Towards developing a phonetically balanced code-mixed speech corpus for Hindi-English ASR

Ayushi Pandey
IIIT-Hyderabad
Hyderabad
email@domain

Brij Mohan Lal Srivastava
Microsoft Research
Bangalore, India
t-brsriv@microsoft.com

S V Gangashetty
IIIT-Hyderabad
Hyderabad
svg@iiit.ac.in

Abstract

This paper presents the ongoing process in the design of the first phase of the phonetically balanced code-mixed corpus of Hindi-English speech (PBCM-Phase I). The reference corpus is a large code-mixed (LCM) newspaper corpus selected from the sections that contain frequent English insertions in a matrix of Hindi sentence. From a phonetically transcribed corpus, compulsory inclusion of lowest frequency triphones has been ensured, with the assumption that high frequency phones may automatically be included. A high correlation of 0.81 with the representative large corpus has been observed. A small scale speech corpus of 5.6 hours has been collected, by the contribution of 4 volunteer native Hindi speakers. The recording has been conducted in a professional recording studio environment. As a second contribution, this paper also presents a baseline recognition system with pooled monolingual and code-mixed speech datasets as training and testing environments.

1 Introduction

Code-mixing is a frequently encountered phenomenon in day-to-day communication in multilingual and bilingual communities. The phenomenon is so widespread that is often considered a different, emerging variety of the language. In India, English has been granted the status of an official language by the constitution. Additionally, there are complex diglossic patterns existing between most of the regional languages and English, where English is usually the language of prestige. Indian bilingual speakers therefore, show abundant mixing and switching between their regional language and English. Computational modeling of the phenomenon of code-mixing and code-switching assumes particular relevance with the advancement of social media. However, computational studies for both textual and speech processing of code-mixing suffer from a sincere disadvantage: lack of data.

To investigate the problem in a controlled environment, the paper presents the first phase of a Phonetically Balanced Code-Mixed (PBCM-Phase I) read speech corpus.

The design of the paper is as follows: Section 2 elaborates the popular methods in the area of corpus design. Section 3 details the metric in use for designing the speech corpus. Section 4 details the recording procedure and information about speakers. Section 5 provides a brief introduction to DNN based acoustic modeling, language modeling and an adaptive implementation of both in bilingual speech recognition. Section 6 presents the results and concludes the paper.

2 Related studies in corpus develo

It is commonly believed that the quality of the training data for nearly all speech processing systems, largely determines the success of the systems. Adequate phonemic coverage with minimal redundancy is crucial in corpus design, to allow for a wide coverage of common phonetic forms in a variety of their contexts. A large and usually diverse text corpus serves as a **reference** corpus, from which a set of phonetically rich and/or balanced sentences are selected. Phonetically rich sentences (Radová and Vopálka, 1999) contain a homogeneous frequency distribution of

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 95–101,
Kolkata, India. December 2017. ©2016 NLP Association of India (NLPAI)

Table 1: Genre-wise distribution: LCM corpus

Section	**Number of sentences**
Lifestyle	9,495
Sports	11,202
Gadgets and Technology	11,342

all phonemes in the language. In a phonetically rich corpus, adequate training instances of almost all phones, or at least one instance of every phone are compulsorily included. In a phonetically balanced corpus, the distribution of phones is modeled to be proportionate to the natural phonemic distribution in the concerned language. Once the phonetic transcriptions are made available along with the speech recordings, the *add-on* procedure is a popular method (Falaschi, 1989). From the reference corpus, a set of sentences are randomly selected as the seed corpus. Thereafter, sentences with frequency scores proportionate to those of the already selected corpus are chosen. Speech databases designed especially for recognition studies benefit from a context-sensitive phone; for example a triphone or another subword unit like a syllable or a diphone. Santen et al (Van Santen and Buchsbaum, 1997) note that a training corpus requires to be prepared towards less frequent phonetic units. To optimize coverage of all phonetic units, ASR studies usually implement a sentence selection approach with weighted frequencies of triphones, where the weights are actually the inverse of frequencies. This ensures an inclusion of rare phones in the corpus, while the high frequency phones are collected inadvertently. (Van Santen and Buchsbaum, 1997). In India, there has been heavy investment on developing corpora that are both phonetically rich and/or balanced. But most of these have been designed for monolingual speech recognition purposes, and do not cover the scope of code-mixing. The next section details our approaches in design and development of PBCM-Phase I, the first phase of a phonetically balanced code-mixed speech corpus for an Indian language pair.

