An Approach to Mixed Language Automatic Speech Recognition

Kiran Kumar Bhuvanagiri, Sunil Kumar Kopparapu
TCS Innovation Labs - Mumbai
Yantra Park, Pokharan Road 2, Thane(West), Maharastra, INDIA
{KiranKumar.Bhuvanagiri,SunilKumar.Kopparapu}@TCS.Com

Abstract

Use of mixed language in day to day spoken speech is becoming common and is being accepted as being syntactically correct. However recognition of mixed language spoken speech is a challenge to a speech recognition engine. Though sparse, there have been studies on how to enable recognition of mixed language spoken speech. At one extreme is to use acoustic models of the complete phone set of the mixed language to enable recognition while on the other extreme is to use a language identification module followed by a language dependent speech recognition engine to recognize mixed language. Each of this has its own implications. In this paper, we approach the problem of mixed language recognition by constraining ourselves to use readily available resources and show that by (a) suitably modifying the language model to use mixed language and (b) by constructing a pronunciation dictionary, one can achieve a good recognition of mixed language spoken speech.

1. Introduction

Mixed language arises through the fusion of two or more, usually distinct, source languages, normally in situations of thorough bilingualism, so that it is not possible to classify the resulting language as belonging to either of the language families that were its source [17], [1] [2]. With urbanization and geography shift of people the ability to converse simultaneously in many languages is becoming common. A very large population use mixed language in everyday conversation without actually being aware of its usage, especially the young urban. Though mixed language is defined as a mixture of two distinct languages void of the knowledge of which language is mixed into which, at least in the Indian context, the native language is the primary language and the non-native language (usually English) is the mixed or the secondary language. Primary language can be loosely defined as that language in the mixed language which is spoken in majority. In other words there are a majority of words from that language in a given sentence and a relatively smaller number of words are from the secondary language. One can observe that often the words uttered in the secondary language are *keywords* or *foreign* words or phrases which have colloquial acceptance. As a result, the rate of language change within a spoken sentence is very frequent. Thus recognition of mixed language speech requires in our opinion an entirely different approach.

Consider a human agent based inquiry service in a metropolitan city. which has to cater to people speaking different languages. In such a scenario, the agent needs to be able to converse (understand and reply) in multiple languages which is very unlikely. A possible solution can be to ascertain the language of the speaker and then direct the call to an agent who can converse in that language expertly. Similarly, a speech solution for multiple languages can be built by developing separate recognition engines of each language. Having identified the language of the speaker, the speaker could be directed to that language specific recognition engine. Clearly, this system though can address multiple languages, it cannot work in the scenario where people used mixed language speech even if one knew the specific mix of the languages in use because the language segments are short and the change very frequent. Recently there has been an increased interest in mixed language recognition (for example [2][3]) although the work in literature is restricted to a mix of Mandarin and Taiwanese. As such work in mixed language speech recognition is in its nascent stages of research and to the best of our knowledge there is no work reported in literature for India specific language mix.

There are two major frameworks to build mixed language automatic speech recognition (ML-ASR). One is the multi-pass framework while the other is a one-pass framework. Typically in a multi-pass ML-ASR, the exact instances in spoken speech where language switch occurs is determined and the language of the speech segment is identified. Once the segment of speech and its language is found, a corresponding language dependent ASR is used to decode or recognize the speech segment. On the other hand in a typical one-pass approach, an acoustic model, pronunciation dictionary and language model

are built to encompass both the languages in the mixed language. This enables recognition on mixed language speech. This is relatively a simpler approach compared to the multi-pass approach because (a) there is no need to specifically identify the language and (b) employ several language specific ASRs. However one-pass approach to ML-ASR poses problems in the form of a need to collect sufficiently large mixed language speech and mixed language text corpus which can be used to build the acoustic models and the language model (LM) required for ML-ASR In this paper, we hypothesize that one could use the available resources (for example acoustic models for one of the languages that is part of the mixed language) and carefully construct the pronunciation dictionary and the language model to enable a ML-ASR. We conduct a number of experiments on mixed language speech where the primary language is Hindi and the secondary language is English. It should be noted that this approach can be used as it is for any other Indian languages taking the place of Hindi in our experiments by appropriate mapping of the phone in that language to English phones.The rest of the paper is organized as follows. A short review on multi-pass and one-pass frameworks is discussed in Section 2, followed by a discussion on the corpus used in our experiments highlighting our approach in Section 3. In Section 4, we discuss the experimental results and conclude in Section 5.

