Adapting monolingual resources for code-mixed Hindi-English speech recognition

by

Ayushi Pandey, Brij Mohan lal, Suryakanth V Gangashetty

in

21st International Conference on Asian Language Processing (IALP 2017)
(*IALP-2017*)

Singapore

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2017

Adapting monolingual resources for code-mixed Hindi-English speech recognition

Ayushi Pandey
Speech and Vision Lab, IIIT-Hyderabad
Hyderabad, India
ayushi.pandey@research.iiit.ac.in

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

Suryakanth V Gangashetty
Speech and Vision Lab, IIIT-Hyderabad
Hyderabad, India
svg@iiit.ac.in

***Abstract*—The paper presents an automatic speech recognition (ASR) system for code-mixed read speech in Hindi-English, developed upon the extrapolation of monolingual training resources. A monolingual Hindi acoustic model, mixed with code-mixed speech data has been implemented to train a neural network based speech recognition framework. The testing corpus also follows a similar structure: containing data from both monolingual and code-mixed speech. The shared phonetic transcription, captured in WX notation has been exploited to harness the commonality between the pooled phonesets of Hindi and English. The experiments have been conducted in two separate formulations of a trigram based language model 1) In the first experiment, the language model contains no out-of-vocabulary words, as the test utterances are included in the training of the language model. The word error rate in this case has been obtained to be 10.63 %. 2) In the second experiment, the testing utterances have been excluded from the training language model. The word error rate in this case has been obtained to be 41.66 %.**

***Keywords*-code-mixed speech recognition, low-resource acoustic modeling**

I. Introduction

Bilingual and multilingual speech communities recognize code-switching and code-mixing as predominant phenomena in conversational speech. While code-switching is regarded as an inter-sentential alternation between two languages, code-mixing is a word-level embedding of one language in the matrix of another. The phenomenon holds particular relevance in speech communities where the mother tongue and the medium of instruction are different languages. In India, English has been accorded an official language status by the constitution, and is maintained widely as a medium of education in schools. As English has maintained a ubiquitous status of the prestige language, Indian bilingual speakers show abundant usage of mixing and switching between their regional language and English. The scope and relevance of this phenomenon serves as reason for shifting the tradition from monolingual automatic speech recognition (ASR) studies and accommodating instead, for code-mixed speech recognition. The paper presents an automatic speech recognition system which is trained on a combination of monolingual Hindi and code-mixed Hindi-English speech. The testing has been conducted on 1 code-mixed speaker, and 3 monolingual Hindi speakers. We observe that such a mixing enhances the phonetic coverage of the acoustic model used for training.

The design of the paper is as follows: Section 2 describes the speech corpus used for experimentation. Section 3 provides a brief introduction to DNN based acoustic modeling, and an overview of trends in the acoustic modeling of bilingual speech recognition. Similarly, Section 4 outlines the concept of language models and outlines their implementation in code-mixed speech recognition. Section 5 presents the results, and Section 6 concludes the paper.

II. Description of the test corpus

This section outlines the development of a phonetically balanced speech corpus of code-mixed Hindi-English speech. The section also presents details about the monolingual Hindi speech corpus, which is used in combination with the code-mixed corpus to train the acoustic model.

A. Development of the code-mixed corpus

Although the dynamic nature of code-mixing is best captured in informal spoken communication, there is a significant body of print media employing code-mixed diction. Portions of newspaper dedicated to technology, sports, gadgets and fashion trends report frequent word-level insertions. The Phonetically Balanced Code Mixed (PBCM) corpus is a set of 2,694 sentences, sampled from a Large Code Mixed (LCM) corpus of 23,666 sentences. The LCM has been mined from the popular Hindi newspaper DainikBhaskar (http://epaper.bhaskar.com/). To ensure domain independent phonetic diversity, three different sections (Gadgets & Technology, Lifestyle and Sports) of the newspaper were chosen.

Example:
पार्ट्स खरीद कर टेक्नीशियन से बदलवा सकते हैं ।

Gloss:
[parts-ENG] [buy-HIN] **[technician-ENG]** [case

marker-HIN] [change(causative)-HIN] [can-HIN]

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

The example above represents the word level English insertion in the matrix of a Hindi sentence.

