# Understanding forced alignment errors in Hindi-English code-mixed speech – a feature analysis

*Ayushi Pandey*[1]*, *Pamir Gogoi*[2]*, *Kevin Tang*[2]

[1]Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]Department of Linguistics, University of Florida, Gainesville, FL, U.S.A., 32611-5454

pandeya@tcd.ie, pgogoi@ufl.edu, tang.kevin@ufl.edu

## Abstract

Forced alignment methods have recently seen great progress in the fields of acoustic-phonetics studies of low-resource languages. Code-mixed speech however, presents complex challenges to forced-alignment techniques, because of the longer phonemic inventory of bilingual speakers, the nature of accented speech, and the confounding interaction of two languages at a frame level. In this paper, we use the Montreal Forced Aligner to annotate the Phonetically Balanced Code-Mixed read-speech corpus (7.4 hours; 113 speakers) in 3 different training environments (code-mixed, Hindi and English). Additionally, we present an analysis of alignment errors using phonological and data-driven features using Random Forest and Linear mixed effects models. We find that contextual influence of neighbouring phonemes influences the error in alignment most significantly, when compared against any other features. Many of the alignment errors by phonological features can be explained by their acoustic distinctiveness. Additionally, the amount of training data by phone type also contributed to lowering their respective error rates.

**Index Terms**: code-mixing, code-switching, forced alignment, error analysis, speech recognition

## 1. Introduction

Forced alignment techniques have become increasingly popular in the analysis and description of speech data. After successful performance of forced aligners on large scale monolingual speech resources [1, 2, 3], a variety of non-traditional speech resources [4, 5, 6] have also been explored in the past decade. However, code-mixed speech may present a challenge for automatic speech alignment, given the complex interaction of two languages at the frame level within the same utterance.

Code-mixed speech is frequently observed in bilingual and multilingual communities all across the world. Speakers exhibit alternation either within the utterance (code-mixing) or at a clausal or phrasal level (code-switching). In light of its relevance and complexity, techniques on processing [7, 8, 9, 10] code-mixed speech have actively been explored in the past few years. Until a few years ago, prevalent practices in the domain have included two-pass approaches; first detecting language and then recognizing using appropriate acoustic models. At the same time, adapting monolingual speech resources have also seen active investment [11, 12]. More recently, dedicated neural network architectures [13, 14, 8], augmentation of existing datasets [10] and several fine-grained problems [15, 16] in speech recognition are being unearthed, underscoring the relevance of these studies even further.

With the rise in computational speech processing, theoretical questions surrounding this phenomena may also arise, as would a discussion on their tools and techniques. In this paper, we present a step in this direction. We compare the performance of Montreal Forced Aligner [1] against gold-standard annotations over isolated English words from the Phonetically Balanced Code Mixed corpus of Hindi-English read speech [17]. We present results in three word-level training environments; a) with pooled Hindi and English data, b) with monolingual Hindi data, and c) within corpus monolingual English data. Finally, using Random Forests and Linear Mixed Effects models, we present an analysis of forced alignment errors, against a set of phonological and data-driven predictors.

The organization of the paper is as follows: Section 2 describes the acoustic data, the pronunciation lexicon, and the gold-standard annotation procedure. Section 3 details the experimental procedure, with descriptions of the training datasets for alignment, and the analytical models for evaluation. Section 4 discusses the results of the analysis, and presents comparison between each type of experiment over different models. Section 5 concludes the paper.

## 2. Data

### 2.1. Acoustic data

The Phonetically Balanced Code Mixed (PBCM) corpus [17] contains 6,941[1] phonetically balanced sentences, recorded at IIIT-Hyderabad. Sentences for this corpus were compiled using an optimal selection procedure from selected sections of a prominent Hindi newspaper, Dainik Bhaskar. The speech data for this corpus was recorded by 113 native Hindi speakers (58 male, and 55 female), all of whom were fluent in English.

| | Hindi | English |
|---|---|---|
| word (types) | 4790 | 3754 |
| word (tokens) | 54961 | 18839 |
| phoneme (types) | 73 | 52 |
| phoneme (tokens) | 194672 | 97137 |

Table 1: *Distribution of Hindi-English words and phonemes in the PBCM corpus.*

### 2.2. Pronunciation data

The PBCM corpus was originally transcribed using the Wx notation [18], which is a popular transcription metric for Indian languages, especially for NLP and related purposes. For conducting acoustic-phonetic studies however, we compared the pronunciations generated by Espeak, Epitran and Wx. We

*authors of equal contribution

[1] more phonetically balanced sentences were added after the original publication, which reports 6,126 utterances

found Espeak (http://espeak.sourceforge.net/) to be the best pronunciation scheme, particularly for cases where pronunciation was not predictable by orthography. Some errors, however, still persisted. Errors of nasalization and syllabification (irregular schwa insertion) were corrected through a combination of manual edits and phonological rules that ensured homorganic nasal-consonant clusters.

