To appear in Proc. of INTERSPEECH 2025

Oral Reading Errors by Grade 3 Children in Indian Schools: A Hindi-English Perspective

Sneha Raman, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India

sneharaman@ee.iitb.ac.in, prao@ee.iitb.ac.in

Abstract

We present an analysis of reading errors on a manually transcribed dataset of Grade 3 students (N=595) who read aloud level-appropriate passages in English and Hindi. Substitutions are categorised as word or non-word errors. Further, substitutions are analysed using grapheme/phoneme sequence matching and assigned typical reading error types such as initial part matches, final part matches and scaffolding errors. We compare the distribution of error categories for the two languages and discuss underlying language-dependent decoding strategies. We also apply the analysis methodology to identify the percentage scaffolding errors per utterance given its usefulness for reading assessment. Finally, we extend our work to its practical application for diagnostics by testing an automatic phone recognition system on our task.

Index Terms: oral reading fluency, hindi, devnagiri script

1. Introduction

Achieving foundational literacy is one of the most important goals of primary school education. The ability to read a given level-appropriate text fluently and with understanding is critical to the process of learning in later years, and is directly linked to long-term academic success, competence, and social well-being. Assessments of oral reading fluency (ORF) are widely used to benchmark reading ability in early school years [1]. Word recognition, reading pace, and the proper use of expression contribute to ORF. While words correct per minute (WCPM) serves as a composite metric, the further analysis of word recognition errors can be used to accurately and efficiently identify students' letter-to-sound understanding leading to informed instructional strategies. Moreover, the use of connected text, as in ORF assessments, serves to identify struggling readers more reliably than word list reading measures [2].

The development of word reading skills in alphabetic orthographies requires the use of a decoding strategy that relates graphemes and phonemes. Ehri [3] proposes four phases in the development of accurate and fast word reading skills: pre-alphabetic, partial alphabetic, full alphabetic and consolidated alphabetic. While the first phase is considered optional, children in the second and third phases are actually establishing sight vocabulary and phonic decoding processes [4]. Detecting the nature of word recognition errors in these stages, therefore, offers insights that are useful in predicting future ability, and has been widely researched in the language pedagogy community.

Stuart and Coltheart [5] attempted to group word substitution errors into six error groups based on letter sequence comparisons of prompted words and the transcription of pronounced words. These groups consisted of: beginning letter(s) used (cat/car), final letter(s) used (hat/cat), both end letters used (bird/bad), target included in error (looks/look), letters or letter segments used (milk/like) and partial/irrelevant information used (look/baby). They found that the 'both end letters used' group or "scaffolding errors" as termed by Savage et al. [6] correlated with overall letter-sound knowledge and reading age, i.e. accurate word reading ability. Boundary consonants are known to be particularly salient to young readers [3].

Savage et al. [6] used a similar approach to categorize errors as did Stuart and Coltheart [5], but chose phoneme-based error groups rather than letter-based ones, a choice motivated by their previous study [7]. They called the categories errors preserving initial phoneme (shower/chef), errors preserving final phoneme (mesh/fish), scaffolding errors (cholera/camera, sad/salad), errors sharing orthographic overlap (look/milk) and unrelated errors (last/milk). Their results indicated that children making fewer scaffolding errors ($< 25\%$ across all error types) at age 6 were poorer readers at age 8 compared to those who made a higher percentage of scaffolding errors.

Early reading development studies are primarily focussed on English, a language with opaque orthography (low letter-sound correspondence). Early readers of the English language face the challenge of dealing with complex orthographic irregularities (know, so/do, nation/sheep, though/through/tough etc.), which is not the case for orthographically transparent languages [8]. Given our interest in Hindi, an orthographically transparent language (high letter-sound correspondence), we note that it has several complex elements in its orthography inherent to the Devanagari script such as conjunct consonants and diacritics [9]. Therefore, early readers of Hindi may have the benefit of regularity in orthography, but they face the challenge of deciphering the complex script and its multiple consonant-vowel and consonant-consonant graphemes.