After the necessary steps of pre-processing, the LCM corpus was carefully sampled into the PBCM corpus. While Sports and the Gadgets and Technology section have prominent technical vocabulary borrowings, this content is not always limited to lack of parallel vocabulary in the matrix language. Figure 1 displays the lexical diversity in the respective genres.

The PBCM corpus was recorded by 2 male and 2 female volunteer speakers, in a professional recording studio environment. The utterances were recorded at a sampling rate of 44.1kHz, but were downsampled to 16 kHz for the ASR experiments. A silence of 1 second was appended to each of the utterances, both in the beginning and the end of the utterance.

B. Desciption of the training dataset

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 duration of the training dataset is 7.5 hours (with 3 hours of monolingual speech and 4.5 hours of code-mixed speech). The testing dataset also follows a similar structure, with a duration of 2 hours. (1 hour monolingual and 1 hour of code-mixed speech). The acoustic model, described in the next section, was developed on such an arrangement.

III. The acoustic modeling component

In processing of code-mixed speech, several ideas for acoustic modeling have been put forward. An acoustic model establishes statistical relationship between speech segments and the corresponding text.

In general, let $O = \{x_1, ..., x_T\}$ be the acoustic observations and $w = \{w_1, ..., w_T\}$ be the corresponding word sequence. Then the DNN must learn the conditional distribution of words given acoustic observations, represented by $p(w|O)$. DNN acts as a discriminative classifier which classifies the tied-state phoneme classes (*senones*) given the acoustic observations O. The acoustic model decodes the 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 the correct hypothesis and boost the accuracy of the overall system.

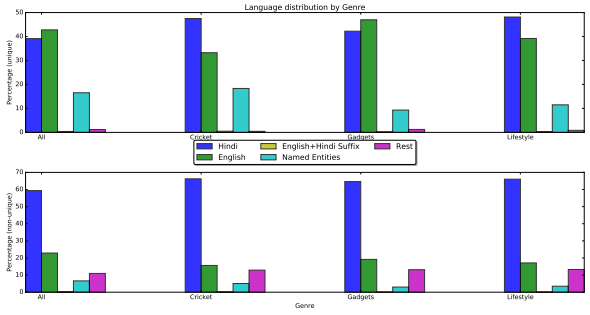

Figure 1. Stacked graphs displaying the Unique (above) vs Non-unique (below) frequency of distribution of English embeddings, Named-Entities and Rest contained in the PBCM corpus.

Traditionally, there have been two methods of approaching multilingual ASR studies. In the first approach, a language identification system is implemented at the front-end. [1] Word-level speech segments are identified based on their language, and then sent to their respective monolingual speech recognizers for phoneme decoding. Usually, such multipass frameworks suffer from error-propagation from the language identification front-end to the speech recognizer system at the back-end. To combat the error-propagation, the second approach usually does away with the word-level language identification component. [2] Bhuvanagiri [3] et al present a one-pass speech recognition approach for a code-mixed corpus of 225 spoken utterances. They exploit the resources already made available for large-scale speech recognition, such as the acoustic model trained on monolingual English corpus. An approximation of missing Hindi phonemes is achieved using a combination of existing English phones (from CMU dictionary). Similarly, Fung et al [4] develop three sets of language-independent phone merging, and use existing monolingual corpora to develop a code-mixed speech recognition system. More approaches to combining phonesets can be seen in an interpolation of two monolingual speech corpora, and a combined phoneset. [5].

In the present study, we use language independent phones, as described in the succeeding section.

