Phonetically Balanced Code-Mixed Speech Corpus for Hindi-English Automatic Speech Recognition

Ayushi Pandey[1], B M L Srivastava[2], Rohit Kumar[3]*, B T Nellore[1], K S Teja[4]*, S V Gangashetty[1]
IIIT-Hyderabad[1], Microsoft Research[2], NIT Patna[3], MIT Manipal[4]
ayushi.pandey@research.iiit.ac.in, t-brsriv@microsoft.com, svg@iiit.ac.in

Abstract

The paper presents the development of a phonetically balanced read speech corpus of code-mixed Hindi-English. Phonetic balance in the corpus has been created by selecting sentences that contained triphones lower in frequency than a predefined threshold. The assumption with a compulsory inclusion of such rare units was that the high frequency triphones will inevitably be included. Using this metric, the Pearson's correlation coefficient of the phonetically balanced corpus with a large code-mixed reference corpus was recorded to be 0.996. The data for corpus creation has been extracted from selected sections of Hindi newspapers.These sections contain frequent English insertions in a matrix of Hindi sentence. Statistics on the phone and triphone distribution have been presented, to graphically display the phonetic likeness between the reference corpus and the corpus sampled through our method.

Keywords: code-mix speech, phonetic balance, newspaper corpus

1. Introduction

Code-mixing is a frequently encountered phenomenon in day-to-day natural language communication, especially in multilingual and bilingual communities. Code-switching is considered to be the phenomenon of alternating languages at the sentential or clausal level, and code-mixing is the word-level insertions from one language into the sentential frame of another. The phenomenon is particularly prevalent in speech communities where the native language and medium of education are recognized as two separate languages. According to the census of 2001, 12.1% of the speakers in India are speakers of English as their second or third language. Additionally, the popularity of English in social media, print media, and also entertainment make English widely accessible to most such bilingual speakers. The ubiquitous prestige associated with English in the diglossic Indian situation also motivates Indian bilinguals to show abundant code-mixing and code-switching patterns between English and other regional languages. The widespread usage and growth of this phenomenon of code-mixing mandates a shift in paradigm from monolingual automatic speech recognition (ASR) studies into code-mixed speech recognition. However, computational studies for both textual and speech processing of code-mixing suffer from a sincere disadvantage: lack of data.

In this paper, we present a Phonetically Balanced Code Mixed (PBCM) speech corpus, sampled from a standardized code-mixed text corpus, the Large Code Mixed (LCM) corpus. An optimal text selection procedure has been used to extract 6,126 utterances from the LCM. The PBCM corpus is currently in the process of being recorded and post-processed for speech recognition purposes at IIIT-Hyderabad.

The paper is organized as follows: Section 2 describes some popular methods in corpora creation, and also mentions the development of code-mixed corpora for various language pairs. Section 3 details the procedure of optimal text selection that we employed to design the PBCM corpus. Section 4 describes the recording procedure, and the progress of speech recording so far. Section 5 presents the conclusion, and Section 6 discusses some future directions.

2. Prevalent methods in corpus design

It is popularly believed that the success of the recognition and/or synthesis system depends significantly on the quality of the speech corpus. Careful attention therefore, has been paid to designing corpora that ensure a phonemic distribution appropriate for training and testing of the system. Ensuring minimal redundancy in phonetic coverage is also crucial to optimize the time consumed in post-processing. From a large and usually diverse textual database, a set of either phonetically rich or phonetically balanced sentences are selected. Phonetically rich sentences (Radová and Vopálka, 1999) contain an approximately uniform distribution of all phonemes in the language. Phonetically balanced sentences, on the other hand, represent the frequency distribution of phonemes proportionate to the "natural" phonemic distribution in the concerned language. For a phonetically transcribed corpus, the *add-on* procedure is a popular method (Falaschi, 1989). The sentence with a frequency distribution score proportionate to that of the already selected sentences gets added on to the corpus. Corpora designed for speech recognition studies require a context-sensitive phone; a triphone or another subword unit containing sequence information. For synthesis systems, corpora must contain adequate distribution of word-joins, in addition to a phonetic coverage. It is also common to optimize phonetic coverage based on a suprasegmental feature vector such as lexical stress, pitch, prosody etc (Black and Lenzo, 2001). Santen et al emphasize the importance of the preparedness of a system towards rare phonetic units (Van Santen and Buchsbaum, 1997) . To optimize coverage of all phonetic units, ASR studies

