Corpora*

In terms of the fixed effects, the two predictors of interest are the self-information of a word using its unconditional probability, $I(word) = -log(p(word))$, and that of a word using its conditional probability, $I(word|context) = -log(p(word|context))$. In addition, we included two predictors which are the controls. They are *Word Length* (estimated as the number of IPA segments in a word), and *Proper Name* (whether a word is a proper name). These control predictors were added to control for confounds, because word length is known to affect processing cost (the longer the word, the higher the cost), proper names are known to behave differently from non-proper names during lexical retrieval (Valentine, Brennen, and Brédart, 1996),

and their probability estimates might be inaccurate (e.g. the lexical frequency of *Harvard* is likely to be differ more between individuals than that of *apple*.)

In terms of the random effects, three variables were included, *Utterances*, which is the unique number given to each utterance (which is each instance of misperception), *Utterance Length*, which is the number of words in each utterance, and *Corpora*, which is the seven subcorpora used to construct the combined corpus: Browman (1978), Bird (1998), Labov (2010), Bond (Adult) (1999), Bond (Children) (1999), Nevins (2009) and Nevins (2010). These random effects would allow us to control for the variability of word errors in specific utterances, length of utterance and corpora.

To reduce collinearity, all continuous predictors, $I(word)$, $I(word|context)$ and Word Length, were standardised by scaling and centering as z-scores. The standardized predictors are henceforth referred to as $z[I(word)]$, $z[I(word|context)]$ and $z[Word\ Length]$. Following the recommendation of Rogerson (2001), any predictors with a variance inflation factor (VIF) over 5 will indicate collinearity in the model. Since all our predictors have VIF < 5 (the highest VIF was 4.97), collinearity is unlikely to be a problem.

We first fitted a model with the single terms of the predictors as fixed effects and two random intercepts.

Superset model:

$$\begin{aligned} Word\ Error \sim\ & z[I(word)] + z[I(word|context)] + z[Word\ Length] + \\ & Proper\ Name + (1|Utterances) + (1|Utterance\ Length) + (1|Corpora) \end{aligned}$$

We then performed a series of nested model comparisons on the fixed effects using ANOVA (test $= \chi^2$, $\alpha = 0.05$). The removal of terms was justified by whether a significant improvement to the model was made. If there were multiple subset models that resulted in p-values exceeding the α-level in their nested model comparisons with the superset model, the subset model with the strongest evidence (the highest p-value) was selected. We arrived at the following best model.

Best model:

$$Word\ Error \sim\ z[I(word)] + z[I(word|context)] + z[Word\ Length] + (1|Utterances) + (1|Utterance\ Length) + (1|Corpora)$$

4.5.2 Analyses

The complete summary of the best model is shown in Table 4.17. The model suggests the following significant predictors: $z[I(word)]$, $z[I(word|context)]$ and $z[Word\ Length]$. The predictor *Proper Name* was dropped during the nested model comparison, indicating that it makes an insignificant contribution to the model, and as such it is not a useful predictor of word errors.

In the table, a positive estimate means that the corresponding predictor has a positive relationship to the likelihood of a word error, and therefore a negative estimate means that there is a negative relationship. Given that we have converted our continuous predictors to z-scores, the absolute values of the estimates are now comparable between predictors.

From the table, the strongest to the weakest predictors are $z[I(word)]$, with an estimate of 1.1247, the control predictor $z[Word\ Length]$, with an estimate of -0.4144, and $z[I(word|context)]$, with an estimate of 0.2764.

Firstly, both measures of self-information survived the nested model comparison. This means that they are both important predictors of whether a word will be misheard. Secondly, the signs of the estimates of the $z[I(word)]$ and $z[Word\ Length]$ are both positive. This supports the prediction that there is a relationship between processing cost and self-information, such that a word with high self-information is more likely to be misheard. The prediction that word errors are dependent on phonetic reduction (due to low self-information) is rejected.

Secondly, the control predictor, $z[Word\ Length]$, has an estimate of -0.4144, which suggests that the longer the word is, the less likely the word would be misper-

ceived. This is consistent the findings in previous studies such as Wiener and Miller (1946) and Felty et al. (2013). Given that the word confusions in Felty et al. (2013) were induced by presenting participants with words in isolation, the fact that the word length effect is also found in our naturalistic corpus suggests that the length effect is robust not only in isolation but also in words that are presented together with other words.

Finally, R^2_{GLMM}, the percentage of variance explained by the model, is calculated (Nakagawa and Schielzeth, 2013; Johnson, 2014; Bartoń, 2014) with marginal R^2_{GLMM} being 24% and conditional R^2_{GLMM} being 34%. Marginal R^2_{GLMM} represents the variance explained by fixed effects and conditional R^2_{GLMM} represents the variance explained by both fixed and random effects. The difference between conditional R^2_{GLMM} and Marginal R^2_{GLMM} indicates that the random effects did capture a sizeable portion (10%) of the variance. From Table 4.17, we see the variance of the *Utterance Length* is the highest of all three random intercepts, and therefore it contributes most towards the 10% of the variance. In other words, there is a considerable amount of variation in predicting word errors that depends on the length of the utterance. Interestingly, the variance of *Corpora* was 0.0718 which is a lot lower than that of *Utterance Length*. This indicates that there is a high level of consistency across corpora. Most importantly, the fixed effects capture twice as much variance as the random effects, highlighting the strong relationship between the fixed effects and the likelihood of word errors.

4.5.3 Conclusion

This section examined the effect of self-information on word errors. Two types of self-information were tested. The first type is based on the unconditional probability of a word, namely token frequency – $I(word)$. The second type is based on the conditional probability of a word given its preceding context, $I(word|context)$.

Fixed effects	Estimate	SE	z	$p(>\|z\|)$
(Intercept)	-1.7563	0.1867	-9.410	$< 2 \times 10^{-16***}$
$z[I(word)]$	1.1247	0.0394	28.540	$< 2 \times 10^{-16***}$
$z[I(word\|context)]$	0.2764	0.0315	8.786	$< 2 \times 10^{-16***}$
$z[Word\ Length]$	-0.4144	0.0269	-15.410	$< 2 \times 10^{-16***}$

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$

Random effects	Variance
Utterances (Intercept)	0.0475
Utterance Length (Intercept)	0.3775
Corpora (Intercept)	0.0718

Data size	N
Observations	19,840
Utterances	3,739
Utterance Length	26
Corpora	7

Table 4.17: Best logistic mixed-effects model: predicting word errors with self-information

Having controlled for potential confounds, such as word length and whether the word is a proper name, and allowed for variations in corpora and utterances, both types of self-information were still strong and significant predictors of whether a word will be misperceived in an utterance. The amount of self-information of a word is positively related to the likelihood of a word error. In other words, the less predicable a word is, the more likely it is that it will be perceived. This confirmed the processing cost account, in the respect that high self-information words have a higher processing cost. The findings also rejected the phonetic reduction account, due to the words being phonetically more reduced if they have low self-information, and as such phonetically reduced words are hardly to perceive.

4.6 Conclusion

The focus of this chapter was to examine the top-down lexical factors that play a role in naturalistic misperception. Top-down factors were tested from linguistic units of

various sizes – segments, syllables, words, and utterances – and the strength of their effects was evaluated.

Section 4.2 examined the effect of segmental frequency on two different aspects of segmental confusions. Firstly, the target bias and the response bias are strongly dependent on segmental frequency as confirmed by the strong to very strong correlations. This frequency bias is true for substitutions, insertions and deletions. In a segmental misperception, the probability of a given segment being the target (the intended segment) or the response (the perceived segment) is dependent on the probability of this segment occurring in the language, i.e. its frequency. The more frequent a segment is, the more likely it is that it will be the intended segment that gets misheard, and the perceived segment that is the result of a misperception.

Secondly, the difference in segmental frequency was found to be a significant predictor of the direction and strength of the asymmetrical confusions. For a given pair of segments, the confusion pattern tends to be in the direction of the more frequent segment, such that the less frequent segment is perceived as the more frequent segment more often than the reverse.

In Section 4.3, the three syllable factors – syllable constituency (onset, nucleus, and coda), syllable position (initial, medial, and final), and stress (unstressed and stressed) – were found to have a definite effect on the likelihood of a segment error. However, the effect of syllable constituency and that of stress are different between monosyllabic words and polysyllabic words. In monosyllabic words, coda is more erroneous than both onset and nucleus – Coda > [Onset, Nucleus]. This pattern is consistent with the findings from previous confusion experiments (Wang and Bilger, 1973; Redford and Diehl, 1999; Benkí, 2003).

In polysyllabic words, the constituency pattern is, however, different, with onset being more erroneous than both nucleus and coda – Onset > [Nucleus, Coda]. This is robust across syllable positions and stress conditions. I argued that this mis-

match between monosyllabic words and polysyllabic words is expected, since all the external evidence was based on monosyllables. One explanation was proposed using arguments of predictability and additional transitional cues. Segments become more predictable incrementally towards the end of a word, and this effect should be stronger for polysyllabic words than monosyllabic words, given the findings on uniqueness point (Luce, 1986a). Codas in word initial and medial syllables could have extra transitional cues from sonorant onsets on the right. Assuming that the true effect is the one found with monosyllabic words, Coda > [Onset, Nucleus], both of these factors could decrease the error rates of coda and nucleus and as a result onset becomes more erroneous.

For polysyllabic words, stressed syllables were less erroneous than unstressed syllables, which supports the idea that stressed syllables are "islands of reliability" (Pisoni, 1981). However, the effect was reversed for monosyllabic words, which can be explained in three different ways – a reporting bias, a differing definition of stress between monosyllabic and polysyllabic words, and lexical frequency.

Syllable position has a robust effect, in that word initial syllables are more erroneous than medial syllables, which in turn are more erroneous than final syllables. This effect is stronger in unstressed syllables and is attenuated in stressed syllables. This attenuation was argued to be a result of a ceiling effect caused by the high perceptual salience of stress overshadowing other factors.

Section 4.4 examined the frequency relationship between intended and perceived words. A strong and significant relationship was found between the frequency of the intended word and that of the perceived word – $Freq._{Perceived} = f(Freq._{Intended})$. This relationship remained robust after controlling for potential confounds, such as the choice of the corpus, the shape of the word (monosyllabicity), the removal of duplicates and the removal of pairs with a different number of syllables. This is consistent with the large-scale experimental confusion study by Felty et al. (2013).

The frequency of the perceived word and the frequency of the intended word are not significantly different in the combined corpus, and the pattern is inconsistent across subcorpora. This suggests that there is not a robust frequency difference – $Freq._{Perceived} \approx Freq._{Intended}$. In addition, the frequency relationship between the intended and perceived words cannot be reduced to that of the intended and perceived segments, as indicated by the weak correlation between the frequency of the intended segment and that of the perceived segment in substitution and the inconsistency between consonants and vowels — the frequency of the perceived segment is significantly higher than that of the intended segment only for consonants, and not for vowels.

Section 4.5 evaluated the effect of self-information on the likelihood of a word error in an utterance. After controlling for word length, the conditional and unconditional self-information of a word were both strong and significant predictors of word errors. High self-information words were more erroneous than low self-information words. This supports the processing cost account, which states that words that are harder to retrieve/process are more erroneous. The findings also suggest that listeners are sensitive to the conditional probability of a word, as opposed to only the unconditional probability (i.e. token frequency) when processing speech on an utterance level. This highlights the fact that in naturalistic speech perception, we do not process words in isolation, but in context. This casts doubt on the ecological validity of confusion studies which present words in isolation (Cooke, 2009; Felty et al., 2013; Tóth et al., 2015).