2. ML-ASR Literature Review

Recognition of mixed language speech is still in its initial stages of research. There are primarily two approaches reported in literature. One being multi-pass framework [4] and other is the one-pass framework [3]. Multilingual speech recognition, is an area of research which has closely related to ML-ASR. In multilingual speech recognition, the spoken speech is although in a single language the main challenge is that one does not know *apriori* the identity of the language that is being spoken. So the first task in multilingual ASR is to identify the language. This problem of identifying language is well addressed in literature [5] for almost two decades now. Cimarusti et al [5] used LPC based acoustic features to identify language on eight different languages with reasonable success while Foil [7] used prosodic features for language identification. In 1992, Nagawaka [8] compared four different methods and concluded that continuous HMM based method works best for language identification. Later in 1995 Yan [9] applied acoustic, phonotactic and prosodic information for language identification. Naratil and others [10] successfully used phonotactic-acoustic features to identify language. Many recognizers like Gaussian Mixture Model (GMM), single language phone recognition followed by language modeling (PRLM), parallel PRLM (PPRLM), GMM tokenization [6] and Gaussian Mixture Bi-gram Model (GMBM) [11] have also been studied in literature for multilingual speech recognition. To use multilingual approaches in mixed language speech recognition, one needs to find the exact time instants at which a switch from one language to another occur and follow it up with language identification. Automatic segmentation of speech of different languages within an utterance had been addressed by Chung-Hsien Wu et al. [4]. They apply Bayesian Information criteria (BIC) on Δ-MFCC features of each frame and group frames based on the scores. In another work, Chi-Jiun et al, [12] used statistical approach to segment and identify language boundaries and language identification. They used MAP estimate to find the boundary segments and latent semantic analysis based GMM with VQ based bi-gram language model to do language identification.

Mixed language speech recognition using multi-pass framework can be realized using the following steps. The mixed language speech input is divided into segments by identifying instants at which change in language occurs. Then each segment is mapped to a corresponding language using a language identification module. Then a language dependent speech recognizer is used to decode that particular segment of speech. These three steps are shown in Fig 1. The recognition performance of multi-pass approach depends on (a) performance of the language boundary detection, (b) language identification block and (c) the actual performance of the language dependent ASR. So a poor performance by any one of the three blocks affects the overall performance of the ML-ASR system. The one-pass framework [3] avoids this drawbacks of a multi-pass system by building a pronunciation dictionary, *super*[1] acoustic models and a language model to encompass both the languages . Super acoustic model is an acoustic model generated for the combined phone set of the languages in the mixed language. The advantage of this approach (shown in Fig. 2) is that it is not inhibited by language boundary detection and language identification blocks. It is direct and simple (along the lines of an single language ASR) as we build acoustic models, pronunciation dictionary and language models for the mixed language. However this approach needs explicit access to mixed language speech and text corpus.

1 Meaning a super set of phonemes of both the languages forming the mixed language

In our approach, we used one-pass framework however we constrained ourselves to use the acoustic phoneme models of a single language (which was readily available) instead of trying to undertake the herculean task of collecting speech corpus and transcribing it to build acoustic models for the super phone set which encompasses both the languages. We however built a small database of mixed language speech (to test our approach) and text corpus to (construct the language model to handle mixed language recognition [16]). We describe our approach next .