*participated in project while interning with IIIT-Hyderabad

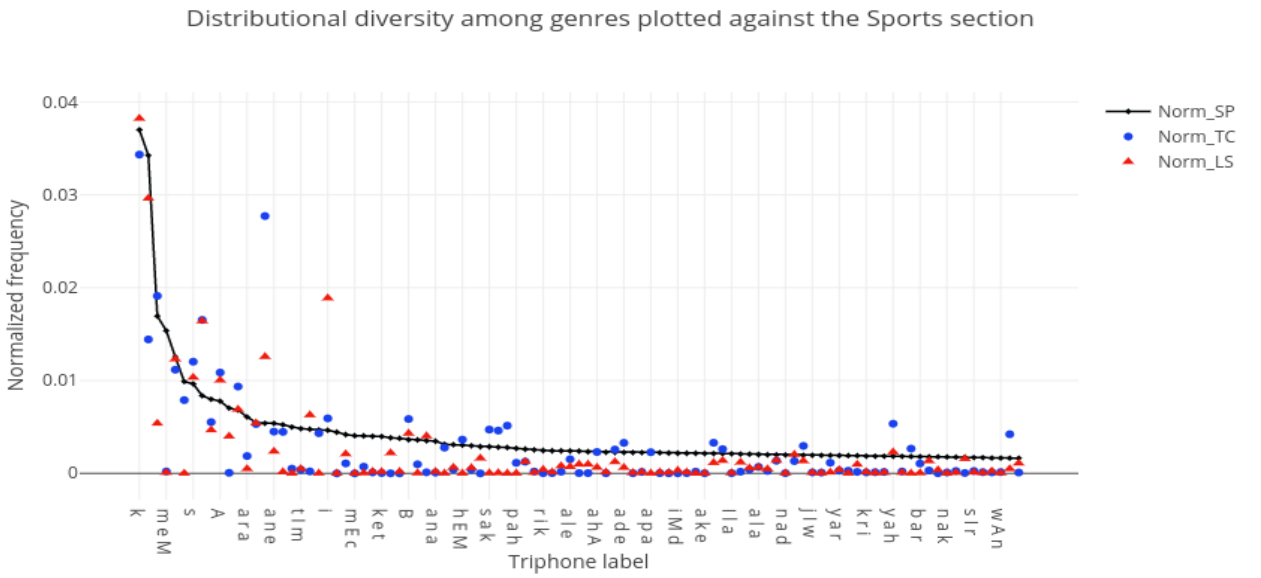

Figure 1: Plot between the disproportionate triphone coverage across two genres (Lifestyle and Technology), plotted against the third (Sports).

benefit from a greedy sentence selection approach with weighted frequencies of triphones, where the weights are the inverse of frequencies. This ensures an inclusion of rare phones in the corpus. (Van Santen and Buchsbaum, 1997). In India, there has been consistent effort to develop both phonetically rich and phonetically balanced corpora for Indian languages. (Kumar et al., 2005), (Godambe and Samudravijaya, 2011), (Arora et al., 2010), (Samudravijaya et al., 2000), (Upadhyay and Riyal, 2010) However, large-scale development of code-mixed corpora still needs attention. There have been several attempts to create speech corpora for language pairs like Mandarin-English, Cantonese-English, Frisian-Dutch, Swahili-English and so on.(Yılmaz et al., 2016), (Chan et al., 2005), (Lyu et al., 2015), (Lyu et al., 2010) (van der Westhuizen and Niesler, 2016), (Kleynhans et al., 2016) As this research field remains in the nascent stage of investigation, a read speech corpus can provide insightful contribution into modeling the acoustic properties of code mixing. A corpus designed in such a manner could offer enormous control on the lexical content, optimal phonetic coverage, choice of speakers, recording environments and reduce the dependence on post-processing. However, one of the largest challenges in approaching the development of an large vocabulary read speech corpus, is the lack of standardized code-mixed text data. The following section describes our approaches towards selecting standardized textual material which reflects patterns of Hindi-English code-mixing in print media.

3. Design of the data corpus

Conversational communication between bilingual speakers represents the dynamic nature of code-mixing in nearly all its entirety. However, there are large sections of print media now that employ recurrent patterns of code-mixing, if not switching. Columns specifically dedicated to content like technology, sports, gadgets and fashion trends show frequent word-level English embeddings in the matrix of a Hindi sentence.