To conclude, the four sets of analyses in this chapter have demonstrated that naturalistic misperception is dependent on top-down factors from a range of linguistic units – segments, syllables, words, and utterances. This complements our findings in Chapter 3 in which phonological and phonetic factors were found. On the whole, we successfully replicated findings by Bird (1998), Browman (1978), Bond (1999),

Vitevitch (2002), and Tang and Nevins (2014) which used much smaller sets of naturalistic data. Crucially, this chapter presented, for the first time, a wide range of analyses of top-down factors using the largest naturalistic corpus of misperception.

Chapter 5

Conclusion

This chapter discusses limitations of this thesis and perspectives for future work. The limitations are considered for each analysis in Chapter 3 and Chapter 4, and potential solutions to overcome these limitations are proposed. Finally, I propose how the findings of this thesis can be integrated for examining the interactions between top-down and bottom-up factors, and how naturalistic misperception could be further examined cross-linguistically and cross-modally.

5.1 Accent

Recall in Chapter 2, Section 2.2.5 the choice of accent transcription was discussed. It was decided that both intended and perceived utterances are transcribed with the utterer's accent. There were two alternative choices: 1) transcribe both intended and perceived utterances with the perceiver's accent, and 2) transcribe both intended utterance with the utterer's accent and the perceived utterance with the perceiver's accent. The implication of the current choice is that the effect the interaction between the accents of the utterer and the perceiver has on misperceptions (specifically of the vowels) was ignored. In other words, the current analyses have underestimated the role of dialect on misperceptions.

The first solution is to create another set of transcriptions of the corpus by transcribing both the intended utterance with the utterer's accent and the perceived utterance with the perceiver's accent, even though this would overgenerate errors as mentioned in Chapter 2, Section 2.2.5.

The second solution is to create another set of transcriptions by transcribing both intended and perceived utterances phonemically. This is in fact how Labov (2010b) transcribed his own corpus. The accent interaction would have to be incorporated during the analyses, rather than in the transcriptions.

The third solution is to estimate the effect of accent interaction for each instance of misperception. This was done by Labov (2010b) on his own corpus. Labov (2010b) conducted an analysis of the relative contribution of five linguistic factors (lexicon, dialect, phonology, pragmatics and syntax). Using a tertiary scoring scheme, each instance of misperception was scored for whether each of these five factors was inhibiting, promoting or neutral to the misperception. Dialect came second as the promoter for 27% of the instances, after phonology. By applying the same scoring method to the entire mega corpus, each instance would have a score of the estimated effect of dialect (as a promoter or an inhibitor). The estimated dialect score can then be used to subset the mega corpus at different levels, and these subcorpora can then be cross examined to find out if the dialect/accent interaction has an impact on given analyses. One drawback is that the estimation process is subjective. Therefore, it should be done by multiple dialectologists or sociolinguists, with a set of well-defined criteria, in order to to evaluate the internal consistency of the estimated scores.

5.2 Vowel analyses

The way the vowels were analysed in this thesis is that each of the following IPA segments is a vowel: [e, ɛ, a, ɑ, ɒ, ʌ, ɔ, o, u, ʉ, ə, ɜ, ɪ, ʊ, æ, i].

The implication of this is that the offglides [j, w] were not considered with the nucleus portions of the vowels; and long vowels such as [ɑː] were treated as two vowels [ɑɑ].

It is not immediately clear what effect this would have on the vowel confusion analyses in this thesis. In the future, one could reparse the alignments and analyse the confusions in terms of the lexical sets e.g. FLEECE, KIT etc, which is effectively a phonemic analysis of vowel confusions. Such phonemic analyses of the vowels could be used to replicate the findings by Labov (1994b) and Labov (2010b) which supports the concepts of subsystems and to reanalyse the comparisons with experimental vowel confusions obtained by Cutler et al. (2004) in Chapter 3, Section 3.7.

5.3 Consonant analyses

Section 3.7 in Chapter 3 examined the ecological validity of experimental misperception data. Section 3.6 in Chapter 3 examined the amount of phonetic bias in naturalistic misperception. Both sets of analyses employed the agglomerative hierarchical clustering method (Rokach and Maimon, 2005) to examine the data on a structural level.

Agglomerative hierarchical clustering is a bottom-up approach. Every phone is assumed to belong to its own cluster. These clusters are then merged iteratively to form a hierarchy until there is only one cluster.

There is an alternative method of clustering, called divisive hierarchical clustering (Rokach and Maimon, 2005), which is a top-down approach. All phones are assumed to belong to one cluster. The cluster is then split iteratively to form a hierarchy until every phone belongs to its own cluster. In the future, the divisive hierarchical clustering method should also be tested to examine if the results are independent of the chosen approach.

In fact, the divisive hierarchical clustering method is similar to the successive division algorithm (SDA) of Dresher (2008). This algorithm is for specifying contrasts by a feature hierarchy. It can construct a hierarchical structure of phonemes using their feature specifications. It starts by assuming that all sounds form one phoneme, then the set is divided up into small sets with each selected distinctive feature, until every phoneme belongs to its own set. The output is a hierarchical tree.

In Chapter 3, Section 3.6, the phonetic bias of consonants was examined on a structural level. The featural based structure, which was compared to the perceptual based structure, was created by first computing the featural distances between phones using Frisch's similarity metric (Frisch, 1996; Frisch, Broe, and Pierrehumbert, 1997) (Section 3.6.1), before being projected as a hierarchical structure using the agglomerative hierarchical clustering method. In the future, SDA can be used as an alternative method to form the featural based structure.

5.4 Ecological validity

Section 3.7 in Chapter 3 compared naturalistic misperception data with experimental misperception data.

One limitation is that the phonological environments of the naturalistic confusions did not match those of the experimental confusions. For instance, Miller and Nicely (1955) tested 16 consonants embedded in a CV syllable with the vowel [ɑː]; therefore, the naturalistic confusions could be restricted to match the experimental environments. The complication is that there are many ways of matching the environments, e.g. 1) consonants that are in a (stressed) CV syllable with the vowel [ɑː], 2) word initial consonants that are in a (stressed) CV syllable with the vowel [ɑː], 3) word initial consonants that are in a (stressed) CV syllable with any open vowels, etc. A preliminary exploration of subsetting the context-free confusions by different

environments found that some subsets of the matrices were too sparse (i.e. too many zeros); therefore, there is a need for balancing the sparsity of a naturalistic matrix and the specificity of its environments when comparing naturalistic misperception data with experimental misperception data.

Alternatively, experimental data with less restricted environments may be better candidates for comparing with the naturalistic data. For instance, segmental confusions can be extracted from word level confusion data such as those obtained by Felty et al. (2013), or even from sentence level confusion data (i.e. present participants with sentences that are masked with noise).

One aspect of the experimental data that was considered in the thesis is that the confusions occurred with no pragmatic contexts. Specifically, they are not part of a conversation; therefore, listeners have no communicative need. Listeners were required to provide a response to what they thought they heard after being presented with a stimulus masked with noise. Confusion data that were obtained in a conversational setting should therefore be more ecologically valid and should yield higher correlation values with the naturalistic data. A potential source of such data is the Diapix corpus[1] (Hazan and Baker, 2011; Baker and Hazan, 2011). The corpus contains recordings of conversations between participants who engaged in pairs in 'spot the difference' picture tasks in both quiet and noise conditions. Having briefly examined the corpus, we can now identify confusion data from the conversations. For instance, participant *A* told participant *B*, "There is a *dog* at the bottom left of the picture". Participant *B* replied, "I cannot see a doll, do you mean a dog?". Therefore, it is possible to infer that participant *B* misheard *dog* as *doll*.

[1] I thank Michele Pettinato for suggesting this.

5.5 Segmental frequency

Section 4.2 in Chapter 4 examined the effect of segmental frequency on segmental confusions. However, only naturalistic data were analysed. It is yet to be known whether the findings can be extended to experimental data. One might even speculate that the effect of frequency would be stronger in experimental data because they are not affected by other top-down effects, e.g. word frequency. Therefore in the future similar analyses could be conducted on the experimental data described in Section 3.7.1, Chapter 3.

5.6 Syllable factor

Section 4.3 in Chapter 4 examined the effect of syllable constituency, syllable position, and stress. Four limitations are discussed below.

Firstly, in a similar analysis by Browman (1980) using the Browman data (a subset of the combined corpus), the author divided the segmental errors into two types – acoustic errors, and lexical errors. An acoustic error is defined as having one feature difference between the intended segment and the perceived segment, while a lexical error is defined as having multiple feature differences. The author argued that the errors with multiple feature differences are a reflection of a failure in lexical decision. However, this was not incorporated in the analyses in the thesis because such a distinction seems arbitrary. It is possible that a single feature difference is caused by a failure in lexical decision. However, it is equally possible that multiple feature differences are caused by acoustic misanalysis. Nonetheless, the author found that after separating the errors, a different pattern merged especially with the lexical errors. Therefore, in future analyses the number of feature differences can be incorporated by entering it as a random effect in our statistical model to capture the variation of the severity of the errors.

Secondly, our analyses found that segments in stressed monosyllables were more erroneous than those in unstressed monosyllables. This surprising result was attributed to a reporting bias in the naturalistic corpus, the differing definitions of stress between polysyllabic words and monosyllabic words and/or a lexical frequency effect.

One experiment could potentially test whether there is a reporting bias in reporting more stressed monosyllabic words than unstressed monosyllabic words. The reporting bias argument is that mishearing content words (stressed monosyllabic words) would disrupt communication more than mishearing function words (unstressed monosyllabic words), and therefore it is more noticeable when content words are misheard. Based on this idea of noticeability, an experiment[2] could be devised to test whether participants can notice more stressed monosyllabic words than unstressed monosyllablic words being misheard in a conversation. The stimuli would be dialogues with two interlocutors. During the conversation, one speaker would partially repeat what the other has just said but with one word (which could be either a content word or a function word) being different. For instance, *Speaker A* said to *Speaker B*, "So I'll bring my black **bag** with me tonight"; and *Speaker B* replied, "Why are you bringing a black **cat**?". The participants are asked to spot any instances of a misperception by one of the two interlocutors. If the participants are able to detect misperceptions of unstressed monosyllabic words and stressed monosyllabic words equally well, then the reporting bias explanation can be rejected.

Thirdly, the decrease in error rates across syllable constituents (Onset > [Nucleus, Coda]) and across syllable positions (Initial > Medial > Final) was argued to be the result of a predictability effect (Luce, 1986a). Concretely, the predictability of a segment increases with the number of preceding segments. To confirm this potential effect, one could create a language model on the level of segments, and the

[2]I thank Michael Becker for the idea of testing the noticeability of misperceptions experimentally.

conditional probability of a segment given its previous segments can be computed for all the segments in the corpus, and the segmental conditional probability predictor can then be entered in a mixed-effects model to see a) if it is a significant predictor and b) whether syllable constituency and syllable positions are still useful predictors after taking into account the segmental conditional probability.

Fourthly, the syllable constituency effect was different for monosyllables (Coda > [Nucleus, Onset]) and for polysyllables (Onset > [Nucleus, Coda]). In one of the explanations, I proposed that the true effect is the one with monosyllables, and the error rates of the nucleus and coda consonants get lowered by other factors. One of the factors is that the coda consonant in word initial and medial syllables could have extra transitional cues from a sonorant onset. The extra transitional cues would lower the error rates of the codas, such that they are even lower than the rates of the onsets. A further analysis can be done to examine this factor by removing the codas that are followed by a sonorant onset. Furthermore, the current analysis combined all the consonants in an onset cluster as onsets, and all the consonants in a coda cluster as codas. There are in fact more fine-grained positional effects within consonant clusters. In intervocalic C1C2 clusters, C1 is less prominent than C2, such that C1 is more likely to undergo phonological processes (such as place/voicing assimilation) than C2 (Jun, 2011). Further analyses should examine within-cluster positions of onsets and codas, and take into account their phonological environments.