3 Design of the data corpus

The nature of code-mixing has best been seen reflected in conversational communication, because the practice of code-mixing is still frowned upon in formal registers. However, owing to an increasing readership, selected sections (like Sports, Technology, Lifestyle) of newspapers offer enormous coverage of code-mixing. In addition to a wide and diverse coverage, these sections have also introduced a standardization into code-mixed diction. In this paper, we design a representative corpus, the Large Code-Mixed (LCM) Corpus as a large and diverse textual database, scraped from three sections, namely Gadgets and Technology, Lifestyle and Sports from the popular Hindi newspaper DainikBhaskar (http://epaper.bhaskar.com/). Details of these sections are given in Table 1. Figure 1 displays the code-mixing distributions in the respective genres.

Upon preliminary observation, we note that while the Sports and the Gadgets and the Technology sections have prominent technical vocabulary borrowings, this content is not always limited to lack of parallel vocabulary in the matrix language.

Example:
पार्ट्स खरीद कर टेक्नीशियन से बदलवा सकते हैं ।

Translation:
“One could buy parts and get them replaced by a technician.”

3.1 Selecting sentences based on triphone frequency

In development of most ASR corpora, frequency of the triphone has been given specific importance. This is primarily because of the ability of the triphone to be sensitive to both the 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

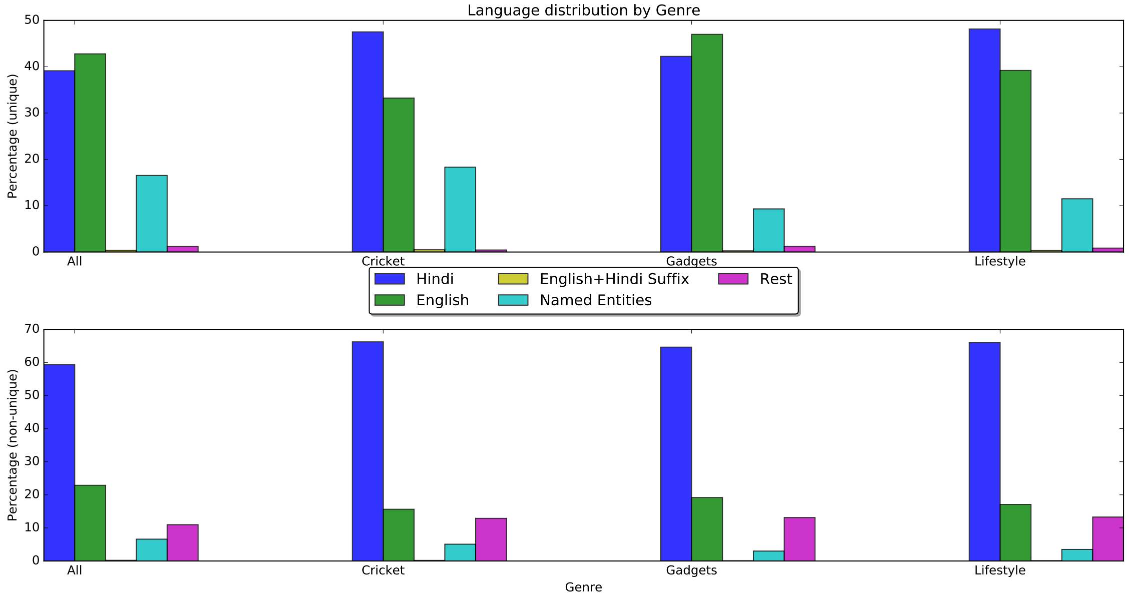

Figure 1: Stacked graphs displaying the Unique (above) vs Non-unique (below) frequency of distribution of English embeddings, Named-Entities and Rest contained in the PBCM corpus.

they occurred in.

Word-internal triphones were chosen as opposed to word-adjacent triphones, because word-internal triphone sequence can aid the identification of language at a word-level.

The design on the optimal text selection was created using the following steps:

1. As a pre-processing step for creation of a read-speech corpus, sentences only of length 5-12 words were selected.

2. Phoneme sequences for unique Roman words in the corpus were generated using a grapheme to phoneme (G2P) converter trained on 7000 words using sequence-to-sequence (Yao and Zweig, 2015) learning approach implemented in Tensorflow.

3. Phoneme sequences for uniqueDevanagari words in the corpus were converted to their corresponding WX notation representation. (Bharati et al., 1995)

4. Word-internal triphones were collected and arranged based on the descending order of their frequency of occurrence in the corpus.