Gupta and Jamal [10] investigated reading errors of dyslexic children (age 7-10 years) from English medium schools in India, in Hindi and English. Overall, there was a higher accuracy in Hindi compared to English. They categorized errors as phonological (felt/filt) or orthographic (huge/hug) errors. They found that Hindi errors were largely phonological and English errors were phonological as well as orthographic, suggesting that Hindi readers followed the sub-lexical or grapheme-phoneme correspondence route and English readers used both the sub-lexical as well as lexical routes (visual cues of the word). Look at the dual route cascading model [11] for details on the lexical and sublexical processes of reading. In addition, they used the same phoneme-based error categorization method as Savage et al. [6]. They found a high proportion of scaffolding errors for both English and Hindi, followed by errors preserving the initial phoneme, errors preserving final phoneme and errors sharing overlap. They also categorized errors based on whether they were word (last/lost) or

non-word (last/las) responses. They found a higher proportion of non-word responses for both Hindi and English, but the gap of word/non-word errors was smaller for English.

A follow up study by Gupta and Jamal [12] extended the dyslexic readers cohort by recruiting age matched normally progressing readers. They found more non-word errors compared to word errors for dyslexic Hindi and English readers as well as normally progressing Hindi readers. However, for normally progressing readers, they found more word errors in English.

All of the aforementioned studies use word lists to study word decoding errors. Flynn et al [2] argue that while word list reading has the advantage of the elimination of contextual influences, they pose a problem for educators who use level-appropriate connected texts for ORF assessment. In their study, they found that word list-based measures correlated with connected text-based measures for monosyllabic words (the typical case in word list-based studies) but not for polysyllabic words. Therefore, the analysis of word recognition errors encountered in ORF testing with passages is considered to be a more streamlined and efficient choice for assessing reading ability.

In this work, we present a comparative study of reading errors of Grade 3 in English and Hindi. The study is based on audio recordings of read-aloud connected text obtained in the course of ORF assessments implemented in an Indian school network where both languages are part of the school curriculum since Grade 1. We identify and characterise the observed word substitution errors in each language in terms of the above-reviewed error categories that are of interest to the prediction of future reading development. We carry out our study on manual transcriptions of the speech from across 400 paragraph-sized utterances by about 300 speakers in each language. In the next section, we present more details on our dataset and methodology. This is followed by a summary of our observations that are discussed from the perspective of the previous works on reading errors by early learners. We also carry out an utterance-level evaluation of the percentage scaffolding errors, given its significance in reading assessment, and test the feasibility of using an available phone recognition system for the automatic identification of at-risk students from their ORF recordings.

2. Dataset

We had access to data collected from across India as part of a benchmarking exercise for reading levels in elementary schools [13]. Ethics clearance was obtained for the recordings, with speaker information anonymized except for age and gender. For our reading error analyses, we selected a subset of each of the English and Hindi ORF recordings by Grade 3 students in the age-group 7-9 years, so that the data for the two languages are broadly comparable in terms of overall reading skill.

2.1. Text Prompts

The ORF assessment employed level-appropriate stories, one each for Hindi and English. Each story comprised of two paragraphs that contained 60 to 70 words (included in supplementary material [1]). The English story was selected from the reading cards created by the Central Institute of English and Foreign Languages (CIEFL), India [14]. The Hindi story was chosen from a Hindi textbook used by the Board of Secondary Education Rajasthan [15], a different school board to the one the children studied in, to ensure that the story is unseen.

[1] https://doi.org/10.5281/zenodo.15525203

2.2. Audio Recordings and Transcription

The text prompts were presented to the child via a mobile phone application. The school teacher recorded the child reading aloud using the 'record' and 'stop' controls in the application. The audio recordings were manually transcribed by a team of transcribers who were fluent in the respective languages. The transcribers labelled each word based on their perception of the uttered word. For Hindi utterances, all transcription was done in the Devanagari script using Microsoft's Indic Language Input Tool (ILIT), which was installed on the transcriber's machine. For English utterances, recognisable English words were transcribed in the English alphabet and intelligible English non-words in the Devanagari script. Devanagari script has high grapheme-to-phoneme correspondence, enabling easy mapping of the transcribed grapheme sequences to phoneme sequences. All transcriptions underwent a quality check by a second transcriber.

Only those utterances where the student attempted at least 70 percent of the text prompt were chosen for this study. For each utterance, we computed the standard fluency metric of Words Correct Per Minute (WCPM) and accuracy (words correct $\div$ number of words in the text prompt). These metrics were used to select a subset of the available transcribed utterances for this reading errors study.