A. Language independent phones

Both in multilingual and code-mixed speech recognition, a global phoneset development is a typical method for expanding phonemic coverage. Phones of low-resource language are combined or mapped to the closest approximations of phones in a high-resource language through various adaptation methods [3], [4] or clustering techniques based on monolingual decoding of target language utterances. [6] In the PBCM corpus, the English insertions are predominantly in Devanagari. However, some sentences do contain word-insertions in Roman script. To ensure phonetic con-

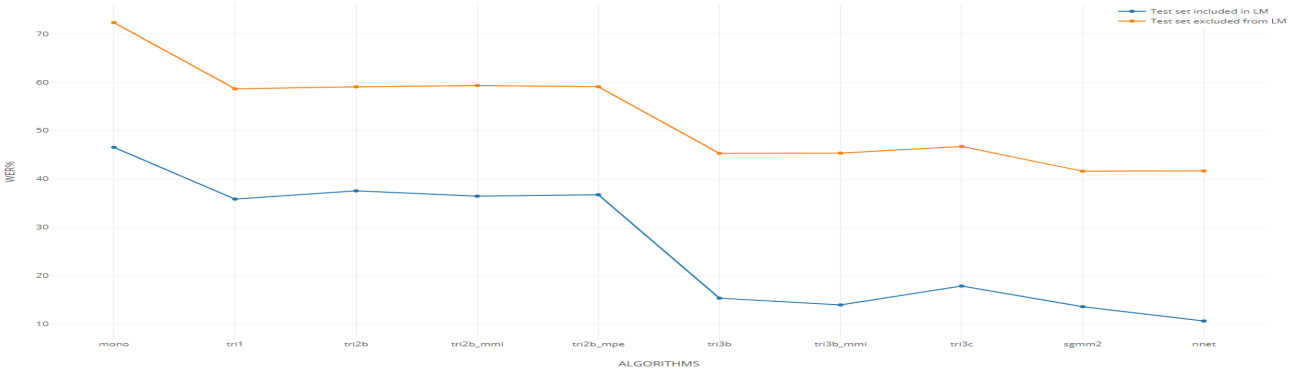

Figure 2. Word Error Rate (WER) evolution plotted against the set of acoustic modeling algorithms in Kaldi. The yellow curve plots the WER in Expt 2 (without testing prompts) and the blue curve plots Expt 1 (with testing prompts).

sistency among all the transcripts, we use automatic transliteration [7] to convert the words in Roman script into their respective Devanagari representation. The Devanagari was then converted to its corresponding WX notation.

B. Feature selection and extraction

We propose to use WX notation as the phoneset because it provides better coverage for the code-mixed phones. As has been described above, the corpus contains transliterated English words adapted through the monolingual phonetic representation of Hindi. Experimenting with WX representation of English phones is advantageous to most Indian languages, as code-mixing with English is a prevalent phenomenon across India.

The DNN model is trained over features obtained initially by concatenating ± 4 frames of MFCC 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.

As described in Section II, the training dataset is rich in acoustic variability in containing speech recordings from 20 different speakers. The existing code-mixed data has been supplemented with this resource, and the testing has been performed.

IV. The language modeling component

Language models are frequently used in ASR studies to provide word-level probability scores derived from the sequential structure of sentences. Standard n-gram based models are developed on the assumption that the probability of a given word $p(w_t)$ can be accurately predicted based on the history h_t of the word-sequence that immediately precedes it.

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

In recent literature, Vu et al [8], approach acoustic modeling through modifying and refining the language model. Grammatical constraints (equivalence and government constraint) are modeled and implemented in [9], to predict the well-formedness of a sequence. In another approach, class-based language models [10], [11] are used to circumvent the limitations of low-resources in data. In following the dynamic nature of code-mixing, remarkable biases in code-mixing have been identified, where code-mixing has been characterized as a speaker-specific phenomenon. Language models adapted by [12] to model speakers-specific attitudes, display a significant increase in accuracy.

For the purpose of this study, we use a trigram based language model, developed on the PBCM corpus concatenated with the monolingual Hindi corpus. Details of the experimental setup are presented in the succeeding section.

The PBCM corpus, having been extracted from a popular newspaper does not suffer from a dearth of corpus for language model. Combining the monolingual acoustic model with a large code-mixed language model can be considered for future work.

V. Results and discussion

Acoustic models were trained according to Dan's NNET2 setup [13].

We conduct two sets of experiments with respect to the language model.