These corrections could still not accommodate the speaker-specific variation between the **/ʤ-z/** and **/p$^{h}$-f/** contexts, which were not always distinct in the orthography. Manual inspection revealed very little speaker variation in the /p$^{h}$-f/ context, with most speakers preferring the /f/ variant. But for the /ʤ-z/ context, speakers were found to compensate for the orthographic inconsistency, resulting in variant pronunciations of the ʤ words. To overcome the problem of lesser represented phonemes, bootstrapping techniques are prevalent in the forced alignment literature [5, 6, 19, 20, 21]. In such models, target phonemes in the lexicon are mapped to more frequent and closely resembling phonemes. Therefore, we decided to create several bootstrapped versions of each of the variants (in /ʤ-z/ and /p$^{h}$-f/) lexicon, and allowed the Montreal Forced Aligner [1] to pick the most appropriate variant. The pronunciation selected by the best performing bootstrapping model was chosen to be listed in the lexicon. Thus, we created a variant free lexicon where the /p$^{h}$-f/ variation was mapped entirely to /f/ and each pronunciation of the /ʤ/ words were given a unique identity.

### 2.3. Gold-standard alignment data

After the development of a variant-free lexicon, each word in the dataset was manually given a) Hindi, b) English and c) part-Hindi tags, by two Hindi speakers (one fluent, one native). Part-Hindi tags referred to switches between Hindi and English at a morphological level [22] (e.g, *amerik-i*, for American). Such words were removed from the analysis. Among the English words, 10% of the words were selected for gold-standard annotation. To maintain consistent speaker variation, 6 sentences from each speaker's set of sentences were chosen. Word and phoneme level boundaries for each of these words were manually annotated and cross-evaluated by the two Hindi speakers. Forced alignment results obtained from each training environment (discussed Section 3.1) will be analyzed against the words in this dataset.

## 3. Experimental setup

This section describes the experimental procedure for the analysis of force aligned English words. First, we discuss the 3 training environments using each of which, the isolated English words were aligned. We then introduce the phonological and data driven features that were used as predictors for the evaluation of the forced alignment.

### 3.1. Acoustic models

As a preliminary step, the complete PBCM corpus was used as training as well as alignment data. This initial alignment resulted in word and phoneme boundaries for every sentence in the corpus. Using this dataset, and the TextGridTools [23] package, we separated the data into word level chunks. Subsets of this data were created in the following way:

- **Code-mixed Words (CoM-W):** All the extracted words from the aligned sentences were used as training environment for this experiment. This included Hindi words, English words, as well as those English words that carried Hindi inflection (for example: "amerik-i"). The purpose of this sub-experiment was to maximize coverage for each phoneme within the training dataset, and pool Hindi and English word-level data for the alignment.
- **Hindi Words (Hin-W)** From the aforementioned word-level dataset, monolingual Hindi words were separated and a new dataset was created. The purpose of this sub-experiment was to assess the reliability of forced alignment in absence of any English data. This type of setup also probes the question of the accuracy of alignment when all the phonemes are well represented, but the phonotactic information of English words is withdrawn.
- **English Words**: In the same vein as the previous setup (Hin-W), monolingual English words separated from the CoM-W dataset, and an English-only model was created. The purpose of this setup was to evaluate the role of monolingual English data in identical training and alignment environments.

Each of these datasets was then sent to align the isolated English words dataset (Eng-W trained and aligned on itself), using the speaker-adapted triphone model. From the obtained forced alignment timestamps, Midpoint (the central timestamp between left and right boundary) for each phoneme were extracted. Similarly, this Midpoint value was extracted for the gold-standard annotations as well. The absolute difference between the Gold-standard Midpoint and the Forced-Aligner Midpoint will serve as the main dependent variable for each of the subsequent analyses.