5. To ensure the coverage of rare phones, all the sentences containing words that were composed of the triphones lower in frequency than the threshold, were selected.

The Phonetically Balanced Code-Mixed (PBCM) corpus of 2,694 sentences was created based on the low frequency of the rarest triphones.

3.2 Correlation computation

After the selection process of sentences, we wanted to ensure that this is truly representative of the distribution of phones present in natural language. Unique phones from both the LCM corpus and the PBCM corpus were collected, and a Pearson's correlation was computed between them. The Pearson's correlation coefficient between the two vectors was found to be 0.81. A high correlation value display a proportionate distribution of phones between the sampled corpus PBCM, and the reference corpus LCM.

3.3 Text annotations

We are in the process of annotating this data at the following four levels: 1) language identity, 2) word-identity, 3) part-of-speech and 4) word-identity in the phonetic form. Manual annotation of language identity on the corresponding text corpus has been completed for 1,760 sentences. A percentage

Table 2: Details of speakers for the PBCM corpus

Speaker ID	Sentences	Age
FEMALE-1	675	32
FEMALE-2	673	25
MALE-1	676	28
MALE-2	671	22

distribution of English embeddings across genres is illustrated in Figure 2.

As can be clearly seen from Panel 1 of Figure 1, the Gadgets and LifeStyle section display a wider distribution of unique English word embeddings in their respective content.

4 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. The duration of recorded utterances collected so far is **5.6 hours**.

4.1 Speakers description

Speech recordings were collected from 4 volunteer speakers (2 male and 2 female), who were each a native speaker of Hindi and had received education in English medium schools. The age range of these speakers was between 20-35 years. Sentences from the designed PBCM corpus was equally divided among the speakers, so that every speaker recorded around 675 sentences. Exact details of speakers can be found in Table 2.

At this stage, the corpus reflects a balance in phonetic coverage, but low acoustic variability in terms of speakers. We are planning to develop this corpus into a large corpus of about 100 speakers, with a more vast collection of speech utterances.

4.2 Recording environment and equipment

The recording of the speech utterances of the PBCM corpus was conducted in a professional voice recording studio (Deepali Studio, Lucknow, Uttar Pradesh). The recordings were administered through the Nuendo speech processing software. The equipment consisted of an integrated SoundCraft Digital-Mixer, a high fidelity noise free Sennheiser microphone and two Yamaha studio speaker systems.

The volunteer speaker was instructed to maintain a distance 10-12 inches from the microphone. Each speaker conducted recordings in a set of 20 sentences, after which they were given a water-break and vocal rest of 2-5 minutes. Before each recording session, each volunteer speaker was primed by having the sentences read out aloud to them, in order to minimize hesitation while speaking. After every 100 sentences, the speaker was given a vocal rest for 10-15 minutes.

4.3 Post-processing of audio files

The files recorded through the session were then post-processed as a final step. A long sound file of 20 utterances each was manually split into a one sound file per sentence format, using Praat. Repetitions and non-verbal sounds were also manually removed, and only noise-free sentences were compiled. For preparing the data for use for speech recognition, we gave each sound file a unique ID, which contained the speaker information and the serial number of recording. A silence of 1 second was appended to each sound file, both before and after the utterance. The sound files, initially recorded at 44 kHz and 24-bit resolution, were also downsampled to 16 kHz and a 16-bit resolution.

5 Baseline Automatic Speech Recognition

Computational modeling of the phenomenon of code-mixing assumes particular relevance with the advancement of social media in multilingual and bilingual communities. In processing of code-mixed speech, several ideas have been put forward.

5.1 The acoustic modeling component

Acoustic model aims to establish statistical relationship between speech utterances and the corresponding text.

In general, let $\boldsymbol{O} = \{x_1, ..., x_T\}$ be the acoustic observations and $\boldsymbol{w} = \{w_1, ..., w_T\}$ be the corresponding word sequence. Then the DNN must learn $p(\boldsymbol{w}|\boldsymbol{O})$, which is the conditional distribution of words given acoustic observations. DNN acts as a discriminative classifier which classifies tied-state phoneme classes (*senones*) given the acoustic observations $\boldsymbol{O}$. The acoustic model decodes speech utterance and proposes a directed acyclic graph (*lattice*) of phonemes with edges as transition probabilities. The lattice is then searched for contesting legal hypotheses. In order to correct the errors made by the DNN acoustic model, we multiply the probabilities from the existing knowledge in form of language model. This process is called lattice rescoring. By devising statistical language models which can mimic the original structure of language, we can supplement the probability of correct hypothesis and boost the accuracy of the overall system.