- **Expt 1:** The speech utterances that had been covered in the spoken corpus were included in the language model training. This setup was designed so as to remove any instance of out-of-vocabulary words, and purely testing the performance of a largely monolingual acoustic model.
- **Expt 2:** In this experiment, we exclude the speech utterances that have been reported as prompts

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 I
TABLE WITH WORD ERROR RATE (WER) OF DIFFERENT ACOUSTIC MODELS IMPLEMENTED WITH THE TWO LANGUAGE MODELING SETUPS

of the testing corpus. The design of this setup allowed us to evaluate the ASR based on the largely monolingual acoustic model, and a purely monolingual language model.

As a result of Expt 2, the WER obtained over the mixed (3 monolingual, 1 code-mixed) test set evolved from 72.34% for monophone training to 41.63% for Dan's NNET2. Canceling out the out-of-vocabulary words significantly reduced the WER, as the monophone baseline started much lower at 46.54 %. Figure 2 displays the evolution of WER obtained per system as we trained using Kaldi. We can clearly notice the decreasing WER as we progress the training from monophone training towards subspace Gaussian mixture models and further, to neural network based models. The word error rate for the neural network based decoding for Expt 1 returns the lowest WER, i.e. 10.63 %. Table 1 displays the different acoustic modeling algorithms in combination with the two different setups of language modeling.

VI. Conclusion

In this work, we present the development of an automatic speech recognition system for code-mixed read speech, through mixing the available monolingual corpora (17 speakers/ 3 hours) with some amount of code-mixed speech (3 speakers/4.5 hours). The testing dataset follows a similar structure, comprising of data from 3 monolingual and 1 code-mixed speaker. A set of acoustic models have been implemented in combination with a language model, in two separate experimental setups. The word error rate for the neural network based setup has reported the best results in a neural network based framework, primarily when the language model contains the test utterances. The word error rate in this case has been found to be 10.63 %. The lowest word error rate in the second experiment is reported to be 41.6%.

References

[1] C.-H. Wu, Y.-H. Chiu, C.-J. Shia, and C.-Y. Lin, "Automatic segmentation and identification of mixed-language speech using delta-bic and lsa-based gmms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 266–276, 2006.

[2] D.-C. Lyu, R.-Y. Lyu, Y.-c. Chiang, and C.-N. Hsu, "Speech recognition on code-switching among the chinese dialects," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[3] K. Bhuvanagiri and S. Kopparapu, "An approach to mixed language automatic speech recognition," *Oriental COCOSDA, Kathmandu, Nepal*, 2010.

[4] M. C. Yuen and F. Pascale, "Using english phoneme models for chinese speech recognition," in *International Symposium on Chinese Spoken language processing*. Citeseer, 1998, pp. 80–82.

[5] J. Y. Chan, H. Cao, P. Ching, and T. Lee, "Automatic recognition of cantonese-english code-mixing speech."

[6] Y. Li, P. Fung, P. Xu, and Y. Liu, "Asymmetric acoustic modeling of mixed language speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5004–5007.

[7] I. A. Bhat, V. Mujadia, A. Tammewar, R. A. Bhat, and M. Shrivastava, "Iiit-h system submission for fire2014 shared task on transliterated search," in *Proceedings of the Forum for Information Retrieval Evaluation*, ser. FIRE '14. New York, NY, USA: ACM, 2015, pp. 48–53. [Online]. Available: http://doi.acm.org/10.1145/2824864.2824872

[8] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4889–4892.

[9] Y. Li and P. Fung, "Code-switch language model with inversion constraints for mixed language speech recognition." in *COLING*, 2012, pp. 1671–1680.

[10] C. F. Yeh, C. Y. Huang, L. C. Sun, and L. S. Lee, "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*. IEEE, 2010, pp. 214–219.

[11] T.-L. Tsai, C.-Y. Chiang, H.-M. Yu, L.-S. Lo, Y.-R. Wang, and S.-H. Chen, "A study on hakka and mixed hakka-mandarin speech recognition," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 199–204.

[12] N. T. Vu, H. Adel, and T. Schultz, "An investigation of code-switching attitude dependent language modeling," in *International Conference on Statistical Language and Speech Processing*. Springer, 2013, pp. 297–308.

[13] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.