### 3.2. Predictors

In this subsection, we discuss each of the features (both phonological, and data-driven) that were used as predictors to evaluate the accuracy of the forced alignment. To consistently maintain the number of features per phoneme, all phonemes in the corpus were uniquely specified for each of these features.

#### 3.2.1. Consonant-specific feature set

The following set of features were specified for each consonant in the inventory:

- Manner of articulation
- Place of articulation
- Voicing type

Existing literature on error evaluation of speech recognition models and forced aligners have shown that some of these features are recognised better than others [24], [6]. Mel-frequency cepstral coefficients (MFCCs) are the standard acoustic speech signal representations in speech recognition, and indeed in forced alignment models. [24] compared that the performance of an MFCC-based recognition system with an articulatory based system trained on German speech, some articulatory features were suboptimally encoded by MFCCs, such as labial, coronal, dental, palatal, velar, fricative, –round, high, back and –voice. Similarly, [6] finds that a cross-lingual forced alignment of non-English speech using English models performed better on natural classes of stops, fricatives, nasals and affricates. These results supported our motivation for separating our consonants and vowels into sets of natural classes of place, manner and voicing features. In general acoustic terms, consonant specific features, fricatives and stops were hypothesized to be well aligned given their prominent acoustic characteristics. For instance, the boundaries of fricatives are marked by a section of

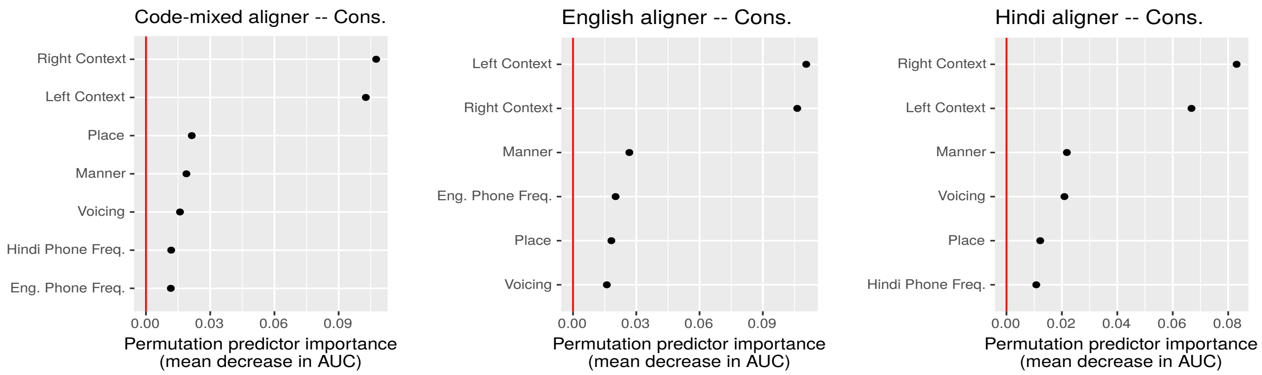

Figure 1: *Variance importance for predicting Error$_{consonant}$ using consonant specific features across the CoM-W, Hin-W and Eng-W training environments*