2.3. Reading Miscue Distribution

In contrast with isolated word reading studies, identifying reading errors in connected text is complex, and requires a prior step of word-level alignment between the prompt and spoken words. This was accomplished via word sequence matching using a nuanced Levenshtein distance between the pronunciations of reference words and the manually transcribed uttered words. This nuanced Levenshtein distance accounts for word merges and break-ups that can occur due to a child's speaking style [16, 13]. Each reference word was compared with its corresponding realization in the manual transcript to be labeled as one of Correct, Deleted, or Substituted. Further, an uttered word is considered an Insertion if it does not correspond to any prompt word (e.g. false starts).

Metrics (Per utterance)	English	Hindi
WCPM	90 ± 25	97 ± 20
Accuracy	90.73 ± 6.83	91.89 ± 5.31
Substitutions	5.21 ± 3.26	4.98 ± 3.10

Table 1: *Metrics per utterance (Mean $\pm$ SD) for the 422 English utterances and 413 Hindi Utterances.*

Although the Hindi and English datasets did not have overlapping speakers, we ensured that they were of comparable reading fluency levels i.e. comparable WCPM, accuracy and number of word substitutions made when reading. Since our study focuses on comparing substitution errors made by children in English and Hindi, obtaining a similar distribution of word substitutions was a key factor when sampling utterances from English and Hindi readers. Mann-Whitney U test confirmed an equal distribution of the number of substitutions (same central tendency) per utterance (U = 90396, p = 0.347). Table 1 shows the mean WCPM, accuracy and substitutions for the Hindi and English utterances for the eventually selected subset of 422 English utterances from 310 unique readers and 413 Hindi utterances from 285 unique readers. We present the reading miscue distribution for our dataset next, followed by the analysis of substitution errors in the next section.

Table 2 presents a summary of the overall reading miscues across all the utterances in each language. In the current work on reading errors, we are interested in the substituted words.

Tags	English		Hindi	
	Count	%	Count	%
Correct words	24140	90.75	25198	91.89
Deleted words	263	0.99	165	0.60
Substituted words	2199	8.27	2058	7.51
Inserted words	727	2.73	816	2.98
total prompt words	26602	100	27421	100

Table 2: *Reading error distributions across 422 English utterances and 413 Hindi utterances.*

3. Analysis Methods and Observations

In this section, we present our methods for reading error analyses, observations and results.

3.1. Reading error categorization and distributions

We investigated two types of reading errors - word/non-word errors and phoneme/letter sequence based errors.

3.1.1. Word/nonword errors

Word/non-word tagging for English was done by checking the substituted words against the pyenchant dictionary package [2]. For Hindi, two different sources were used, the Shabd database [17] and an additional list of stop words [18], since many short words were not included in the Shabd database. For Hindi, we found more nonword errors than word errors, and for English, marginally more word errors (see Table 3).

Language	English		Hindi	
	Count	%	Count	%
Word errors	**1133**	**51.5**	762	37.0
Non-word errors	1066	48.5	**1296**	**63.0**

Table 3: *Word and non-word errors in English and Hindi*

As the word length (length of phonemes in reference word) increased, the ratio of word errors to non-word errors changed for both Hindi and English (Figure 1). For shorter reference words, there are more word errors than non-word errors, likely due to many common sight words of short length. For longer reference words there were more non-word than word errors. Examples - shepherd/sheephurt, counted/counten, मुलायम/ लिआमन (mulɑːyəm / lɪɑːmən), निर्भर/ नीबंदर (nɪrbʰər/ nɪːbəndər).

3.1.2. Phoneme Sequence Generation

Phoneme sequences were generated for every reference word in the text prompt and the corresponding substituted word/non-word in the transcription. For English, we used the g2p python library [3] to get the phone sequence. It uses the CMUSphinx phone set and dictionary [19]. We added Indian English pronunciations to the lexicon to account for variations in Indian English. For example, the use of v for w (wet and vet have the same pronunciation) [20]. For Hindi, since there is high grapheme to phoneme correspondence, the unicode characters