Multilingual and code-mixed ASR have seen two major trends. The first approach uses a language identification system implemented at the front-end, and a monolingual speech recognizer at the back end. Once the language has been identified at the word level, the segments (words) are passed as input to monolingual speech recognizers for phoneme decoding. However, such two-pass approaches return inferior results, owing to an unpredictable error-propagation from the language identifier at the front end, to the speech recognizer at back end. To circumvent this error, a one-pass approach is chosen, wherein the language identification component is completely removed. Some such efforts have been made by Bhuvanagiri (Bhuvanagiri and Kopparapu, 2010) et al, where they exploit an adaptation of the existing monolingual (English) training resources for code-mixed Hindi-English speech recognition. An approximation of the missing Hindi phonemes is achieved using either a direct mapping or a combination of existing English phones. Similar monolingual training resource extrapolation studies have been conducted by Fung et al, in experimenting with three sets of phonemic adaptation. (Yuen and Pascale, 1998). More approaches to merging phonesets can be seen in an interpolation of two monolingual speech corpora. (Chan et al.,). Model adaptation of monolingual corpora for code-mixed speech recognition has also been augmented with a model reconstruction with accented speech. (Li et al., 2011)

5.2 Language independent phones

One of the primary stages in the design of a code-mixed or multilingual speech recognition system, is the development of a combined phoneset. A combined phoneset allows for the recognition system to be prepared for all phones of the participating languages. If one of the languages is low in resources, then its phones are mapped to the closest approximations of phones in a high-resource language. A variety of phonemic adaptation methods have been explored, for example rule-based, manual (Bhuvanagiri and Kopparapu, 2010), (Yuen and Pascale, 1998) or clustering. (Li et al., 2011) For the purpose of this experiment, the speech utterances that are contained in the PBCM-Phase I are extracted from a Hindi national newspaper. The English embeddings are predominantly in Devanagari. However, the corpus does have a sizeable collection of sentences that have word-insertions in Roman script. To ensure phonetic consistency among all the transcripts, we use automatic transliteration (Bhat et al., 2015) to convert the words in Roman script into their respective Devanagari representation.

This experiment can further be refined through evaluating correspondence between the resulted phonetic transcription of the WX and the actual English phonetic transcription, and then intervening with a rule-based mapping. For this, however, an LID for Devanagari would need to be prepared, which is beyond the scope of the present study.

5.3 Feature selection and extraction

We propose the usage of the WX notation (Bharati et al., 1995) for establishing a grapheme to phoneme representation of the Devanagari. For the Roman utterances, the sequence-to-sequence converter has been implemented, and the phonetic representation belongs in the IPA.

The DNN model is trained over features obtained initially by concatenating ±4 frames of MFCC followed by followed by Linear Discriminant Analysis (LDA). The features thus obtained have unit variance. These features are subjected to Maximum Likelihood Linear Transform (MLLT). MLLT is a feature-space transform with the objective function which is defined as the sum of the average per-frame log-likelihood of the transformed features given the model, and the log determinant of the transform.

In the end, we apply feature-space Maximum Likelihood Linear Regression (fMLLR), which is an affine feature transform of the form $x \rightarrow Ax + b$. We finally obtain the 40-dimensional feature set used for DNN training.

5.4 Training and test corpora development

The main objective of the latter part of this study is to be able to utilize and adapt high-resource monolingual corpora for a low-resource setting such as code-mixed speech. In order to achieve such an extrapolation, an adaptation of phonemes has already been explained in the previous section. The training dataset comprises of a combination of monolingual Hindi speech and a small portion of code-mixed speech.The training dataset comprises of monolingual speech corpus containing speech recordings from 17 speakers, collected through the Hindi DD-News channel and Indic speech database and 3 code-mixed speakers, collected through the PBCM corpus. The testing dataset comprises of the 3 monolingual speakers and 1 code-mixed speaker, collected through the PBCM corpus.