5.7 Word frequency

Section 4.4 in Chapter 4 examined the relationship between the frequency of the intended word and that of the perceived word. Three limitations are discussed below.

Firstly, one of the confounds was the difference in word length between the in-

tended words and the perceived words. This was controlled for by matching the number of syllables in the intended word and that in the perceived word. However the number of segments was not controlled for. It is possible that two words have the same number of syllables, but one is longer than the other. Future analyses should incorporate the number of segments and the number of syllables of both the intended words and the perceived words. This could potentially account for the difference between polysyllabic words and monosyllabic words, such that monoysyllabic words correlated more strongly than polysyllabic words.

Secondly, word pairs that are the results of juncture errors were not considered, because the errors involved multiple words being misperceived as a single word and vice-versa. It is unclear which word amongst the multiple words should be chosen for comparing with the single word. One possible solution is to consider the frequency of the multiple word sequence. For instance, *how big is it?* is misperceived as *how bigoted*. The tri-gram frequency of *big is it* could be compared with the frequency of *bigoted*.

Third, the word frequency relationship was found to be strong and robust, while the segmental frequency relationship was found to be weak and inconsistent. How about the frequency relationship between the intended and perceived syllables, as opposed to words and segments? An analysis of syllable frequency could support the view that the syllable is a unit in perception (Bertoncini and Mehler, 1981; Cutler and Norris, 1988; Cutler and Butterfield, 1990) and listeners are sensitive to syllable frequency (see Carreiras, Alvarez, and Devega (1993), Perea and Carreiras (1998), and Conrad, Grainger, and Jacobs (2007) for a syllable frequency effect in visual word recognition). Furthermore, should we find a strong frequency relationship of syllables, then this analysis could potentially address the fact that the word frequency relationship is stronger for monosyllabic words than for polysyllabic words, because monosyllabic words are the same as single syllables, while polysyllabic words consist

of multiple syllables.

5.8 Interactions between top-down and bottom-up factors

While it is clear that bottom-up and top-down factors are at work as demonstrated in Chapter 3 and Chapter 4, the interaction between the two sets of factors remains unexplored in naturalistic misperception. Even distant linguistic levels such as pragmatics and phonetics are known to interact in perception (Rohde and Ettlinger, 2012). To bridge the gap between bottom-up and top-down factors, one approach would be to construct a word-recognition model, building on works by Marslen-Wilson and Welsh (1978) and Norris and McQueen (2008). In such a model, the lexical candidates are selected using the output of bottom-up acoustic analyses as well as top-down lexical constraints. The findings in the thesis can be entered as biases; for instance, the confusion matrices can be used to bias the acoustic analyses, and our lexical frequency findings can restrict the lexical candidates to those that are of similar frequency as the intended word. The performance of the model can serve as an indicator of the relative contribution of the factors by systematically including/excluding specific factors.

5.9 Beyond misperception of conversational speech

This thesis focused on naturalistic misperception data of conversational speech. As mentioned in Chapter 1, Section 1.2.5, Mondegreens and misperception of conversational speech differ in the production of language, the listening environment, the perception mechanism, and the available context. Given these differences, do listeners use the same strategies when listening to sung speech?

On the level of segmental confusions, Hirjee and Brown (2010) (as summarised in Chapter 1, Section 1.3.2) showed that the segmental confusion matrices derived from Mondegreens were able to capture our knowledge of perceptual similarity. Even though a direct comparison between Mondegreen and misperceptions of conversational speech has not been made, Hirjee and Brown's (2010) findings are encouraging.

On the syllable level, in misperception of conversational speech of English, the pattern of juncture misperceptions indicates how the listeners would use the structure of their language to form a strategy for speech segmentation. For English, strong syllables are most likely to be the beginning of a content word, whereas weak syllables are either non-initial syllables or function words (Cutler and Norris, 1988; Cutler and Butterfield, 1992). Given how function words can be stressed or lengthened in sung speech, it would be interesting to examine whether the listeners still use rhythmic cues to aid the speech segmentation process. In fact, this is supported by recent work on cross-linguistic Mondegreens (errors produced by a L2 listener) (Kentner, 2015).

Mondegreens can also be induced experimentally. Beck, Kardatzki, and Ethofer (2014) have identified two top-down factors that influence Mondegreens. They found that the severity of the misperception is a function of the perceived wittiness of the misperceived utterance and the vocabulary size of the listeners. It is worth examining whether these factors also play a role in misperceptions of conversational speech.

It is clear that a thorough comparison between Mondegreens and misperceptions of conversational speech is likely to yield promising findings on the cross-modal nature of misperception. I have compiled $\approx$ 130,000 instances of English Mondegreens, which will be compared in future research with the current corpus presented by this thesis.

5.10 Beyond misperception of English

As highlighted by Bond (1999, p. 134), it would be valuable to develop naturalistic misperception corpora for other languages besides English. A comparison between English and other languages would allow us to examine whether the findings with English are universal or language specific.

In fact, Mondegreen corpora are available for Japanese (Otake, 2007), German (Kentner, 2015), and possibly other languages. However, conversational data other than English are scarce. The closest data of a conversational nature is Voss's (1984) corpus of experimental German misperceptions of sentence level stimuli. For this reason, I have compiled $\approx$ 2,000 instances of conversational and Mondegreen data for Mandarin Chinese.

There are many aspects that are worth comparing between misperceptions of Mandarin Chinese and English, e.g. the difference between lexical tones in Mandarin Chinese and syllable stress in English. Section 4.3 in Chapter 4 found that unstressed syllables are more likely to be misperceived than stressed syllables. How about tones? Are level tones more likely to be misperceived than contour tones? How about strategies for word segmentation? Given that English listeners use stressed syllables to mark the beginning of a word, do Mandarin Chinese listeners use particular tones (e.g. neutral tone) to mark the beginning/end of a word? Using my developing corpus of Mandarin Chinese, these cross-linguistic questions can therefore be addressed.

Bibliography

Albright, Adam (2006). *Segmental similarity calculator*. http://web.mit.edu/albright/www/software/SimilarityCalculator.zip. (accessed 27 June 2015).

— (2007). "Gradient phonological acceptability as a grammatical effect". URL: http://web.mit.edu/albright/www/papers/Albright-GrammaticalGradience.pdf.

Albright, Adam and Bruce Hayes (2003). "Rules vs. analogy in English past tenses: A computational/experimental study". In: *Cognition* 90.2, pp. 119–161. DOI: 10.1016/S0010-0277(03)00146-X.

ANSI (1997). *S3. 5–1997. Methods for the calculation of the Speech Intelligibility Index*. New York: American National Standards Institute.

Asaridou, Salomi S. and James M. McQueen (2013). "Speech and music shape the listening brain: evidence for shared domain-general mechanisms". In: *Frontiers in Psychology* 4, p. 321. DOI: 10.3389/fpsyg.2013.00321.

Assmann, Peter and Quentin Summerfield (2004). "The Perception of Speech Under Adverse Conditions". English. In: *Speech Processing in the Auditory System*. Vol. 18. Springer Handbook of Auditory Research. Springer New York, pp. 231–308. ISBN: 978-0-387-00590-4. DOI: 10.1007/0-387-21575-1_5.

Attneave, Fred (1959). *Applications of Information Theory to Psychology: a summary of basic concepts, methods, and results*. New York: Holt, Rinehart & Winston.

Aylett, Matthew and Alice Turk (2004). "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech". In: *Language and Speech* 47.1, pp. 31–56. DOI: 10.1177/00238309040470010201.

Baayen, Harald R., Richard Piepenbrock, and Leon Gulikers (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Bailey, Todd M. and Ulrike Hahn (2001). "Determinants of wordlikeness: Phonotactics or lexical neighborhoods?" In: *Journal of Memory and Language* 44.4, pp. 568–591. DOI: 10.1006/jmla.2000.2756.

— (2005). "Phoneme similarity and confusability". In: *Journal of Memory and Language* 52.3, pp. 339–362. DOI: 10.1016/j.jml.2004.12.003.

Baker, Frank B. (1974). "Stability of two hierarchical grouping techniques Case I: Sensitivity to data errors". In: *Journal of the American Statistical Association* 69.346, pp. 440–445. DOI: 10.2307/2285675.

Baker, Rachel and Valerie Hazan (2011). "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs". In: *Behavior Research Methods* 43.3, pp. 761–770. DOI: 10.3758/s13428-011-0075-y.

Bartlett, Susan, Grzegorz Kondrak, and Colin Cherry (2009). "On the syllabification of phonemes". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 308–316.

Bartoń, Kamil (2014). *MuMIn: Multi-model inference*. R package version 1.10.0. URL: http://CRAN.R-project.org/package=MuMIn.

Bates, Douglas et al. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-6. URL: http://CRAN.R-project.org/package=lme4.

Beck, Claudia, Bernd Kardatzki, and Thomas Ethofer (2014). "Mondegreens and Soramimi as a Method to Induce Misperceptions of Speech Content Influence of Familiarity, Wittiness, and Language Competence". In: *PLoS ONE* 9.1, e84667. DOI: 10.1371/journal.pone.0084667.

Becker, Michael, Andrew Nevins, and Jonathan Levine (2012). "Asymmetries in generalizing alternations to and from initial syllables". In: *Language* 88.2, pp. 231–268.

Benkí, José R. (2003). "Analysis of English Nonsense Syllable Recognition in Noise". In: *Phonetica* 60.2, pp. 129–157. DOI: 10.1159/000071450.

Bertoncini, Josiane and Jacques Mehler (1981). "Syllables as units in infant speech perception". In: *Infant Behavior and Development* 4, pp. 247–260. DOI: 10.1016/S0163-6383(81)80027-6.

Bird, Helen (1998). "Slips of the ear as evidence for the postperceptual priority of grammaticality". In: *Linguistics* 36.3, pp. 469–516. DOI: 10.1515/ling.1998.36.3.469.

Black, Alan W., Richard Sproat, and Stanley Chen (2000). *Text normalization tools for the Festival speech synthesis system*. http://www.festvox.org/nsw/. (accessed 27 June 2015).

Black, Alan W. et al. (2002). *The Festival speech synthesis system (Version 1.4.2)*. 1.4. University of Edinburgh.

Blevins, Juliette (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Bond, Zinny S. (1999). *Slips of the ear: Errors in the perception of casual conversation*. New York: Academic Press.

Bond, Zinny S., Thomas J. Moore, and Beverley Gable (1996). "Listening in a second language". In: *ICSLP 96: Proceedings of the 4th International Conference on*

Spoken Language Processing. Vol. 4, pp. 2510–2513. DOI: 10.1109/ICSLP.1996.607323.

Botne, Robert and Stuart Davis (2000). "Language games, segment imposition, and the syllable". In: *Studies in Language* 24.2, pp. 319–344. DOI: 10.1075/sl.24.2.04bot.

Bradlow, Ann R. and David B. Pisoni (1999). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors". In: *The Journal of the Acoustical Society of America* 106.4, pp. 2074–2085. DOI: 10.1121/1.427952.