[2] http://pyenchant.github.io/pyenchant/
[3] https://pypi.org/project/g2p-en/

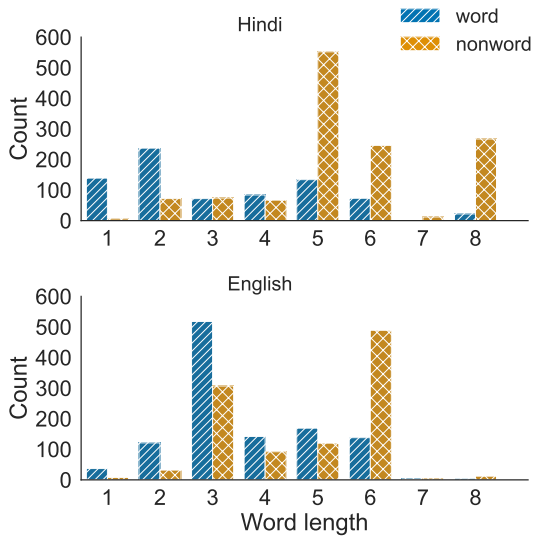


Figure 1: *Change in word/non-word error ratio with increase in phoneme sequence length in Hindi and English*

corresponding to each character was mapped to a phone set. The Hindi phoneset was larger than the English phoneset to accommodate the consonant diversity in Hindi.

3.1.3. Phoneme and letter sequence based errors

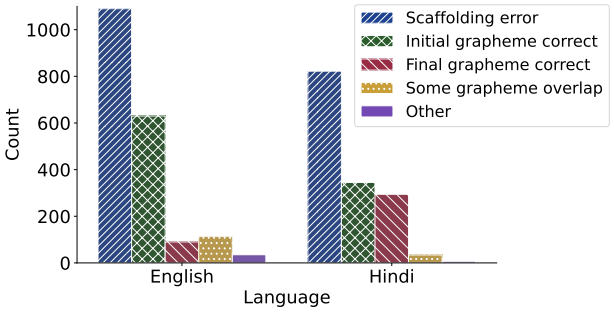
Word	Substitution	Error Category (orthographic)	Error Category (phonological)
sheep	sip	scaffolding	final correct
shepherd	sheephurt (n)	initial correct	initial correct
once	worse	final correct	scaffolding
किया	लिया	final correct	final correct
निर्भर	नीबंदर (n)	scaffolding	scaffolding

Table 4: *Examples of error categories in orthographic and phonological mode. Nonword errors are labelled as (n). IPA for devnagari words: किया (kɪjaː), लिया (lɪjaː), निर्भर (nɪrbʰər), नीबंदर (nɪːbəndər)*

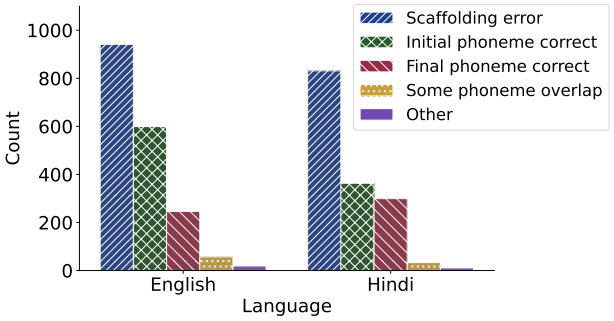
Error categories were assigned based on sequence matching of the substituted word with the reference word in two modalities - orthographic (sequence of letters or graphemes) and phonological (sequence of phonemes). See Table 4 for examples. We only considered cases where the length of the sequence (reference and substituted) was at least 3. A Levenshtein distance was used to compare the substituted word's sequence to the corresponding canonical word's sequence, resulting in a sequence of labels indicating correct or matching units (c), insertions (i), deletions (d) and substitutions (s). Then error categories were assigned based on the following rules:

- Initial correct: 'c' label in the initial but not the final position
- Final correct: 'c' label in the final but not the initial position
- Scaffolding error: 'c' label in the initial as well as final positions
- Some overlap: At least one 'c' label in the comparison sequence
- No overlap: Not a single 'c' label in the comparison sequence

In the phonological mode, scaffolding errors are the most common type of error, followed by initial correct errors and final error correct errors. English had a larger proportion of 'ini-



(a) *Orthographic*



(b) *Phonological*

Figure 2: *Error category comparisons in orthographic and phonological mode for English and Hindi*

tial correct' errors compared to 'final correct' errors, whereas for Hindi, the difference in the number of initial and final correct errors was marginal (See Figure 2b).