Lexical diversity in phonetic coverage is also recognised as a major concern in corpora design, because the coverage of the recognition unit (triphone, syllable etc) may differ significantly from domain to domain. While monolingual corpora achieve this diversity by selecting portions from various genres, designing a code-mixed corpus requires selections that are exhaustive in English insertions. 1 displays the distributional diversity among the genres of the LCM corpus. As a first step, a large body of data was scraped from three sections, namely Gadgets and Technology, Lifestyle and Sports from the newspapers DainikBhaskar (http://epaper.bhaskar.com/) and Sanjeevani (http://www.sanjeevnitoday.com/). The following example represents the word level English insertion in the matrix of a Hindi sentence.

Example:
अनहैल्थी फूड्स को अधिकतर अवॉइड करना चाहिए ।
Gloss:
[unhealthy-ENG] [foods-ENG] [case marker-HIN] [avoid-ENG] [mostly-HIN] [do-HIN] [should-HIN]

Translation:
One should mostly avoid unhealthy foods.

Here, the English insertion has been transcribed in a matrix sentence of Devanagari. The newspaper corpus contains both English words transcribed in Devanagari, as in the example above, but also a sizeable amount of English words in their Roman transcriptions. The size of the scraped corpus is 46,595 sentences and it has been named the Large Code Mixed (LCM) Corpus. The development

of a Phonetically Balanced Code-Mixed Corpus will be detailed in the succeeding paragraphs.

3.1. Sampling corpus through triphone frequency

Triphone, as a recognition unit has been given primary importance in development of most ASR corpora. The primary reason for this consideration is the sensitivity of triphone towards both its preceding and the succeeding context. To obtain an optimal selection of sentences, the corpus needed to be balanced not only in a set of unique phones, but also the contexts that they occurred in.
A common phonetic scheme was required to cover all the possible contexts in this combination of scripts. As a large section of the vocabulary was transcribed in Devanagari, the WX notation [1] was chosen to develop a consistent phonetic representation of the entire corpus. The following paragraphs detail the formation of a bilingual dictionary and a combined phoneset.
The design on the optimal text selection was created using the following steps:

1. *Grapheme to phoneme (Roman):* Phoneme sequences for unique Roman words in the LCM corpus were generated using a grapheme to phoneme (G2P) converter trained using the CMUdict sequence-to-sequence (Yao and Zweig, 2015) model. Using the conversion map described in Table 1, the ARPABet [2] characters were transformed into their respective WX counterparts.
2. *Grapheme to phoneme (Devanagari):* Phoneme sequences for unique Devanagari words in the LCM corpus were generated by converting them to their corresponding WX notation (Bharati et al., 1995), (Bhat, 2016). . Normalization of this phoneset was achieved through pruning out the word-final schwa, and other special characters such as "nukta".
3. *Bilingual dictionary:* Concatenating the two phonetic dictionaries generated in steps 1. and 2., a bilingual pronunciation dictionary was created. The total number of unique phones in the corpus, derived from the combination of WX and the ARPAbet-adapted WX was recorded to be 65.
4. *Preprocessing :* As a pre-processing step for creation of a read-speech corpus, sentences only sentences of length 5-15 words were selected. Punctuations (except '.',' |', ',') were pruned out. Web addresses were replaced by the single word "website". Numerals were converted to their Hindi word expansion equivalents. The size of this cleaned corpus was 23,389 sentences.
5. *Triphone coverage:* The cleaned and pre-processed corpus (in step 4.) was converted to its corresponding phonetic representation (mapped from the dictionary generated in step 3.). Word-internal triphones were collected and arranged based on the descending order of their frequency of occurrence. To ensure the coverage of rare phones, all the unique sentences that contained words that were composed of the triphones lower in frequency than the threshold, were selected and added to the corpus. The threshold was set to 10.
6. *Correlation computation:* After this selection process of sentences, a metric that compared the true distribution in the sampled corpus with the LCM was required. Unique phones (monophones) from both the LCM corpus and the PBCM corpus were collected as vectors, and a Pearson's correlation coefficient was computed.