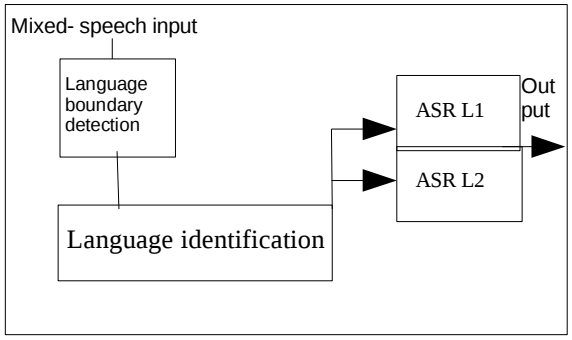

Figure 1 Multi-pass Framework

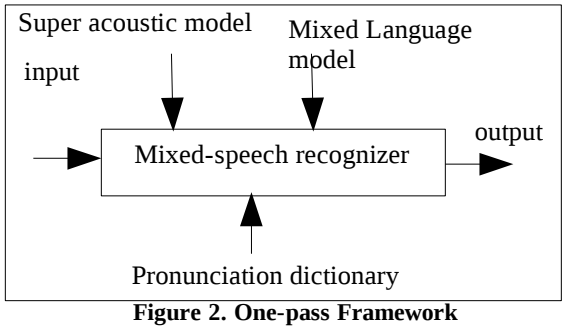

Figure 2. One-pass Framework

3. Our Approach

We have worked on a specific language mix, namely, Hindi-English whose usage is very common in the Indian subcontinent. Specifically Hindi being the native language, is spoken in majority and English is the secondary language. In our corpus we found that a little more than two thirds of the total words were spoken in Hindi while the rest were either spoken in English or were proper nouns. Overall, our database came from 46 different speakers (with sufficient gender and age variability and the speakers came from different metros in India). Each of the speaker uttered three to five different sentences of which at least one sentence uttered by the speaker was elicited speech. The elicited speech gave an indication of the actual mix of the language as spoken in everyday conversation. In all there were 225 different spoken sentences from 46 speakers. In all there were 2071 words. All our discussion and experimental results are based on this corpus. During data collection, the speakers were supplied a speaker sheet (in Hindi script) and were asked to call from a quite environment and the recording was done using a telephony card (Dialogic). All the speech was recorded at 11 kHz and 8 bits per sample using a home grown data collecting application.

Our approach was one-pass and did not require us to segment the speech into language based segments. Further in all our experiments we used the public domain speech recognition engine, Sphinx [15], with the WSJ acoustic (English phones) models which came with the speech engine. The reason for using acoustic models of English was (a) its ready availability for use and (b) building acoustic models for Hindi or mixed language was too cumbersome requiring actual collection of large amount of data to which we did not have access. Ideally as reported in literature Hindi ASR requires 62 phonemes while English requires only 39 phonemes, we approximate those Hindi phonemes which are not in English by a combination of two or more approximate English phonemes [13]. The lexicon or the phonetic dictionary that supports the ASR is constructed using the CMU language toolkit [14] for the English words in the corpus. All Hindi words were first transliterated into English and the pronunciation of this English word was obtained using [14]. So we had all the words in the mixed language expressed using only the English phone set. This allows us to use the default acoustic models that comes with Sphinx [15] in our experiments.

4. Experimental Results

We conducted three sets of experiments to evaluate the performance of our approach to ML-ASR. All the three sets of experiments used the Sphinx ASR [15] and the language model (mixed language model) generated from the mixed language speech corpus that we collected. However the manner of construction of the phonetic lexicon is different in each of the experiments. The distribution of words in the corpus was 62% from Hindi, and 28 % from English and the remaining 10 % of the words were proper nouns. In the first set of experiments (Expt 1), the construction of the phonetic lexicon is done using the CMU dictionary [14] for all the words. While in the second experiment (Expt 2) the phonetic mapping of all the English words and the proper nouns is done using CMU dictionary [14], while the Hindi words are mapped using an approximated English phoneme(s).