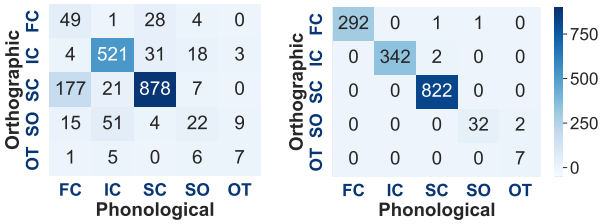
The same trend was observed in the orthographic mode, with the exception of fewer final correct errors in English (see Figure 2a). A confusion matrix of error categories in phonological and orthographic modes for English (Figure 3a) reveals that several final correct cases in the phonological mode are marked as scaffolding errors in the orthographic mode, with its overall higher proportion of scaffolding errors. Some examples of these cases are evening/everything (ˈiːvnɪŋ /ˈɛvriθɪŋ) and thorns/trowns (/θɔːnz/traʊnz). See the supplementary material [4] for more examples. For Hindi, a high agreement in the orthographic and phonological mode of error categorisation (see Figure 3b) is expected, owing to the high grapheme phoneme correspondence.

3.2. Detection of utterance-level scaffolding errors

We present an example of the practical utility of our analysis methodology. As mentioned in Section 1, there is a strong link between a low proportion of scaffolding errors (less than 25% of total errors) in a child's oral reading performance and future reading ability [6]. We therefore compute scaffolding errors at an utterance level as an important feature to identify at-risk students. Using the manual analysis for phonological errors (Section 3.1), we found that 28% of English utterances and 35% of Hindi utterances were characterised by less than 25 percent scaffolding errors.

With the practical application of large scale diagnosis of at-risk students in mind, we further implemented automatic word error detection and categorisation using automatic phone recognition. We used a hybrid ASR system to obtain word segmentation followed by a wav2vec based model for extracting the phone sequence of each uttered word. The system was adapted from the work of Gothi et al. [13] for English, and a similar system made available by the same authors for Hindi. Both systems are trained on adult speech and fine-tuned on children's speech. The phone error rates for the two systems as available from the authors are 9.7% (English) and 8.4% (Hindi).

[4] https://doi.org/10.5281/zenodo.15525203



(a) *English* (b) *Hindi.*

Figure 3: *Confusion matrix of error categories - Final Correct (FC), Initial Correct (IC), Scaffolding (SC),Some overlap (SO), Other (OT) - in orthographic and phonological mode*

Performance metric	English	Hindi
Precision	0.54	0.70
Recall	0.62	0.78
f1-score	0.57	0.74
Accuracy	0.74	0.80

Table 5: *Performance of the automatic system to detect utterances with less than 25 percent scaffolding errors*

There was a moderate-high correlation (Pearson's $r = 0.58, p < 0.001$) between the manual and automatic computation of utterance-level percentage of scaffolding errors. The correlation was higher for Hindi (Pearson's $r = 0.64, p < 0.001$) compared to English (Pearson's $r = 0.53, p < 0.001$).

The performance of the automatic system was tested given the manual binary labels (% scaffolding error $>$ / $<$ 25%) as ground truth. The automatic system performed moderately well with a better performance for Hindi compared to English (See Table 5).

4. Summary and Discussion

We presented an analysis of word reading errors in Hindi and English for connected text. Scaffolding errors were the most common error type, as observed by Gupta and Jamal as well [10]. Our findings of more word errors in English and more non-word errors in Hindi corroborate with the findings of Gupta and Jamal [12], indicating a mixture of lexical and sub-lexical processing for English and a largely sub-lexical processing for Hindi. However, as detailed in Section 3.1.1, the ratio of word/non-word errors was also modulated by word length. This is a cautionary note for word-length choices in future studies on reading errors. Orthographic transparency effects were observed for Hindi, namely, high agreement in the orthographic and phonological mode of error categorisation and better performance of the automatic phone recognition in identifying at-risk students. For English, the differences in the orthographic and phonological mode suggests that we might need to look at both the modalities for an efficient error categorisation system. The English automatic at-risk students detection system would benefit from a character recognition system in addition to the phone recognition system presented in this paper. The methods and results presented in this paper may inform future applications of connected text ORF assessment in Hindi and English.