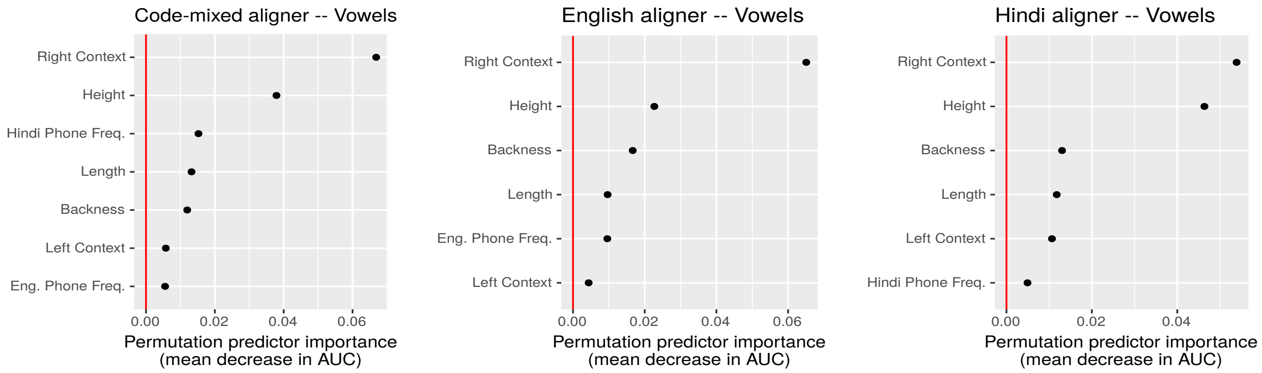

Figure 2: *Variance importance for predicting Error$_{vowel}$ using consonant specific features across the CoM-W, Hin-W and Eng-W training environments*

noise and those of stops are marked by a clear burst in most cases, suggesting better alignments, while approximants poorly so, given their dynamic transitions into adjacent vowels.

### *3.2.2. Vowel-specific feature set*

The following set of features were specified for each vowel in the inventory:

- Vowel height
- Vowel frontness/backness
- Vowel length

Among these, height and backness were expected to influence the alignment quality more significantly, especially with high and back vowels given the findings by [24].

### *3.2.3. Data driven feature set*

The following set of features were specified for each phoneme in the inventory, regardless of its segment type (consonant or vowel). However, the phone frequency are relevant only selectively for particular training environments. For example, the frequency of English phone is not relevant for training environments with Hindi monolingual words.

- Right boundary
- Left boundary
- Phone frequency in Hindi Words
- Phone frequency in English Words

Here, the right and left boundary indicated whether the target phone is surrounded by a vowel, a consonant or is a boundary phone. Due to coarticulatory effects, and the triphone modeling inherent in the Montreal Forced Aligner, the influence of the neighbouring environment was hypothesized to be strong on the alignment quality. Similarly, it was important to analyze the error in terms of the frequency of the phone in the Training environment.

## 3.3. Statistical models

This subsection gives an overview of the statistical models developed for use in the later part of the analysis. We use two statistical models: Random Forests, and Linear mixed effects model to analyze the Error under each training environments.

### *3.3.1. Random forest*

Random forests have emerged as powerful tools in estimating the importance of individual features in the prediction of a dependent variable. To minimize the effect of speaker variation on alignment errors, the predicted variable Error was passed through a per speaker z-score normalization. Then, using the Party package [25, 26] in *R*, a random forest model was used to analyze Error for each of the vowel and consonant type. The entire set of *relevant* features in each case was used as predictor variables for each type of segment (vowel or consonant) and experiment. Therefore, for example:

$$\begin{aligned}Error_{vowel} \sim Height &+ Backness + Roundedness \\ &+ Left.Context + Right.Context \\ &+ Eng.Ph.Freq + Hin.Ph.Freq \\ &+ Length\end{aligned}$$

Here, equation 1 represents the model equation for Error in vowels in the CoM-W context. All the vowel specific features have been specified in the model, in addition to the global features. Their individual influence is described in the Results section, and can be seen in Figure 2.

### *3.3.2. Linear mixed effects model*

Random forest and linear mixed effects models are complementary statistical methods [27]. While random forest will provide the *overall* importance of the variables of interest (factorial or continuous), mixed-effects model will be used to highlight how these variables affect the amount of alignment errors, while being able to capture random-effect factors such as the speakers and words in the sample. The model structure will be largely the same as 3.3.1 but with the addition of by-speaker and by-word random intercepts.

$$\begin{aligned}Error_{vowel} \sim Height &+ Backness + Roundedness \\ &+ Left.Context + RightContext \\ &+ Eng.Ph.Freq + Hin.Ph.Freq \\ &+ Length + (1|Speaker) \\ &+ (1|Word)\end{aligned}$$

## 4. Results

This section presents the results obtained from comparing the Error computed using the absolute difference in Midpoints per phoneme, from the gold-standard annotations and the forced-aligned annotations. Error thus obtained will be modeled as a *predicted* variable using RandomForest and Linear Mixed Effects modeling. RandomForest and will be presented for vowels and consonants separately.

### 4.1. Descriptive Statistics

Table 2 displays the error tolerance levels (in msec) for each of the training environments (CoM-W, Hin-W and Eng-W). This means that, when trained on code-mixed word level data, 52.58% phoneme boundaries matched gold-standard annotations within <10 ms. A comparison across rows (different training models) shows that CoM-W was the best performing model, with the greatest coverage of phonemes in the <10 ms range.