5.5 The language modeling component

Language models are prevalent in ASR studies for providing word-level probability scores derived from the sequential structure of sentences. N-gram (trigram, 4-gram etc) language models are designed on the assumption that the probability of a given word $p(w_t)$ can be determined based on the context h_t that it is preceded by.

$$p(w_{1:T}) = \prod p(w_t|w_{t-1}w_{t-2}..w_{t-T}) = p(w_t|h_t) \tag{1}$$

Several ideas have been put forward in designing language models for code-mixed speech. Approaches relying on (Vu et al., 2012) acoustic modeling alone have been refined and augmented through modifications of the language model. Grammatical constraints (equivalence and government constraint) are implemented in (Li and Fung, 2012), to predict the correctness or likelihood of a switch in a sentence. Additionally, to circumvent the limitations of low-resources in data, class-based language models (Yeh et al., 2010), (Tsai et al., 2010) are used. Improved language modeling for code-mixed speech recognition have also aided in characterising some of the speaker specific patterns of code-mixing. (Vu et al., 2013)

6 Results and discussion

Acoustic models were trained according to Dan's NNET2 setup (Povey et al., 2014). The featureset implemented for training has been described in detail in section 5.3. We conducted two sets of experiments with respect to evaluating the scalability of the training corpus.

- **Expt 1:** The speech transcripts that belonged in the testing corpus were included in the language model training. This setup was designed so as to remove any instance of an out-of-vocabulary word, and evaluating the performance of an monolingual acoustic model with an adapted phoneset.

- **Expt 2:** The speech utterances that had been covered in the spoken corpus were excluded from the language model training. The design of this setup allowed us to evaluate the ASR based on a monolingual language modeling and monolingual acoustic modeling.

Expt 2 reveals that the WER obtained over the mixed (3 monolingual, 1 code-mixed) test set evolved from 72.34% with respect to monophone training to 41.63% for Dan's NNET2. Removing out-of-vocabulary words significantly reduces the WER, as the basline (monophone) results in a far lower error 46.54 %, when compared with Expt 1.

7 Conclusion

We present the initial design and the ongoing process in the development of a phonetically balanced speech corpus of Hindi-English non-conversational speech. Data has been subsetted from selected sections of a popular Hindi newspaper, DainikBhaskar, and a corpus of

LM	mono	tri1	tri2b	tri2b$_{m}mi$	tri2b$_{m}pe$	tri3b	tri3c	sgmm2	nnet
Expt 1	46.54	35.84	37.54	36.44	36.74	15.35	17.86	13.59	10.63
Expt 2	72.34	58.64	59.05	59.32	59.07	45.29	46.72	41.60	41.66

Table 3: Table with word error rate (WER) of different acoustic models implemented with the two language modeling setups

2,694 sentences have been collected. Sampling from a Large Code-Mixed (LCM) corpus into a Phonetically Balanced Code-Mixed corpus has been designed through a threshold frequency of triphones, with the assumption that high-frequency phones would be accommodated inadvertently through the process. The first phase of the PBCM corpus has successfully been recorded. We also present the development of a baseline automatic speech recognition system, modeled on adapting the available high-resource such as the monolingual Hindi speech corpora.

References

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective.* Prentice-Hall of India New Delhi.

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

K Bhuvanagiri and Sunil Kopparapu. 2010. An approach to mixed language automatic speech recognition. *Oriental COCOSDA, Kathmandu, Nepal.*

Joyce YC Chan, Houwei Cao, PC Ching, and Tan Lee. Automatic recognition of cantonese-english code-mixing speech.

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

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *COLING*, pages 1671–1680.

Ying Li, Pascale Fung, Ping Xu, and Yi Liu. 2011. Asymmetric acoustic modeling of mixed language speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5004–5007. IEEE.

Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. 2014. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455.*

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

Tsai-Lu Tsai, Chen-Yu Chiang, Hsiu-Min Yu, Lieh-Shih Lo, Yih-Ru Wang, and Sin-Horng Chen. 2010. A study on hakka and mixed hakka-mandarin speech recognition. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 199–204. IEEE.

Jan PH Van Santen and Adam L Buchsbaum. 1997. Methods for optimal text selection. In *EuroSpeech.*

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4889–4892. IEEE.

Ngoc Thang Vu, Heike Adel, and Tanja Schultz. 2013. An investigation of code-switching attitude dependent language modeling. In *International Conference on Statistical Language and Speech Processing*, pages 297–308. Springer.

Kaisheng Yao and Geoffrey Zweig. 2015. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1506.00196.*

Ching Feng Yeh, Chao Yu Huang, Liang Che Sun, and Lin Shan Lee. 2010. An integrated framework for transcribing mandarin-english code-mixed lectures with improved acoustic and language modeling. In *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pages 214–219. IEEE.

MA Chi Yuen and FUNG Pascale. 1998. Using english phoneme models for chinese speech recognition. In *International Symposium on Chinese Spoken language processing*, pages 80–82. Citeseer.