Broad, David J. and Ralph H. Fertig (1970). "Formant-Frequency Trajectories in Selected CVC-Syllable Nuclei". In: *The Journal of the Acoustical Society of America* 47.6B, pp. 1572–1582. DOI: 10.1121/1.1912090.

Broecke, M. P. R. van den and L. Goldstein (1980). "Consonant features in speech errors". In: *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*. Ed. by Victoria A. Fromkin. London: Academic Press.

Browman, Catherine P. (1978). "Tip of the tongue and slip of the ear: Implications for language processing". In: 42. UCLA Working Papers in Phonetics. University of California.

— (1980). "Perceptual processing: Evidence from slips of the ear". In: *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*. Ed. by V.A. Fromkin. New York: Academic Press, pp. 213–230.

Brown, Charles R. and Herbert Rubenstein (1961). "Test of Response Bias Explanation of Word-Frequency Effect". In: *Science* 133.3448, pp. 280–281. DOI: 10.1126/science.133.3448.280. URL: http://www.sciencemag.org/content/133/3448/280.abstract.

Brysbaert, Marc and Kevin Diependaele (2013). "Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based

choice". In: *Behavior Research Methods* 45.2, pp. 422–430. DOI: 10.3758/s13428-012-0270-5.

Brysbaert, Marc and Boris New (2009). "Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English". In: *Behavior Research Methods* 41.4, pp. 977–990. DOI: 10.3758/BRM.41.4.977.

Bybee, Joan (1995). "Regular morphology and the lexicon". In: *Language and Cognitive Processes* 10.5, pp. 425–455. DOI: 10.1080/01690969508407111.

— (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan L. and Paul J. Hopper (2001). *Frequency and the emergence of linguistic structure*. Vol. 45. Amsterdam: John Benjamins.

Calhoun, Sasha et al. (2010). "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue". In: *Language Resources and Evaluation* 44.4, pp. 387–419. ISSN: 1574-020X. DOI: 10.1007/s10579-010-9120-1.

Carhart, R., C. Johnson, and J. Goodman (1975). "Perceptual masking of spondees by combinations of talkers". In: *The Journal of the Acoustical Society of America* 58.S1, S35. DOI: 10.1121/1.2002082.

Carreiras, Manuel, Carlos J. Alvarez, and Manuel Devega (1993). "Syllable frequency and visual word recognition in Spanish". In: *Journal of Memory and Language* 32.6, pp. 766–780. DOI: 10.1006/jmla.1993.1038.

Celce-Murcia, Marianne (1980). "On Meringer's corpus of "slips of the ear"". In: *Errors of linguistics performance: Slips of the tongue, ear, pen, and hand*. Ed. by Victoria A. Fromkin. New York: Academic Press, pp. 199–212.

Chambers, Jack K. (2003). *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*. Language in Society. Oxford: Wiley-Blackwell. ISBN: 9780631228820.

Chang, Steve, Madelaine C. Plauché, and John J. Ohala (2001). "Markedness and consonant confusion asymmetries". In: *The Role of Speech Perception in Phonology*. Ed. by Elizabeth V. Hume and Keith Johnson. New York: Academic Press. Chap. 4, pp. 79–101.

Charrad, Malika et al. (2014). *NbClust: NbClust package for determining the best number of clusters*. R package version 2.0. URL: http://CRAN.R-project.org/package=NbClust.

Chen, Stanley F. and Joshua Goodman (1999). "An empirical study of smoothing techniques for language modeling". In: *Computer Speech & Language* 13.4, pp. 359–393. DOI: 10.1006/csla.1999.0128.

Cherry, E. Colin (1953). "Some Experiments on the Recognition of Speech, with One and with Two Ears". In: *The Journal of the Acoustical Society of America* 25.5, pp. 975–979. DOI: 10.1121/1.1907229.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Church, Kenneth Ward and Patrick Hanks (1990). "Word association norms, mutual information, and lexicography". In: *Computational Linguistics* 16.1, pp. 22–29.

Clark, Lynn and Graeme Trousdale (2009). "Exploring the role of token frequency in phonological change: Evidence from TH-Fronting in east-central Scotland". In: *English Language and Linguistics* 13.01, pp. 33–55. DOI: 10.1017/S1360674308002852.

Clarke, Frank R. (1957). "Constant-Ratio Rule for Confusion Matrices in Speech Communication". In: *The Journal of the Acoustical Society of America* 29.6, pp. 715–720. DOI: 10.1121/1.1909023.

Coetzee, Andries W. and Shigeto Kawahara (2013). "Frequency biases in phonological variation". English. In: *Natural Language & Linguistic Theory* 31.1, pp. 47–89.

ISSN: 0167-806X. DOI: 10.1007/s11049-012-9179-z. URL: http://dx.doi.org/10.1007/s11049-012-9179-z.

Conrad, Markus, Jonathan Grainger, and Arthur M. Jacobs (2007). "Phonology as the source of syllable frequency effects in visual word recognition: Evidence from French". English. In: *Memory & Cognition* 35.5, pp. 974–983. ISSN: 0090-502X. DOI: 10.3758/BF03193470.

Cooke, Martin (2006). "A glimpsing model of speech perception in noise". In: *The Journal of the Acoustical Society of America* 119, pp. 1562–1573. DOI: 10.1121/1.2166600.

— (2009). "Discovering consistent word confusions in noise". In: *Proceedings of Interspeech*. Brighton, UK, pp. 1887–1890.

Correlations: Direction and Strength. http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit4/correlationsdirectionandstrength/. (accessed 15 May 2015).

Cutler, Anne (1982). "The reliability of speech error data". In: *Slips of the tongue and language production*. Ed. by Anne Cutler. Amsterdam: Walter de Gruyter/Mouton, pp. 7–28.

— (1997). "The syllable's role in the segmentation of stress languages". In: *Language and Cognitive Processes* 12.5-6, pp. 839–846.

Cutler, Anne and Sally Butterfield (1990). "Syllabic lengthening as a word boundary cue". In: *Proceedings of the 3rd Australian International Conference on Speech Science and Technology*. Ed. by Roland Seidl. Canberra: Australian Speech Science and Technology Association, pp. 324–328.

— (1992). "Rhythmic cues to speech segmentation: Evidence from juncture misperception". In: *Journal of Memory and Language* 31.2, pp. 218–236. DOI: 10.1016/0749-596X(92)90012-M.

Cutler, Anne and Dennis Norris (1988). "The role of strong syllables in segmentation for lexical access". In: *Journal of Experimental Psychology: Human Perception and Performance* 14.1, pp. 113–121. DOI: 10.1037/0096-1523.14.1.113.

Cutler, Anne et al. (2000). "Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons". In: *Memory & Cognition* 28.5, pp. 746–755. DOI: 10.3758/BF03198409.

Cutler, Anne et al. (2004). "Patterns of English phoneme confusions by native and non-native listeners". In: *The Journal of the Acoustical Society of America* 116, pp. 3668–3678. DOI: 10.1121/1.1810292.

Daelemans, Walter and Antal van den Bosch (1992). "Generalization performance of backpropagation learning on a syllabification task". In: *TWLT3: Connectionism and Natural Language Processing*. Ed. by M. F. J. Drossaers and A. Nijholt. Enschede, The Netherlands, pp. 27–37.

Delattre, Pierre (1967). "Acoustic or articulatory invariance". In: *Glossa* 1.1, pp. 3–25.

Dimitropoulou, Maria et al. (2009). "Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior: The Case of Greek". In: *Frontiers in Psychology* 1, p. 218. DOI: 10.3389/fpsyg.2010.00218.

Dresher, B. Elan (2008). "The Contrastive Hierarchy in Phonology". In: *Contrast in Phonology: Theory, Perception, Acquisition*. Ed. by Peter Avery, B. Elan Dresher, and Keren Rice. Berlin: Mouton de Gruyter, pp. 11–33.

Durbin, Richard et al. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Eidhammer, Ingvar, Inge Jonassen, and William R. Taylor (2004). "Pairwise Global Alignment of Sequences". In: *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. Hoboken, NJ, USA.: John Wiley & Sons, Inc., pp. 3–34. DOI: 10.1002/9780470092620.ch1.

Ernestus, Mirjam and Anne Cutler (2014). "BALDEY: A database of auditory lexical decisions". In: *The Quarterly Journal of Experimental Psychology* (just-accepted), pp. 1–45.

Fabricius, Anne (2002). "Weak vowels in modern RP: An acoustic study of happY-tensing and KIT/schwa shift". In: *Language Variation and Change* 14.02, pp. 211–237. DOI: 10.1017/S0954394502142037.

Fant, Gunnar (1962). "Descriptive analysis of the acoustic aspects of speech". In: *LOGOS* 5.1, pp. 3–17.

Felty, Robert Albert et al. (2013). "Misperceptions of spoken words: Data from a random sample of American English words". In: *The Journal of the Acoustical Society of America* 134.1, pp. 572–585. DOI: 10.1121/1.4809540.

Ferber, Rosa (1991). "Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue". In: *Journal of Psycholinguistic Research* 20.2, pp. 105–122. DOI: 10.1007/BF01067878.

Festen, Joost M. and Reinier Plomp (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing". In: *The Journal of the Acoustical Society of America* 88.4, pp. 1725–1736. DOI: 10.1121/1.400247.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 363–370.

Fox, Chris and Roz Combley, eds. (2009). *Longman Dictionary of Contemporary English, Fifth Edition (Paperback + DVD-ROM)*. Harlow: Pearson Longman. ISBN: 1408215330.

French, Norman R. and John C. Steinberg (1947). "Factors Governing the Intelligibility of Speech Sounds". In: *The Journal of Acoustical Society of America* 19.1, pp. 90–119. DOI: 10.1121/1.1916407.

Freyman, Richard L., Uma Balakrishnan, and Karen S. Helfer (2004). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition". In: *The Journal of the Acoustical Society of America* 115.5, pp. 2246–2256. DOI: 10.1121/1.1689343.

Freyman, Richard L. et al. (1999). "The role of perceived spatial separation in the unmasking of speech". In: *The Journal of the Acoustical Society of America* 106.6, pp. 3578–3588. DOI: 10.1121/1.428211.

Frisch, Stefan (1996). "Similarity and frequency in phonology". PhD thesis. Northwestern University.

Frisch, Stefan, Michael Broe, and Janet Pierrehumbert (1997). "Similarity and phonotactics in Arabic". (accessed 27 June 2015). URL: http://roa.rutgers.edu/files/223-1097/roa-223-frisch-2.pdf.

Fromkin, Victoria A. (1973). *Speech errors as linguistic evidence*. 77. The Hague: Mouton.

— (2000). *Fromkin's Speech Error Database*. http://www.mpi.nl/resources/data/fromkins-speech-error-database/. (accessed 27 June 2015).

Fromkin, Victoria, Robert Rodman, and Nina Hyams (2003). *An Introduction to Language (9th edition, international edition)*. Boston: Wadsworth. ISBN: 1439082413.

Gale, William A. and Kenneth W. Church (1994). "What is wrong with adding one?" In: *Corpus-Based Research into Language*. Ed. by Nelleke Oostdijk and Pieter de Haan. Amsterdam: Rodopi, pp. 189–198.

Gale, William A. and Geoffrey Sampson (1995). "Good-turing frequency estimation without tears". In: *Journal of Quantitative Linguistics* 2.3, pp. 217–237. DOI: 10.1080/09296179508590051.

Galili, Tal (2014). *dendextendRcpp: Faster dendrogram manipulation using Rcpp*. R package version 0.5.1. URL: http://CRAN.R-project.org/package=dendextendRcpp.