Table 2: *Comparison of tolerance (in msec) of the three models*

| | **Tolerance (msec)** | | | | |
|---|---|---|---|---|---|
| | **<10** | **<20** | **<30** | **<40** | **<50** |
| **CoM-W** | 52.58 | 83.00 | 94.54 | 97.33 | 98.94 |
| **HIN-W** | 50.75 | 80.37 | 92.56 | 96.14 | 98.16 |
| **ENG-W** | 51.89 | 82.71 | 94.12 | 97.27 | 98.86 |

### 4.2. Random Forest

Figures 1 and 2 display the relative importance of each feature for predicting error. The deviance from the red axis indicates the relative strength of each feature in predicting the error in alignment. Across all the training environments (CoM-W, Hin-W, Eng-W), the Right Context and the Left Context appear to be a lot more influential in the error prediction, compared with the phonological features of the target phone. This reflects the nature of consonants being encoded partly in surrounding phones especially vowels, e.g. stops. Since about a third (34%) of consonants in the corpus are stop consonants, which rely on transitional cues, it is possible that co-articulation may be governing this observation. Consistent patterns like these are not quite so clear among other features. The representation of the target phone in the training corpus is significant, but once again, does not influence the error as much. This indicates that simply increasing the individual presence of a phoneme may not be helpful enough, unless its supporting context is present. The vocalic context appears to be largely influenced by its Right boundaries, but not as much as by its left boundaries. This suggests that the vowels are better supported by their left boundaries, likely from an onset consonant, compared to a coda consonant, a well-established phonetic universal. Let us explore these effects in a more granular fashion in the next subsection.

### 4.3. Linear Mixed-Effects Regression

Using the lmerTest [28] package in *R*, we predicted the Error variable using fully specified models for vowels and consonants separately. Similar to the observations obtained in Random Forest model, we found a significant effect of surrounding phones ($p < 0.001$) on the alignment error in the consonantal context. For each of the training environments, non-boundary phones have higher error than boundary phones. Reverse trends were observed in [6], where non-boundary phones showed better alignment. However, the comparison with our results is not straightforward: because the analysis in [6] was conducted using monophone based aligners (HMAlign and P2FA), and Montreal Forced Align is a speaker-adapted triphone acoustic model. Through our LMER analysis, we found increased error for *consonants* due to both their left and right context, but for *vowels*, but b) vowels appear to be influenced largely by their Right boundaries across all training environments. While the case for consonants is not so clear, there is evidence that the onset (left boundary) cues for the vowel are perceptually more significant, than offset (right boundary) cues [29]. The presence of aspiration causes the formant transitions to be more stable across the initial and steady-state portion of the vowels [30, 31], perceptually supporting the recognition of the vowel.

## 5. Conclusion

In this paper, we conducted forced alignment on code-mixed speech from the PBCM Hindi-English read speech corpus. We created 3 types of acoustic models, code-mixed words (CoM-W), Hindi words (Hin-W) and English words (Eng-W) from the PBCM corpus. A variant-free Hindi-accented lexicon used was consistent across all the training datasets. We found that despite having only half the number of phonemes in the training corpus, the monolingual English word model performs better than the monolingual Hindi word model. This suggests that despite having increased phoneme representation, we may not achieve better alignment quality, if phonotactic information is absent. Similarly, in the RandomForest and linear mixed effect model analysis, we found that contextual information was most significant in influencing the errors of alignment.

# 6. References

[1] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017, pp. 498–502.

[2] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "Fave (forced alignment and vowel extraction) program suite," *URL http://fave. ling. upenn. edu*, 2011.

[3] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[4] T. Knowles, M. Clayards, M. Sonderegger, M. Wagner, A. Nadig, and K. H. Onishi, "Automatic forced alignment on child speech: Directions for improvement," in *Proceedings of Meetings on Acoustics 170ASA*, vol. 25, no. 1. Acoustical Society of America, 2015, p. 060001.