5. Acknowledgements

The authors would like to thank Indian Institute of Technology Bombay's Tata Centre for Technology and Design and Bharti Centre for Communication for financially supporting this research work. We also thank Rahul Kumar and Raj Gothi for their valuable contributions.

6. References

[1] S. White, J. Sabatini, B. J. Park, J. Chen, J. Bernstein, and M. Li, "The 2018 naep oral reading fluency study. nces 2021-025." *National Center for Education Statistics*, 2021.

[2] L. J. Flynn, J. L. Hosp, M. K. Hosp, and K. P. Robbins, "* word recognition error analysis: Comparing isolated word list and oral passage reading," *Assessment for effective intervention*, vol. 36, no. 3, pp. 167–178, 2011.

[3] L. C. Ehri, "Phases of development in learning to read words by sight." *Journal of research in reading*, 1995.

[4] M. Stuart, R. Stainthorp, and M. Snowling, "Literacy as a complex activity: Deconstructing the simple view of reading," *Literacy*, vol. 42, no. 2, pp. 59–66, 2008.

[5] M. Stuart and M. Coltheart, "Does reading develop in a sequence of stages?" *Cognition*, vol. 30, no. 2, pp. 139–181, 1988.

[6] R. Savage, M. Stuart, and V. Hill, "The role of scaffolding errors in reading development: Evidence from a longitudinal and a correlational study," *British Journal of Educational Psychology*, vol. 71, no. 1, pp. 1–13, 2001.

[7] R. Savage and M. Stuart, "Orthographic analogies and early reading: Explorations of performance and variation in two transfer tasks," *Reading and Writing*, vol. 14, pp. 571–598, 2001.

[8] M. M. Schaars, E. Segers, and L. Verhoeven, "Word decoding development in incremental phonics instruction in a transparent orthography," *Reading and Writing*, vol. 30, pp. 1529–1550, 2017.

[9] J. Vaid and A. Gupta, "Exploring word recognition in a semi-alphabetic script: The case of devanagari," *Brain and Language*, vol. 81, no. 1-3, pp. 679–690, 2002.

[10] A. Gupta and G. Jamal, "An analysis of reading errors of dyslexic readers in hindi and english," *Asia Pacific Disability Rehabilitation Journal*, vol. 17, no. 1, pp. 73–86, 2006.

[11] M. Coltheart, K. Rastle, C. Perry, R. Langdon, and J. Ziegler, "Drc: a dual route cascaded model of visual word recognition and reading aloud." *Psychological review*, vol. 108, no. 1, p. 204, 2001.

[12] A. Gupta and G. Jamal, "Reading strategies of bilingual normally progressing and dyslexic readers in hindi and english," *Applied Psycholinguistics*, vol. 28, no. 1, pp. 47–68, 2007.

[13] R. Gothi, R. Kumar, M. Pereira, N. Nayak, and P. Rao, "A dataset and two-pass system for reading miscue detection," in *Proc. Interspeech 2024*, 2024, pp. 4014–4018.

[14] C. I. of English and F. L. (CIEFL), "English 400 reading programme." [Online]. Available: https://www.orientblackswan.com/details?id=9788125015703

[15] "Rajasthan state textbook board, jaipur, rajasthan," state Institute of Educational Research and Training (SIERT), Udaipur, Rajasthan. [Online]. Available: https://rajeduboard.rajasthan.gov.in/

[16] N. Ruiz and M. Federico, "Phonetically-oriented word error alignment for speech recognition error analysis in speech translation," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 296–302.

[17] A. Verma, V. Sikarwar, H. Yadav, R. Jaganathan, and P. Kumar, "Shabd: A psycholinguistic database for hindi," *Behavior Research Methods*, vol. 54, no. 2, pp. 830–844, 2022.

[18] S. Rajwal, "Lihisto: a comprehensive list of hindi stopwords," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 50 047–50 059, 2024.

[19] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The cmu sphinx-4 speech recognition system," in *Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong*, vol. 1, 2003, pp. 2–5.

[20] A. Sahgal and R. K. Agnihotri, "Indian english phonology: A sociolinguistic perspective," *English World-Wide*, vol. 9, no. 1, pp. 51–64, 1988.