Gambell, Timothy and Charles Yang (2005). "Word segmentation: Quick but not dirty". (accessed 27 June 2015). URL: http://www.ling.upenn.edu/~ycharles/papers/quick.pdf.

Garnes, Sara and Zinny S. Bond (1980). "A Slip of the Ear: A Snip of the Ear? A Slip of the Year". In: *Errors of linguistics performance: Slips of the tongue, ear, pen, and hand*. Ed. by Victoria A. Fromkin. New York: Academic Press, pp. 231–240.

Garofolo, John S. et al. (1993). *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. URL: http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

Geumann, Anja (2001). "Invariance and variability in articulation and acoustics of natural perturbed speech". PhD thesis. University Munich.

Goldsmith, John (1976). "Autosegmental Phonology". PhD thesis. Massachusetts Institute of Technology.

Goldwater, Sharon and Mark Johnson (2005). "Representational Bias in Unsupervised Learning of Syllable Structure". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ed. by Ido Dagan and Daniel Gildea. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 112–119. URL: http://www.aclweb.org/anthology/W/W05/W05-0615.

Good, Irving J. (1953). "The population frequencies of species and the estimation of population parameters". In: *Biometrika* 40.3-4, pp. 237–264. DOI: 10.1093/biomet/40.3-4.237.

Gordon, Reyna L. et al. (2010). "Words and Melody Are Intertwined in Perception of Sung Words: EEG and Behavioral Evidence". In: *PLoS ONE* 5.3, e9889. DOI: 10.1371/journal.pone.0009889.

Gorman, Kyle (2013). "Categorical and gradient aspects of wordlikeness". (accessed 27 June 2015). URL: http://www.csee.ogi.edu/~gormanky/papers/gorman-2012a.pdf.

Goslin, Jeremy and Ulrich H. Frauenfelder (2001). "A comparison of theoretical and human syllabification". In: *Language and Speech* 44.4, pp. 409–436. DOI: doi:10.1177/00238309010440040101.

Gower, John C. (1966). "Some distance properties of latent root and vector methods used in multivariate analysis". In: *Biometrika* 53.3-4, pp. 325–338. DOI: 10.1093/biomet/53.3-4.325.

Green, Tim, Andrew Faulkner, and Stuart Rosen (2012). "Variations in Carrier Pulse Rate and the Perception of Amplitude Modulation in Cochlear Implant Users". In: *Ear and Hearing* 33.2, pp. 221–230. DOI: 10.1097/AUD.0b013e318230fff8.

Grieser, Jessica (2010). "The Effect of Dialect Features on the Perception of Correctness in English-Word Voting Patterns on Forvo.com". In: *University of Pennsylvania Working Papers in Linguistics* 16.2, p. 10.

Guy, Gregory R. (1991). "Contextual conditioning in variable lexical phonology". In: *Language Variation and Change* 3.02, pp. 223–239. DOI: 10.1017/S0954394500000533.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pp. 1–8.

Halteren, Hans van (2000). "The Detection of Inconsistency in Manually Tagged Text". In: *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*. Ed. by Anne Abeille, Thorsten Brants, and Hans Uszkoreit. Centre Universitaire, Luxembourg: International Committee on Computational Linguistics, pp. 48–55.

Harrington, Jonathan (2010). *Phonetic Analysis of Speech Corpora*. Malden, MA: Wiley-Blackwell. ISBN: 1405199571, 9781405199575.

Harris, John (1994). *English Sound Structure*. Oxford: Wiley-Blackwell. ISBN: 0631187413.

— (2006). "The phonology of being understood: Further arguments against sonority". In: *Lingua* 116.10, pp. 1483–1494. DOI: 10.1016/j.lingua.2005.07.009.

— (2013). "Wide-domain r-effects in English". In: *Journal of Linguistics* 49.02, pp. 329–365. DOI: 10.1017/S0022226712000369.

Haswell, Richard H. (1988). "Error and Change in College Student Writing". In: *Written Communication* 5.4, pp. 479–499. DOI: 10.1177/0741088388005004005.

Hay, Jennifer, Janet Pierrehumbert, and Mary Beckman (2004). "Speech perception, well-formedness, and the statistics of the lexicon". In: *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Ed. by John Local, Richard Ogden, and Rosalind Temple. Cambridge: Cambridge University Press, pp. 58–74.

Hazan, Valerie and Rachel Baker (2011). "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions". In: *The Journal of the Acoustical Society of America* 130.4, pp. 2139–2152. DOI: 10.1121/1.3623753.

Hazan, Valerie, Souhila Messaoud-Galusi, and Stuart Rosen (2013). "The effect of talker and token variability on speech perception in noise in children with dyslexia". In: *Journal of Speech, Language, and Hearing Research*. DOI: 10.1044/1092-4388(2012/10-0107).

Heafield, Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch et al. Edinburgh, Scotland, United Kingdom, pp. 187–197. URL: http://kheafield.com/professional/avenue/kenlm.pdf.

Heeringa, Wilbert Jan (2004). "Measuring dialect pronunciation differences using Levenshtein distance". PhD thesis. Groningen.

Henikoff, S and J G Henikoff (1992). "Amino acid substitution matrices from protein blocks". In: *Proceedings of the National Academy of Sciences* 89.22, pp. 10915–10919.

Henton, Caroline G. (1983). "Changes in the vowels of Received Pronunciation". In: *Journal of Phonetics* 11, pp. 353–371.

Hillenbrand, James et al. (1995). "Acoustic characteristics of American English vowels". In: *The Journal of the Acoustical society of America* 97.5, pp. 3099–3111. DOI: 10.1121/1.411872.

Hirjee, Hussein and Daniel G. Brown (2010). "Solving misheard lyric search queries using a probabilistic model of speech sounds". In: *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Ed. by J. Stephen Downie and Remco C. Veltkamp. Utrecht, Netherlands: International Society for Music Information Retrieval, pp. 147–152.

Hodge, Milton H. (1967). "Some further tests of the constant-ratio rule". English. In: *Perception & Psychophysics* 2.10, pp. 429–437. ISSN: 0031-5117. DOI: 10.3758/BF03208790.

Hodge, Milton H. and Irwin Pollack (1962). "Confusion matrix analysis of single and multidimensional auditory displays". In: *Journal of Experimental Psychology* 63.2, pp. 129–142. DOI: 10.1037/h0042219.

Holube, Inga and Birger Kollmeier (1996). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model". In: *The Journal of the Acoustical Society of America* 100, pp. 1703–1716. DOI: 10.1121/1.417354.

Hook, Donald (1999). "The apostrophe: use and misuse". In: *English Today* 15.03, pp. 42–49. DOI: 10.1017/S026607840001110X.

Hooper, Joan B. (1972). "The syllable in phonological theory". In: *Language* 48.3, pp. 525–540. DOI: 10.2307/412031.

Horowitz, Leonard M., Margaret A. White, and Douglas W. Atwood (1968). "Word fragments as aids to recall: the organization of a word". In: *Journal of Experimental Psychology* 76.2, Pt. 1, pp. 219–226. DOI: 10.1037/h0025384.

Houston, Ann Celeste (1985). "Continuity and change in English morphology: the variable (ING)". PhD thesis. University of Pennsylvania.

Hsu, Bo-June (2009). *MIT Language Modeling Toolkit 0.4.1*. URL: http://code.google.com/p/mitlm/.

Hsu, Bo-June and James Glass (2008). "Iterative language model estimation: efficient data structure & algorithms". In: *Proceedings of Interspeech*. Vol. 8, pp. 1–4.

Hulst, Harry van der and Nancy Ritter (1999). *The syllable: views and facts*. Vol. 45. Berlin: Mouton de Gruyter.

International Electrotechnical Commission (2003). *Sound System Equipment – Part 16: Objective rating of speech intelligibility by speech transmission index*. IEC 60268–16.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Irwin, Patricia and Naomi Nagy (2007). "Bostonians/r/speaking: A quantitative look at (R) in Boston". In: *University of Pennsylvania Working Papers in Linguistics* 13.2, p. 11.

Iverson, Paul and Melanie Pinet (In press). *Individual differences in talker intelligibility in noise as a function of talker-listener accent similarity*.

Jakobson, Roman, Gunnar Fant, and Morris Halle (1952). *Preliminaries to speech analysis: The distinctive features and their correlates. Acoustics Laboratory*. Tech. rep. Massachusetts Institute of Technology.

Johnson, Keith (2012). *Acoustic and auditory phonetics*. Chichester: Wiley-Blackwell. ISBN: 1405194669.

Johnson, Paul C.D. (2014). "Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models". In: *Methods in Ecology and Evolution*, n/a–n/a. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12225. URL: http://dx.doi.org/10.1111/2041-210X.12225.

Jun, Jongho (2004). "Place assimilation". In: *Phonetically based phonology*. Ed. by Bruce Hayes, Robert M. Kirchner, and Donca Steriade. Cambridge: Cambridge University Press. Chap. 3, pp. 58–86.

— (2011). "Positional effects in consonant clusters". In: *The Blackwell Companion to Phonology, Wiley-Blackwell, Malden, MA*, pp. 1103–1123.

Jurafsky, Daniel and James H. Martin (2008). *Speech and Language Processing*. 2nd. Englewood Cliffs, New Jersey: Prentice Hall. ISBN: 0131873210.

Kahn, Daniel (1976). "Syllable-based generalizations in English phonology". PhD thesis. Massachusetts Institute of Technology.

Kehrein, Wolfgang (2002). *Phonological representation and phonetic phasing: Affricates and laryngeals*. Vol. 466. Tübingen: Max Niemeyer.

Kenstowicz, Michael J. (1994). *Phonology in Generative Grammar*. Cambridge, MA: Blackwell.

Kentner, Gerrit (2015). "Rhythmic segmentation in auditory illusions - evidence from cross-linguistic mondegreens". In: *Proceedings of 18th ICPhS*. Glasgow: ICPhS.

Kerswill, Paul (2001). "Mobility, meritocracy and dialect levelling: the fading (and phasing) out of Received Pronunciation". In: *British studies in the new millennium : the Challenge of the Grassroots. Proceedings of the 3rd Tartu Conference on British Studies*. Ed. by P. Rajamäe and K. Vogelberg. Tartu: University of Tartu.

Kerswill, Paul (2003). "Dialect levelling and geographical diffusion in British English". In: *Social dialectology: in honour of Peter Trudgill*, pp. 223–243.

— (2006). "RP, Standard English and the standard/non-standard relationship". In: *Language in the British Isles*. Ed. by David Britain. 2nd. Cambridge: Cambridge Unive.

Kerswill, Paul and Susan Wright (1990). "The validity of phonetic transcription: Limitations of a sociolinguistic research tool". In: *Language Variation and Change* 2.03, pp. 255–275. DOI: 10.1017/S0954394500000363.

Kessler, Brett (1995). "Computational dialectology in Irish Gaelic". In: *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*. Ed. by Steven P. Abney and Erhard W. Hinrichs. Association for Computational Linguistics. Dublin: Morgan Kaufmann, pp. 60–66.

Kessler, Brett and Rebecca Treiman (1997). "Syllable structure and the distribution of phonemes in English syllables". In: *Journal of Memory and Language* 37.3, pp. 295–311. DOI: 10.1006/jmla.1997.2522.

Keuleers, E. (2006). *Leanlex*. (accessed 27 June 2015). URL: http://crr.ugent.be/programs-data/leanlex.

Keuleers, E., M. Brysbaert, and B. New (2010). "SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles". In: *Behavior Research Methods* 42.3, pp. 643–650. DOI: 10.3758/BRM.42.3.643.