[1] https://en.wikipedia.org/wiki/WX_notation
[2] https://en.wikipedia.org/wiki/ARPABET

Table 1: ARPABet to WX notation conversion table

ARPABet	WX notation	Devanagari
AA	A	आ
AE	E	ऐ
AH	a	अ
AO	O	औ
AW	a u	आउ
AY	a i	आ ई
B	b	ब
CH	c	च
D	d	ड
DH	x	द
EH	e	ए
ER	a r	अ र
EY	e	ए
F	P	फ
G	g	ग
HH	h	ह
IH	i	आ
IY	I	ई
JH	j	ज
K	k	क
L	l	ल
M	m	म
N	n	न
NG	M g	न्ग
OW	o	ओ
OY	O i	औ ई
P	p	प
R	r	र
S	s	स
SH	S	श
T	t	ट
TH	W	थ
UH	u	उ
UW	U	ऊ
V	v	व
W	v	व
Y	y	य
Z	j	ज
ZH	j	ज

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x^2)][n\sum y^2 - (\sum y^2)]}} \quad (1)$$

Equation (1) describes the Pearson's correlation r, where n is the number of pairs to be scored, x is the value contained in the first variable (in our case, the phonetic distribution of the LCM corpus), and y is the value contained in the second variable (phonetic distribution of the PBCM corpus).

The Phonetically Balanced Code-Mixed (PBCM) corpus of 6,126 sentences was created through the culmination of steps described above. Table 1 describes the conversion mapping between the ARPABet and WX notation. Phonemes that lacked a direct equivalent were either approximated to their closest sounding phoneme in WX, or represented as a combination of more than one phone.