[5] K. Tang and R. Bennett, "Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (Mayan)," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 1719–1723. [Online]. Available: https://assta.org/proceedings/ICPhS2019/papers/ICPhS_1768.pdf

[6] C. DiCanio, H. Nam, D. H. Whalen, H. Timothy Bunnell, J. D. Amith, and R. C. García, "Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2235–2246, 2013. [Online]. Available: https://doi.org/10.1121/1.4816491

[7] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint arXiv:1904.00784*, 2019.

[8] S. Sivasankaran, B. M. L. Srivastava, S. Sitaram, K. Bali, and M. Choudhury, "Phone merging for code-switched speech recognition," 2018.

[9] M. Choudhury, K. Bali, S. Sitaram, and A. Baheti, "Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks," in *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 2017, pp. 65–74.

[10] E. Yılmaz, H. v. d. Heuvel, and D. A. van Leeuwen, "Acoustic and textual data augmentation for improved asr of code-switching speech," *arXiv preprint arXiv:1807.10945*, 2018.

[11] K. Bhuvanagirir and S. K. Kopparapu, "Mixed language speech recognition without explicit identification of language," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 92–97, 2012.

[12] A. Pandey, B. M. L. Srivastava, and S. V. Gangashetty, "Adapting monolingual resources for code-mixed Hindi-English speech recognition," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 218–221.

[13] E. Yilmaz, H. v. d. Heuvel, and D. A. van Leeuwen, "Investigating bilingual deep neural networks for automatic speech recognition of code-switching frisian speech," 2016.

[14] A. Biswas, F. de Wet, E. van der Westhuizen, E. Yilmaz, and T. Niesler, "Multilingual neural network acoustic modelling for asr of under-resourced english-isizulu code-switched speech." in *Interspeech*, 2018, pp. 2603–2607.

[15] B. M. L. Srivastava and S. Sitaram, "Homophone identification and merging for code-switched speech recognition." in *Interspeech*, 2018, pp. 1943–1947.

[16] S. Rallabandi, S. Sitaram, and A. W. Black, "Automatic detection of code-switching style from acoustics," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 2018, pp. 76–81.

[17] A. Pandey, B. M. L. Srivastava, R. Kumar, B. T. Nellore, K. S. Teja, and S. V. Gangashetty, "Phonetically balanced code-mixed speech corpus for Hindi-English automatic speech recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[18] V. Chaitanya, R. Sangal, and A. Bharati, *Natural language processing: a Paninian perspective*. Prentice-Hall of India, 1996.

[19] L. M. Johnson, M. Di Paolo, and A. Bell, "Forced alignment for understudied language varieties: Testing Prosodylab-Aligner with Tongan data," *Language Documentation & Conservation*, vol. 12, pp. 80–123, 2018.

[20] T. Kempton, "Cross-language forced alignment to assist community-based linguistics for low resource languages," March 6–7 2017, paper presented at ComputEL-2, Honolulu.

[21] E. Kurtic, B. Wells, G. J. Brown, T. Kempton, and A. Aker, "A corpus of spontaneous multi-party conversation in Bosnian Serbo-Croatian and British English," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[22] M. Diab and A. Kamboj, "Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for hindi english code switched data: A pilot annotation."

[23] H. Buschmeier and M. Wlodarczak, "Textgridtools: A textgrid processing and analysis toolkit for python," in *Conference proceedings of the 24th conference on electronic speech signal processing (ESSV 2013)*, 2013.

[24] K. Kirchhoff, "Integrating articulatory features into acoustic models for speech recognition," *Phonus 5, Institute of Phonetics, University of the Saarland*, pp. 73–86, 2000.

[25] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 25, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/25

[26] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 307, 2008. [Online]. Available: http://www.biomedcentral.com/1471-2105/9/307

[27] S. A. Tagliamonte and R. H. Baayen, "Models, forests, and trees of york english: Was/were variation as a case study for statistical practice," *Language variation and change*, vol. 24, no. 2, pp. 135–178, 2012.

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[29] R. Wright *et al.*, "A review of perceptual cues and cue robustness," *Phonetically based phonology*, vol. 34, p. 57, 2004.

[30] K. N. Stevens, A. S. House, and A. P. Paul, "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation," *The Journal of the Acoustical Society of America*, vol. 40, no. 1, pp. 123–132, 1966.

[31] R. Ahmed and S. Agrawal, "Significant features in the perception of (hindi) consonants," *The Journal of the Acoustical Society of America*, vol. 45, no. 3, pp. 758–763, 1969.