Keuleers, Emmanuel et al. (2012). "The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words". In: *Behavior Research Methods* 44.1, pp. 287–304. DOI: 10.3758/s13428-011-0118-4.

Kiparsky, Paul (2008). "Universals constrain change; change results in typological generalizations". In: *Linguistic universals and language change*, pp. 23–53.

Kirchner, Robert Martin (2001). *An effort based approach to consonant lenition*. London: Psychology Press.

Klatt, Dennis H. (1975). "Vowel lengthening is syntactically determined in a connected discourse". In: *Journal of Phonetics* 3.3, pp. 129–140.

Kneser, Reinhard and Hermann Ney (1995). "Improved backing-off for m-gram language modeling". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 1. Detroit, Michigan, USA: IEEE Computer Society, pp. 181–184. DOI: 10.1109/ICASSP.1995.479394.

Kollmeier, Birger, Thomas Brand, and Bernd Meyer (2008). "Perception of Speech and Sound". English. In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin Heidelberg: Springer, pp. 61–82. ISBN: 978-3-540-49125-5. DOI: 10.1007/978-3-540-49127-9_4.

Kondrak, Grzegorz (2003). "Phonetic Alignment and Similarity". In: *Computers and the Humanities* 37.3, pp. 273–291. DOI: 10.1023/A%3A1025071200644.

Kučera, Henry and Nelson Francis (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Kuperman, Victor and Raymond Bertram (2013). "Moving spaces: Spelling alternation in English noun-noun compounds". In: *Language and Cognitive Processes* 28.7, pp. 939–966. DOI: 10.1080/01690965.2012.701757.

Kyprianou, Marianna (2009). "The Phonetics and User-friendliness of Free Online Dictionaries: An Overview". In: *Selected Papers from the 19th International Symposium on Theoretical and Applied Linguistics (ISTAL 19)*. Ed. by Eliza Kitis et al. Thessaloniki, Greece: Aristotle University of Thessaloniki.

Labov, William (1989). "The exact description of a speech community: Short a in Philadelphia". In: *Language Change and Variation*. John Benjamins Publishing Company, pp. 1–57. DOI: 10.1075/cilt.52.02lab. URL: http://dx.doi.org/10.1075/cilt.52.02lab.

Labov, William (1994a). *Principles of Linguistic Change, Volume 1: Internal Factors*. Language in Society 20. Oxford: Wiley-Blackwell. ISBN: 0631179143.

— (1994b). *Principles of Linguistic Change, Volume 1: Internal Factors*. Language in Society 20. Oxford: Wiley-Blackwell. Chap. 9. ISBN: 0631179143.

— (2001). "The Philadelphia Vowel System". In: *Principles of Linguistic Change: Volume 2: Social Factors*. Oxford: Wiley-Blackwell. Chap. 4, pp. 121–145.

— (2010a). "A Controlled Experiment on Vowel Identification". In: *Principles of Linguistic Change: Volume 3: Cognitive and Cultural Factors*. Oxford: Wiley-Blackwell. Chap. 3, pp. 48–58. ISBN: 9781444327496.

— (2010b). "Natural Misunderstandings". In: *Principles of Linguistic Change: Volume 3: Cognitive and Cultural Factors*. Oxford: Wiley-Blackwell. Chap. 2, pp. 21–47.

— (2010c). "The Gating Experiments". In: *Principles of Linguistic Change: Volume 3: Cognitive and Cultural Factors*. Oxford: Wiley-Blackwell. Chap. 4, pp. 60–86. ISBN: 9781444327496.

Labov, William, Sharon Ash, and Charles Boberg (2005). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton/de Gruyter.

Ladefoged, Peter (1975). *A Course in Phonetics*. Harcourt Brace Jovanovich International editions. Harcourt Brace Jovanovich : New York. ISBN: 9780155151802.

Laferriere, Martha (1977). "Boston short a: Social Variation as Historical Residue". In: *Studies in Language Variation: Semantics, Syntax, Phonology, Pragmatics, Social Situations, Ethnographic Approaches*, pp. 100–107.

Lahiri, Aditi and Henning Reetz (2002). "Underspecified recognition". In: *Laboratory Phonology 7*. Ed. by Carlos Gussenhoven and Natasha Werner. Berlin: Mouton de Gruyter, pp. 637–676.

Laver, John (1970). "The production of speech". In: *New Horizons in Linguistics*. Ed. by John Lyons. Harmondsworth: Penguin, pp. 53–75.

— (1994). *Principles of phonetics*. Cambridge University Press.

Lehiste, Ilse and Gordon E. Peterson (1961). "Transitions, glides, and diphthongs". In: *The Journal of the Acoustical Society of America* 33.3, pp. 268–277. DOI: 10.1121/1.1908638.

Levenshtein, Vladimir Iosifovich (1966). "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics Doklady*. Vol. 10. 8, pp. 707–710.

Levitt, Andrea et al. (1988). "Context effects in two-month-old infants' perception of labiodental/interdental fricative contrasts". In: *Journal of Experimental Psychology: Human Perception and Performance* 14.3, p. 361.

Levy, Roger (2008). "Expectation-based syntactic comprehension". In: *Cognition* 106.3, pp. 1126–1177. DOI: 10.1016/j.cognition.2007.05.006.

Liberman, Isabelle Y. et al. (1974). "Explicit syllable and phoneme segmentation in the young child". In: *Journal of Experimental Child Psychology* 18.2, pp. 201–212. DOI: 10.1016/0022-0965(74)90101-5.

Lidestam, B., J. Holgersson, and S. Moradi (2014). "Comparison of informational vs. energetic masking effects on speechreading performance". In: *Frontiers in Psychology* 5, p. 639. DOI: 10.3389/fpsyg.2014.00639.

Lindblom, Björn E. F. (1968). "Temporal organization of syllable production". In: *Quarterly Progress and Status Report*. Vol. 9. 2–3. Stockholm: Royal Institute of Technology, pp. 1–5.

Lindsey, Geoffrey A. (2012a). *Smoothing, then and now*. http://englishspeechservices.com/blog/smoothing-then-and-now/. Blog.

— (2012b). *The British English vowel system*. http://englishspeechservices.com/blog/british-vowels/. Blog.

Lindsey, Geoffrey A. and Péter Szigetvári (2014). *Current British English searchable transcriptions*. http://seas3.elte.hu/cube/. (accessed 24 November 2014).

Lombardi, Linda (2002). "Coronal epenthesis and markedness". In: *Phonology* 19.2, pp. 219–252. DOI: 10.1017/S0952675702004323.

Luce, Paul A. (1986a). "A computational analysis of uniqueness points in auditory word recognition". In: *Perception & Psychophysics* 39.3, pp. 155–158. DOI: 10.3758/BF03212485.

— (1986b). "Neighborhoods of Words in the Mental Lexicon". PhD thesis. Department of Psychology, Indiana University, Bloomington, Indiana.

Luce, Paul A. and David B. Pisoni (1998). "Recognizing spoken words: The neighborhood activation model". In: *Ear and Hearing* 19.1, pp. 1–36.

Luce, R. Duncan (1963). "Detection and Recognition". In: *Handbook of Mathematical Psychology*. Ed. by R. Duncan Luce, Robert R. Bush, and Eugene Galanter. Vol. 1. New York: John Wiley & Sons. Chap. 3, pp. 103–189.

Macmillan, Neil A. and C. Douglas Creelman (2004). *Detection theory: A user's guide*. 2nd. Hove: Psychology Press.

Macmillan, Neil A. and Howard L. Kaplan (1985). "Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates". In: *Psychological bulletin* 98.1, pp. 185–199. DOI: 10.1037/0033-2909.98.1.185.

MacQueen, James B. (1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by Lucien Marie Le Cam and Jerzy Neyman. Vol. 1. 281–297. Berkeley, California, USA: University of California Press, p. 14.

Maechler, Martin et al. (2013). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source).

Mandera, Paweł et al. (2014). "Subtlex-pl: subtitle-based word frequency estimates for Polish". In: *Behavior Research Methods*, pp. 1–13. DOI: 10.3758/s13428-014-0489-4.

Mantel, Nathan (1967). "The detection of disease clustering and a generalized regression approach". In: *Cancer Research* 27.2 Part 1, pp. 209–220.

Marchand, Yannick, Connie R. Adsett, and Robert I. Damper (2009). "Automatic Syllabification in English: A Comparison of Different Algorithms". English. In: *Language and Speech* 52.1, pp. 1–27. DOI: 10.1177/0023830908099881.

Marian, Viorica et al. (2012). "CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities". In: *PLoS ONE* 7.8, e43230. DOI: 10.1371/journal.pone.0043230.

Marslen-Wilson, William D. and Alan Welsh (1978). "Processing interactions and lexical access during word recognition in continuous speech". In: *Cognitive Psychology* 10.1, pp. 29–63. DOI: 10.1016/0010-0285(78)90018-X.

McCarthy, John J. (1979). "On stress and syllabification". In: *Linguistic Inquiry*, pp. 443–465.

McCarthy, John J (1994). "The phonetics and phonology of Semitic pharyngeals". In: *Phonological structure and phonetic form: Papers in Laboratory Phonology III*. Ed. by Patricia A. Keating. Cambridge University Press.

McClelland, James L. and Jeffrey L. Elman (1986). "The TRACE model of speech perception". In: *Cognitive Psychology* 18.1, pp. 1–86. DOI: 10.1016/0010-0285(86)90015-0.

McClelland, James L., David E. Rumelhart, and Geoffrey E. Hinton (1986). "The appeal of parallel distributed processing". In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Ed. by David E. Rumelhart, James L. McClelland, and The PDP Research Group. Vol. 1. London: MIT Press, pp. 3–44.

McLennan, Conor T. and Paul A. Luce (2005). "Examining the time course of indexical specificity effects in spoken word recognition". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.2, p. 306. DOI: 10.1037/0278-7393.31.2.306.

McMillan, James B (1946). "Phonology of the standard English of east central Alabama". PhD thesis. University of Chicago.

Meringer, Rudolf (1908). *Aus dem Leben der Sprache; Versprechen, Kindersprache, Nachahmungstrieb*. Berlin: B. Behr.

Meringer, Rudolf and Carl Mayer (1895). *Versprechen Und Verlesen: Eine Psychologisch-Linguistische Studie*. Stuttgart: G. J. Goschen.

Metropolis, Nicholas and S. Ulam (1949). "The Monte Carlo method". In: *Journal of the American Statistical Association* 44.247, pp. 335–341.

Miller, George A. and Patricia E. Nicely (1955). "An analysis of perceptual confusions among some English consonants". In: *The Journal of the Acoustical Society of America* 27, pp. 338–352. DOI: 10.1121/1.1907526.

Miller, Virginia Rogers (1953). "Present-day use of the broad a in Eastern Massachusetts". In: *Communications Monographs* 20.4, pp. 235–246. DOI: 10.1080/03637755309375090.

Mirenda, Pat and David Beukelman (1987). "A comparison of speech synthesis intelligibility with listeners from three age groups". In: *Augmentative and Alternative Communication* 3.3, pp. 120–128. DOI: 10.1080/07434618712331274399.

Moisl, Hermann (2007). "Data nonlinearity in exploratory multivariate analysis of language corpora". In: *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*. Association for Computational Linguistics, pp. 93–100.

Munson, Benjamin et al. (2003). "Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability". In: *The*

Journal of the Acoustical Society of America 113.2, pp. 925–935. DOI: 10.1121/1.1536630.

Nábělek, Anna K. (1988). "Identification of vowels in quiet, noise, and reverberation: Relationships with age and hearing loss". In: *The Journal of the Acoustical Society of America* 84.2, pp. 476–484. DOI: 10.1121/1.396880.