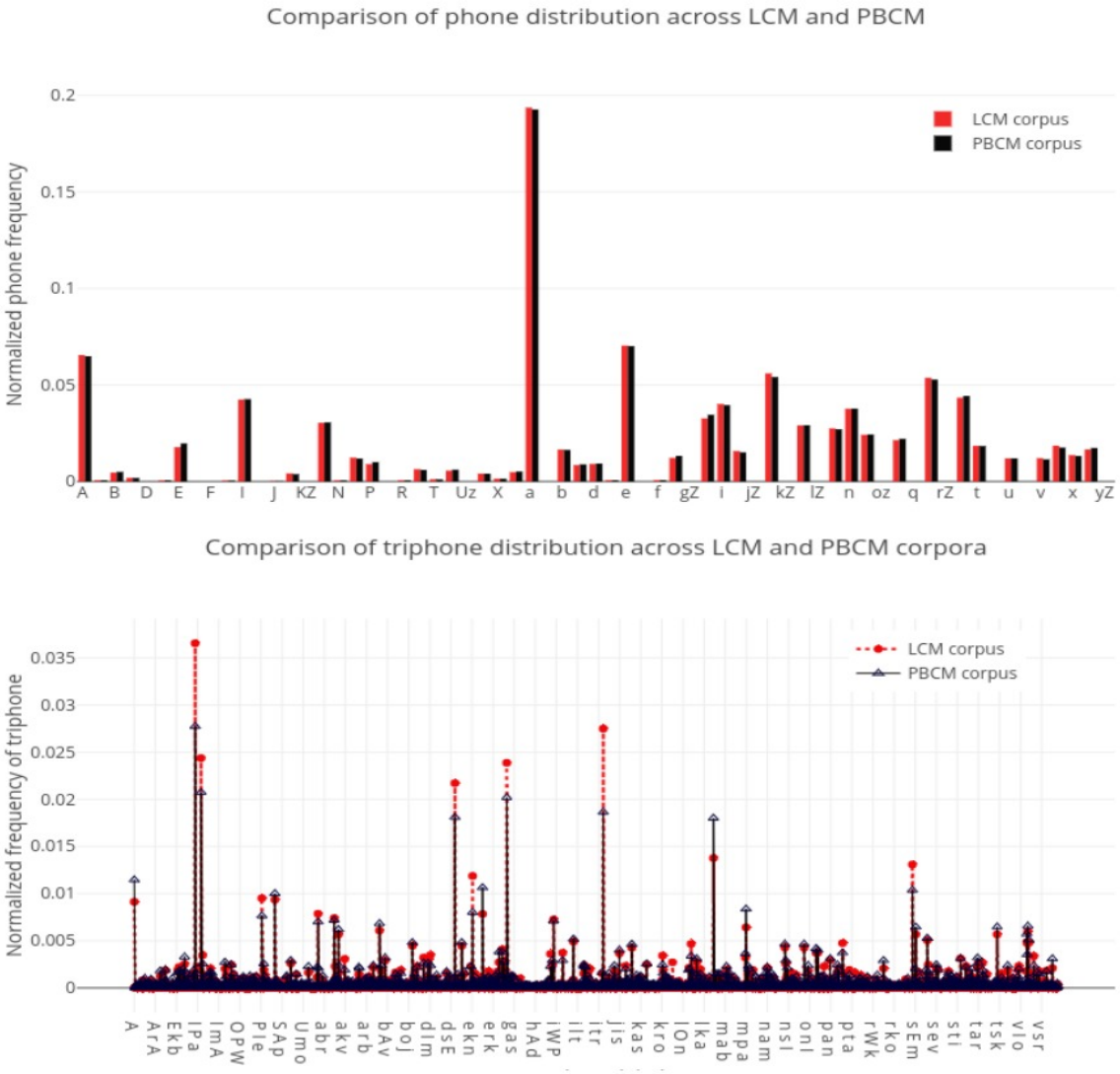

Figure 2: The top panel traces the sum-normalized frequency for phones in the LCM and the PBCM corpus, against the phoneme label. A similar distribution has been described for the triphones. The dotted red curve represents the sum normalized phone/triphone distribution in the LCM corpus, whereas the solid black line represents the same in the PBCM corpus. Both graphs are plotted with the common triphone threshold = 10.

The Pearson's correlation coefficient between the LCM and the PBCM corpora was found to be 0.996. A high correlation value indicates a proportionate distribution of phones between the sampled corpus (PBCM), and the reference corpus (LCM). Figure 1 displays the phone and triphone coverage in the LCM and PBCM corpora. Table 2 displays the comparison between the corresponding sentence, word, triphone and phone units. From Step 4, we observe that the size of the LCM corpus has considerably reduced. This results in the loss of certain phones, giving the phone distribution between LCM and PBCM an imbalance (Table 2, col. 4). It can be observed that a high correlation with the LCM is maintained despite having a non-equal phonetic distribution in the PBCM. We believe that this is because phones that show no occurrence (frequency: 0) in the PBCM are also recorded to be fairly low in frequency in the LCM. It may be observed that a large amount of code-mixed speech utterances have been removed in this way. The purpose of the corpus so created is to be able to achieve maximal phonetic coverage in the least possible amount of recordings.

4. The recording procedure

After the sentence selection procedure is completed, the next step is to conduct the actual recordings. This section presents a detailed description of volunteer speakers, recording environment and the equipment setup utilized for recording.

4.1. Description of speakers

Speech recordings are being collected from 100 volunteer speakers (50 male and 50 female), who are each a native speaker of Hindi and received education in English medium schools. All speakers are students of IIIT-Hyderabad. The age range of these speakers was between 18-35 years. The PBCM corpus is equally divided among the speakers, so that every speaker records around 62 sentences.

4.2. Recording environment and equipment

The recording of the speech utterances of the PBCM corpus are being conducted in a soundproof voice recording studio (Speech and Vision Lab, IIIT, Hyderabad). The recordings are

Table 2: Distribution across the LCM and PBCM

	Sentences	Words	Triphones	Phones
LCM	46,595	30,578	11,370	65
PBCM	6,126	8,683	6,599	57

administered through the OctaCore speech processing software, with a high fidelity noise free microphone. Wavfiles are being recorded through the open-source, cross-platform audio software, Audacity. The recordings are sampled at 48kHz and recorded at 24-bit resolution.

Each volunteer speaker is instructed to maintain a distance 5-6 inches from the microphone. Speakers record 20 sentences in one pass, after which they were given a water-break and vocal rest of 2-5 minutes. Before each recording session, the speaker is *primed* by having the sentences read out aloud to them, in order to minimize hesitation while speaking. After every 100 sentences, the speaker is given a vocal rest for 10 minutes. So far, 78 speakers (40 male and 38 females) have been recorded.