Nagata, Masaaki (1998). "Japanese OCR error correction using character shape similarity and statistical language model". In: *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pp. 922–928.

Nagy, Naomi and Julie Roberts (2004). "New England: phonology". In: *A handbook of varieties of English* 1, pp. 270–81.

Nakagawa, Shinichi and Holger Schielzeth (2013). "A general and simple method for obtaining R2 from generalized linear mixed-effects models". In: *Methods in Ecology and Evolution* 4.2, pp. 133–142. DOI: 10.1111/j.2041-210x.2012.00261.x.

Needleman, Saul B. and Christian D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3, pp. 443–453. DOI: 10.1016/0022-2836(70)90057-4.

Nespor, Marina, Marcela Peña, and Jacques Mehler (2003). "On the different roles of vowels and consonants in speech processing and language acquisition". In: *Lingue e linguaggio* 2.2, pp. 203–230.

New, Boris et al. (2007). "The use of film subtitles to estimate word frequencies". In: *Applied Psycholinguistics* 28.4, pp. 661–677. DOI: 10.1017/s014271640707035x.

Norris, Dennis (1994). "Shortlist: a connectionist model of continuous speech recognition". In: *Cognition* 52.3, pp. 189–234. DOI: 10.1016/0010-0277(94)90043-4.

Norris, Dennis and James M. McQueen (2008). "Shortlist B: A Bayesian model of continuous speech recognition". In: *Psychological Review* 115.2, pp. 357–395. DOI: 10.1037/0033-295x.115.2.357. URL: http://dx.doi.org/10.1037/0033-295X.115.2.357.

Nusbaum, Howard C, David B Pisoni, and Christopher K Davis (1984). "Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words". In: *Research on Speech Perception Progress Report* 10.10, pp. 357–376.

Ohala, Diane K. (1999). "The influence of sonority on children's cluster reductions". In: *Journal of Communication Disorders* 32.6, pp. 397–422. DOI: 10.1016/S0021-9924(99)00018-0.

Ohala, J.J. (1981). "The listener as a source of sound change". In: *Papers from the Parasession on Language and Behavior*. Ed. by C.S. Masek, R.A. Hendrick, and M.F. Miller. Chicago: Chicago Linguistic Society, pp. 178–203.

Ohala, John J. (1989). "Sound change is drawn from a pool of synchronic variation". In: *Language change: Contributions to the study of its causes*. Ed. by Leiv Egil Breivik and Ernst Håkon Jahr. Berlin: Mouton, pp. 173–198.

Ohala, John J. and Elizabeth Shriberg (1990). "Hypercorrection in speech perception". In: *The First International Conference on Spoken Language Processing, ICSLP 1990, Kobe, Japan, November 18-22, 1990*. URL: http://www.isca-speech.org/archive/icslp_1990/i90_0405.html.

Oksanen, Jari et al. (2013). *vegan: Community Ecology Package*. R package version 2.0-10. URL: http://CRAN.R-project.org/package=vegan.

O'Shaughnessy, Douglas (1987). *Speech Communications: Human and Machine (Addison-Wesley Series in Electrical Engineering)*. Reading, MA: Addison-Wesley. ISBN: 0201165201.

Otake, Takashi (2007). "Interlingual near homophonic words and phrases in L2 listening: Evidence from misheard song lyrics". In: *Proceedings of the 16th Inter-*

national Congress of Phonetic Sciences (ICPhS 2007). Ed. by Trouvain. J. and W. J. Barry. Saarbrücken: ICPhS, pp. 777–780.

Padgett, Jaye (2002). "Feature classes in phonology". In: *Language* 78.1, pp. 81–110.

Parker, Stephen G. (2002). "Quantifying the sonority hierarchy". PhD thesis. University of Massachusetts Amherst.

Perea, Manuel and Manuel Carreiras (1998). "Effects of syllable frequency and syllable neighborhood frequency in visual word recognition". In: *Journal of Experimental Psychology: Human perception and performance* 24.1, pp. 134–144. DOI: 10.1037/0096-1523.24.1.134.

Peterson, Gordon E. and Harold L. Barney (1952). "Control methods used in a study of the vowels". In: *The Journal of the Acoustical Society of America* 24.2, pp. 175–184. DOI: 10.1121/1.1906875.

Peterson, Gordon E. and Ilse Lehiste (1960). "Duration of syllable nuclei in English". In: *The Journal of the Acoustical Society of America* 32.6, pp. 693–703. DOI: 10.1121/1.1908183.

Phatak, Sandeep A. and Jont B. Allen (2007). "Consonant and vowel confusions in speech-weighted noise". In: *The Journal of the Acoustical Society of America* 121.4, pp. 2312–2326. DOI: 10.1121/1.2642397.

Phatak, Sandeep A., Andrew Lovitt, and Jont B. Allen (2008). "Consonant confusions in white noise". In: *The Journal of the Acoustical Society of America* 124, pp. 1220–1233. DOI: 10.1121/1.2913251.

Pierrehumbert, Janet (2006). "Syllable structure and word structure: a study of tricomonantal clusters in English". In: *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology* 3, pp. 168–188.

Pinet, M., P. Iverson, and P. Evans (2011). "Perceptual Adaptation for L1 and L2 Accents in Noise by Monolingual British English Listeners". In: *The 17th International Congress of Phonetic Sciences*, pp. 1602–1605.

Pinet, Melanie and Paul Iverson (2010). "Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience". In: *The Journal of the Acoustical Society of America* 128.3, pp. 1357–1365.

Pinet, Melanie, Paul Iverson, and Mark Huckvale (2011). "Second-language experience and speech-in-noise recognition: Effects of talker–listener accent similarity". In: *The Journal of the Acoustical Society of America* 130, p. 1653.

Pisoni, David B. (1981). "Some current theoretical issues in speech perception". In: *Cognition* 10.1, pp. 249–259. DOI: 10.1016/0010-0277(81)90054-8.

Plag, Ingo (2006). "The variability of compound stress in English: structural, semantic, and analogical factors". In: *English Language and Linguistics* 10.01, pp. 143–172. DOI: 10.1017/S1360674306001821.

Plauché, Madelaine, Cristina Delogu, and John J. Ohala (1997). "Asymmetries in consonant confusion". In: *Eurospeech*. DOI: 10.1121/1.417051.

Pollack, Irwin (1975). "Auditory informational masking". In: *The Journal of the Acoustical Society of America* 57.S1, S5–S5. DOI: 10.1121/1.1995329.

Pollack, Irwin, Herbert Rubenstein, and Louis Decker (1960). "Analysis of incorrect responses to an unknown message set". In: *The Journal of the Acoustical Society of America* 32.4, pp. 454–457. DOI: 10.1121/1.1908097.

Prichard, Hilary and Meredith Tamminga (2012). "The impact of higher education on Philadelphia vowels". In: *University of Pennsylvania Working Papers in Linguistics* 18.2, p. 11.

Przedlacka, Joanna (2001). "Estuary English and RP: Some recent findings". In: *Studia Anglica Posnaniensia* 36, pp. 35–50.

Pulgram, Ernst (1970). *Syllable, Word, Nexus, Cursus*. 81-85. The Hague: Mouton.

Quirk, Randolph et al. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Redford, Melissa A. and Randy L. Diehl (1999). "The relative perceptual distinctiveness of initial and final consonants in CVC syllables". In: *The Journal of the Acoustical Society of America* 106, pp. 1555–1565. DOI: 10.1121/1.427152.

Régnier, Marion S and Jont B Allen (2008). "A method to identify noise-robust perceptual features: Application for consonant/t/". In: *The Journal of the Acoustical Society of America* 123, pp. 2801–2814. DOI: 10.1121/1.2897915.

Rhebergen, Koenraad S., Niek J. Versfeld, and Wouter A. Dreschler (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise". In: *The Journal of the Acoustical Society of America* 120, pp. 3988–3997. DOI: 10.1121/1.2358008.

Rogerson, Peter (2001). *Statistical methods for geography*. London: Sage.

Rohde, Hannah and Marc Ettlinger (2012). "Integration of pragmatic and phonetic cues in spoken word recognition." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38.4, p. 967.

Rokach, Lior and Oded Maimon (2005). "Clustering methods". In: *Data mining and knowledge discovery handbook*. Springer, pp. 321–352. DOI: 10.1007/0-387-25465-X_15.

Rosen, Stuart, Alan Adlard, and Heather K. J. van der Lely (2009). "Backward and simultaneous masking in children with grammatical specific language impairment: no simple link between auditory and language abilities". In: *Journal of Speech, Language and Hearing Research* 52.2, pp. 396–411. DOI: doi:10.1044/1092-4388(2009/08-0114).

Schleef, Erik and Michael Ramsammy (2013). "Labiodental fronting of/θ/in London and Edinburgh: a cross-dialectal study". In: *English Language and Linguistics* 17.01, pp. 25–54.

Schütze, Carson T. and Victor S. Ferreira (2007). "What Should We Do With Our Speech Error Corpora? Notes from the Panel Discussion". In: *MIT Working Papers in Linguistics* 53, pp. 383–393.

Selkirk, Elizabeth O. (1984). "On the major class features and syllable theory". In: *Language Sound Structure: Studies in Phonology*. Ed. by Mark Aronoff and Richard T. Oehrle. Cambridge: MIT Press, pp. 107–136.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Shattuck-Hufnagel, Stefanie and Dennis H. Klatt (1979). "The limited use of distinctive features and markedness in speech production: Evidence from speech error data". In: *Journal of Verbal Learning and Verbal Behavior* 18.1, pp. 41–55. DOI: 10.1016/S0022-5371(79)90554-1.

Shepard, Roger N. (1958). "Stimulus and response generalization: Deduction of the generalization gradient from a trace model". In: *Psychological Review* 65.4, pp. 242–256.

— (1972). "Psychological representation of speech sounds". In: *Human Communication: A Unified View*. Ed. by Edward E. David and Peter B. Denes. New York, pp. 67–113.

— (1987). "Toward a universal law of generalization for psychological science". In: *Science* 237.4820, pp. 1317–1323. DOI: 10.1126/science.3629243.

Shuyo, Nakatani (2010). *Language Detection Library for Java*. (accessed 27 June 2015). URL: http://code.google.com/p/language-detection/.

Simpson, Sarah A. and Martin Cooke (2005). "Consonant identification in N-talker babble is a nonmonotonic function of N". In: *Journal of the Acoustical Society of America* 118.5, pp. 2775–2778. DOI: 10.1121/1.2062650.

Singh, Riya and Jont B. Allen (2012). "The influence of stop consonants' perceptual features on the Articulation Index model". In: *The Journal of the Acoustical Society of America* 131.4, pp. 3051–3068. DOI: 10.1121/1.3682054.

Smith, Nathaniel J. and Roger Levy (2008). "Optimal processing times in reading: a formal model and empirical investigation". In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Ed. by Brad C. Love, Ken McRae, and Vladimir M. Sloutsky. Austin, TX: Cognitive Science Society, pp. 595–600.

Smolensky, Paul (1993). "Harmony, markedness, and phonological activity". In: *Paper presented at Rutgers Optimality Workshop 1*. Rutgers University (ROA-87).

Sneller, Betsy (2014). "Antagonistic Contact and Inverse Affiliation: Appropriation of/TH/-fronting by White Speakers in South Philadelphia". In: *University of Pennsylvania Working Papers in Linguistics* 20.2, pp. 169–178.