4.3. Post-processing of audio files

At this stage, the data is completely unsegmented, which means that there is only wavefile per speaker. After completing 100 speakers and exhausting all the utterances of the PBCM speech corpus, the data will be post-processed as a final step. A long sound file of 62 utterances will be manually split into one sound file per sentence format, using Praat. Non-verbal sounds and repetitions will also be manually removed, and only noise-free sentences will be compiled. For preparing the data suitable for use for speech recognition, we plan to give each sound file a unique ID, which will contain the speaker information and the serial number of recording. A silence of 1 second will be appended to each sound file, both before and after the utterance. The sound files, initially recorded at 48 kHz and 24-bit resolution, will also downsampled to 16 kHz and a 16-bit resolution.

5. Conclusion

The paper presents a phonetically balanced read speech corpus for code-mixed Hindi-English automatic speech recognition. The PBCM corpus has been sampled from a Large newspaper Corpus (LCM), which contains rich lexical insertions from English in a matrix of Hindi sentences. The inclusion of rare triphones in the sampled corpus has resulted in a high phonetic coverage (correlation: **0.996**), even with a small number of sentences. To the best of our knowledge, the PBCM can be safely proposed as one of the first phonetically balanced corpus of code-mixed speech in an Indian language pair. Recordings through the contribution of 100 Hindi-English bilinguals is aimed for, of which 78 speakers have been recorded. Once post-processed, the PBCM corpus will be made available for research and related purposes.

6. Future direction

We have observed that increasing the threshold value increases the number of sentences, while maintaining a steadfast correlation. We hope to provide an adaptive measure for selecting sentences by choosing the appropriate threshold within a given range. We also hope to compare the results of our proposed metric for selecting sentences, with the already existing methods of phonetic balance in terms of optimum corpus size, and correlation measure.

7. References

Arora, S., Saxena, B., Arora, K., and Agarwal, S. (2010). Hindi asr for travel domain. In *Oriental COCOSDA*.

Bharati, A., Chaitanya, V., Sangal, R., and Ramakrishnamacharyulu, K. (1995). *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.

Bhat, I. A. (2016). indic-wx-converter. https://github.com/ltrc/indic-wx-converter.

Black, A. W. and Lenzo, K. A. (2001). Optimal data selection for unit selection synthesis. In *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

Chan, J. Y., Ching, P., and Lee, T. (2005). Development of a cantonese-english code-mixing speech corpus. In *Ninth European Conference on Speech Communication and Technology*.

Falaschi, A. (1989). An automated procedure for minimum size phonetically balanced phrases selection. In *Speech Input/Output Assessment and Speech Databases*.

Godambe, T. and Samudravijaya, K. (2011). Speech data acquisition for voice based agricultural information retrieval. In *Proc. Of 39th All India DLA Conference, Punjabi University, Patiala, June*.

Kleynhans, N., Hartman, W., van Niekerk, D., van Heerden, C., Schwartz, R., Tsakalidis, S., and Davel, M. (2016). Code-switched english pronunciation modeling for swahili spoken term detection. *Procedia Computer Science*, 81:128–135.

Kumar, R., Kishore, S., Gopalakrishna, A., Chitturi, R., Joshi, S., Singh, S., and Sitaram, R. (2005). Development of indian language speech databases for large vocabulary speech recognition systems. In *Proceedings of SPECOM*.

Lyu, D.-C., Tan, T. P., Chng, E., and Li, H. (2010). Seame: a mandarin-english code-switching speech corpus in south-east asia. In *INTERSPEECH*, volume 10, pages 1986–1989.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin–english code-switching speech corpus in south-east asia: Seame. *Language Resources and Evaluation*, 49(3):581–600.

Radová, V. and Vopálka, P. (1999). Methods of sentences selection for read-speech corpus design. In *International Workshop on Text, Speech and Dialogue*, pages 165–170. Springer.

Samudravijaya, K., Rao, P., and Agrawal, S. (2000). Hindi speech database. In *Proceedings of International Conference on Spoken Language Processing*.

Upadhyay, R. and Riyal, M. (2010). Garhwali speech database. *Proceedings of O-COCOSDA*.

van der Westhuizen, E. and Niesler, T. (2016). Automatic speech recognition of english-isizulu code-switched speech from south african soap operas. *Procedia Computer Science*, 81(Supplement C):121 – 127. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Van Santen, J. P. and Buchsbaum, A. L. (1997). Methods for optimal text selection. In *EuroSpeech*.

Yao, K. and Zweig, G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1506.00196*.

Yılmaz, E., van den Heuvel, H., and van Leeuwen, D. (2016). Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science*, 81(Supplement C):159 – 166. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.