Sokal, Robert R. and F. James Rohlf (1962). "The Comparison of Dendrograms by Objective Methods". In: *Taxon*, pp. 33–40. DOI: 10.2307/1217208.

Sproat, Richard et al. (2001). "Normalization of non-standard words". In: *Computer Speech & Language* 15.3, pp. 287–333. DOI: 10.1006/csla.2001.0169.

Steeneken, Herman J. M. and Tammo Houtgast (1980). "A physical method for measuring speech-transmission quality". In: *The Journal of the Acoustical Society of America* 67, p. 318. DOI: 10.1121/1.384464.

Steriade, Donca (2001). "The phonology of perceptibility effects: the P-map and its consequences for constraint organization". In: *Ms., UCLA*.

Stevens, Kenneth N. (1972). "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data". In: *Human Communication: A Unified View*. Ed. by Edward E. David and Peter B. Denes. New York, pp. 51–56.

Stevens, Kenneth N. (2002). "Toward a model for lexical access based on acoustic landmarks and distinctive features". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1872–1891. DOI: 10.1121/1.1458026.

Stevens, Stanley Smith, John Volkmann, and Edwin B. Newman (1937). "A scale for the measurement of the psychological magnitude pitch". In: *The Journal of the Acoustical Society of America* 8.3, pp. 185–190. DOI: 10.1121/1.1915893.

Stuart-Smith, Jane (2005). "Is TV a contributory factor in accent change in adolescents". In: *Final Report to the ESRC. Grant No. R000239757*.

— (2007). "The influence of the media". In: *The Routledge Companion to Sociolinguistics. London: Routledge*, pp. 140–148.

Stuart-Smith, Jane and Claire Timmins (2006). "'Tell her to shut her moof': the role of the lexicon in TH-fronting in Glaswegian". In: *The Power of Words: Essays in Lexicography, Lexicology and Semantics: In Honour of Christian J. Kay*. Ed. by C. J. Kay et al. Costerus: new series 163. Amsterdam, Netherlands: Rodopi, pp. 171–183. URL: http://eprints.gla.ac.uk/8990/.

Sundberg, Johan (1970). "Formant structure and articulation of spoken and sung vowels". In: *Folia Phoniatrica et Logopaedica* 22.1, pp. 28–48.

Surendran, Dinoj and Gina-Anne Levow (2004). "The functional load of tone in Mandarin is as high as that of vowels". In: *Proceedings of Speech Prosody 2004*. Nara, Japan, pp. 99–102.

Tang, K. (2012). "A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research". In: *UCL Working Papers in Linguistics* 24, pp. 208–214.

Tang, Kevin and Andrew Nevins (2014). "Measuring Segmental and Lexical Trends in a Corpus of Naturalistic Speech". In: *Proceedings of the 43rd Meeting of the North East Linguistic Society*. Ed. by Hsin-Lun Huang, Ethan Poole, and Amanda Rysling. Vol. 2. GLSA (Graduate Linguistics Student Association), pp. 153–166.

Thomas, Charles Kenneth (1958). *Introduction to the Phonetics of American English*. New York: The Ronald Press Company.

— (1961). "The phonology of new England English". In: *Communications Monographs* 28.4, pp. 223–232. DOI: 10.1080/03637756109375321.

Toscano, Joseph C. and Jont B Allen (2014). "Across-and Within-Consonant Errors for Isolated Syllables in Noise". In: *Journal of Speech, Language, and Hearing Research* 57.6, pp. 2293–2307. DOI: 10.1044/2014_JSLHR-H-13-0244.

Tóth, Máté Attila et al. (2015). "A corpus of noise-induced word misperceptions for Spanish". In: *J. Acoust. Soc. Am.* 137.2, EL184–EL189. DOI: 10.1121/1.4905877. URL: http://dx.doi.org/10.1121/1.4905877.

Traunmüller, Hartmut (1990). "Analytical expressions for the tonotopic sensory scale". In: *The Journal of the Acoustical Society of America* 88.1, pp. 97–100. DOI: 10.1121/1.399849.

Turk, Alice E. and Stefanie Shattuck-Hufnagel (2007). "Multiple targets of phrase-final lengthening in American English words". In: *Journal of Phonetics* 35.4, pp. 445–472. DOI: 10.1016/j.wocn.2006.12.001.

Valentine, Tim, Tim Brennen, and Serge Brédart (1996). *The Cognitive Psychology of Proper Names*. London: Routledge. ISBN: 0415135451.

van Heuven, Walter J. B. et al. (2014). "SUBTLEX-UK: A new and improved word frequency database for British English". In: *The Quarterly Journal of Experimental Psychology* 67.6, pp. 1176–1190. DOI: 10.1080/17470218.2013.850521.

van Son, Rob J. J. H. and Jan P. H. van Santen (2005). "Duration and spectral balance of intervocalic consonants: A case for efficient communication". In: *Speech Communication* 47.1, pp. 100–123. DOI: 10.1016/j.specom.2005.06.005.

van Vugt, Floris, Bruce Hayes, and Kie Zuraw (2012). *Pheatures Spreadsheet*. http://www.linguistics.ucla.edu/people/hayes/120a/Pheatures/.

Vihman, Marilyn May (1996). *Phonological development: The origins of language in the child*. Oxford: Blackwell.

Vitevitch, Michael S. (2002). "Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear". In: *Language and Speech* 45.4, pp. 407–434. DOI: 10.1177/00238309020450040501.

Voss, Bernd (1984). *Slips of the ear: investigations into the speech perception behaviour of German speakers of English*. Tübingen: Narr.

Wagner, Suzanne Evans (2008). "Language change and stabilization in the transition from adolescence to adulthood". In: *Philadelphia, PA: University of Pennsylvania dissertation*.

Wales, Katie (1994). "Royalese: the rise and fall of The Queen's English". In: *English Today* 10.03, pp. 3–10.

Wang, Marilyn D. and Robert C. Bilger (1973). "Consonant confusions in noise: A study of perceptual features". In: *The Journal of the Acoustical Society of America* 54.5, pp. 1248–1266. DOI: 10.1121/1.1914417.

Warren, Richard M. (1970). "Perceptual restoration of missing speech sounds". In: *Science* 167.3917, pp. 392–393. DOI: 10.1126/science.167.3917.392.

Watson, Charles S., William J. Kelly, and Henry W. Wroton (1976). "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty". In: *The Journal of the Acoustical Society of America* 60.5, pp. 1176–1186. DOI: 10.1121/1.381220.

Weide, R. L. (2014). *The Carnegie Mellon Pronouncing Dictionary [cmudict. 0.7a]*. URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Wells, John C. (1982a). *Accents of English. 3 vols*.

— (1982b). *Accents of English: Volume 1: An Introduction*. Cambridge University Press. ISBN: 0521297192.

Wells, John C. (1982c). *Accents of English: Volume 2: The British Isles*. Cambridge University Press. ISBN: 0521285402.

— (1982d). *Accents of English: Volume 3: Beyond the British Isles*. Cambridge University Press. ISBN: 0521285410.

— (1990). "Syllabification and allophony". In: *Studies in the pronunciation of English: a commemorative volume in honour of AC Gimson. London: Routledge*, pp. 76–86.

— (1996). "Why phonetic transcription is important". In: *The Journal of the Phonetic Society of Korea* 31–32, pp. 239–242.

— (2006). *English Intonation PB and Audio CD: An Introduction*. Cambridge University Press. ISBN: 0521683807.

— (2008). *Longman Pronunciation Dictionary, Paper with CD-ROM (3^{rd} Edition)*. Harlow: Pearson Longman. ISBN: 1405881186.

— (2009). *More about e and ɛ*. `http://phonetic-blog.blogspot.co.uk/2009/03/more-about-e-and.html`. Blog.

— (2010). *fronted GOOSE*. `http://phonetic-blog.blogspot.co.uk/2010/06/fronted-goose.html`. Blog.

— (2011). *adding stress [written in IPA]*. `http://phonetic-blog.blogspot.com/2011/06/wenzdiz-bl-n-fnetk-trnskrpn-simz-tu-v.html`. Blog.

Wieling, Martijn, Eliza Margaretha, and John Nerbonne (2012). "Inducing a measure of phonetic similarity from pronunciation variation". In: *Journal of Phonetics* 40.2, pp. 307–314.

Wieling, Martijn and John Nerbonne (2011). "Measuring linguistic variation commensurably". In: *Dialectologia: revista electrònica*, pp. 141–162.

Wieling, Martijn, Jelena Prokić, and John Nerbonne (2009). "Evaluating the pairwise string alignment of pronunciations". In: *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences,*

Humanities, and Education. Association for Computational Linguistics, pp. 26–34.

Wieling, Martijn et al. (2014). "Measuring Foreign Accent Strength in English: Validating Levenshtein Distance as a Measure". In: *Language Dynamics and Change* 4.2, pp. 253–269. DOI: 10.1163/22105832-00402001.

Wiener, F. M. and George A. Miller (1946). "Some characteristics of human speech". In: *Transmission and reception of sounds under combat conditions. Summary Technical Report of Division 17*. Washington, DC: National Research Committee, pp. 58–68.

Wingfield, Arthur et al. (1985). "Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time". In: *Journal of Gerontology* 40.5, pp. 579–585.

Witten, Ian H. and Timothy Bell (1991). "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression". In: *IEEE Transactions on Information Theory* 37.4, pp. 1085–1094. DOI: 10.1109/18.87000.

Wood, Elizabeth (2003). "TH-fronting: The Substitution of f/v for T/D in New Zealand English". In: *New Zealand English Journal* 17, pp. 50–56.

Wright, Charles E. (1979). "Duration differences between rare and common words and their implications for the interpretation of word frequency effects". In: *Memory & Cognition* 7.6, pp. 411–419. DOI: 10.3758/BF03198257.

Wright, Richard (2004). "A review of perceptual cues and cue robustness". In: *Phonetically based phonology*. Ed. by Bruce Hayes, Robert M. Kirchner, and Donca Steriade. Cambridge: Cambridge University Press. Chap. 3, pp. 34–57.

Wright, Sylvia (1954). "The Death of Lady Mondegreen". In: *Harpers Magazine* 209.1254, pp. 48–51.

Xu, Yi (2010). "In defense of lab speech". In: *Journal of Phonetics* 38.3, pp. 329–336. DOI: 10.1016/j.wocn.2010.04.003.

Yang, James H. (2010). “Phonetic evidence for the nasal coda shift in Mandarin”. In: *Taiwan Journal of Linguistics* 8.1, pp. 29–55.

Yavaş, Mehmet (2011). *Applied English Phonology*. Malden, MA: Blackwell.

Yilmaz, Hüseyin (1967). “A theory of speech perception”. In: *Bulletin of Mathematical Biophysics* 29.4, pp. 793–825.

— (1968). “A theory of speech perception: II”. In: *Bulletin of Mathematical Biophysics* 30.4, pp. 455–479.

Yu, Alan C. L. (2013). “Individual differences in socio-cognitive processing and the actuation of sound change”. In: *Origins of sound change: Approaches to phonologization*. Oxford: Oxford University Press, pp. 201–227.

Yu, Alan C. L. et al. (2011). “Effects of working memory capacity and autistic traits on phonotactic effects in speech perception”. In: *Proceedings of the International Congress of the Phonetic Sciences XVII, Hong Kong: International Congress of the Phonetic Sciences*, pp. 2236–2239.

Zatorre, Robert J. and Shari R. Baum (2012). “Musical melody and speech intonation: Singing a different tune”. In: *PLoS Biology* 10.7, e1001372. DOI: 10.1371/journal.pbio.1001372.