In the third experiment (Expt 3), the phonetic lexicon of English words is created using CMU dictionary [14] while both the proper nouns and the Hindi words are done using approximated phoneme set. As mentioned earlier for all the experiments we used Sphinx [15] with WSJ acoustic models in the same configuration. We have presented word error rates (WER) on Train database (Table 1) and Test database (cross validation on three sets) in Table 2. In case of Train data, we generated LM from transcriptions of train database, while performing experiments on Test data, we generated LM from complementary sets. It can be seen that the overall WER of the ML-ASR is less when the pronunciation lexicon for the Hindi words and proper nouns is built using the approximated English phones (Expt 3: Train-47.39%, Test-53.73%) compared to pronunciation lexicon built using CMU dictionary (Expt 1: Train-65.04%, Test-76.16%) and (Expt 2: Train-48.58%, Test-54.91%). This suggests that representing non-English words using approximate English phonemes decreases WER without actually having to construct a mixed language phoneme set and building the super acoustic models. Also note that the performance in Expt 1 even for English words is far poor than the performance in Expt 2 or Expt 3 compare (52.45% to 37.62% and 35.81% on Train database), (57.11% to 38.63% and 37.79% on Test database). This is essentially because of the non perfect representation of Hindi (or proper nouns) words in Expt 1 which results in misrecognition of English words preceding or succeeding the Hindi words (3-gram representation of the mixed language in the LM that we used).

Table 1. Train-data Results in word error rate

	English words	**Hindi words**	**Proper nouns**	**Overall**
Expt 1	52.45%	69.62%	71.64%	65.04%
Expt 2	37.62%	51.25%	66.91%	48.58%
Expt 3	35.81%	52.56%	41.84%	47.39%

Table 2. Test-data Results in word error rate

	English words	**Hindi words**	**Proper nouns**	**Overall**
Expt 1	57.11%	70.28%	80.86%	76.16%
Expt 2	38.63%	50.67%	71.28%	54.91%
Expt 3	37.79%	51.12%	59.52%	53.73%

5. Conclusion

Mixed language automatic speech recognition (MLASR) is gaining increasing popularity because of its wide spread use and more importantly its acceptance in the society. In this paper we have shown a usable approach to enable mixed language speech recognition by making use of the available resources (acoustic models) and (a) carefully constructing the pronunciation dictionary for the mixed language words and (b) constructing a mixed language model (MLM) from a small mixed language corpus. The advantage of our approach is that (a) there is no actual need to segment speech and identify a language which in conversational speech is very difficult because the switch from one language to another is very fast, (b) it does not require one to collect extensive speech data to construct the mixed language super acoustic models.. It should be noted that this approach can be used as it is for any other Indian language taking the place of Hindi in our experiments by appropriate mapping of the phone in that language to English phones.

6. References

[[1]Chien-Lin Huang and Chung-Hsien Wu., "Generation of phonetic units for mixed language speech recognition based on acoustic and contextual analysis". IEEE Transactions on Computers, 56:1225–1233, 2007.

[2] Po-Yi Shih, Jhing-Fa Wang, Hsiao-Ping Lee, Hung-Jen Kai, Hung-Tzu Kao, and Yuan- Ning Lin. "Acoustic and phoneme modeling based on confusion matrix for ubiquitous mixed language speech recognition", In SUTC '08: Proceedings of the 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, pages 500–506, Washington, DC, USA, 2008.

[3] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu, "Speech Recognition on Code-Switching Among the Chinese Dialects", of IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, May. 2006

[4]Chung-Hsien Wu,Yu-Hsein Chie, Chi Jiun Shia, Chun-Yu Lin „ "Automatic segmentation and identification of mixed language speech using Delta-BIC and LSA based GMMs", ICASSP 06, vol 14, No 1, 266-276.

[5] Cimarusti, D., Ives, R.B. "Development of an Automatic Identification System of Spoken Languages: Phase 1". Proc. ICASSP'82, pp. 1661-1664, May 1982.

[6]P. A. Torres-Carrasquillo, Elliot singer, Mars A Kohler, Richard J Greene, Douglas A Reynolds, and J R Deller Jr,, "Approaches to language identification using Gaussian mixture models and shifted delta ceptral features", in Proc.ICSLP'02, 2002, pp. 89–92.

[7]Foil, J.T. Language Identification Using Noisy Speech, Proc. ICASSP'86, pp. 861-864, April 1986.

[8]Nakagawa, S., Ueda, Y., Seino, T. "Speaker-independent, Text-independent Language Identification by HMM". Proc. ICSLP'92, pp. 1011-1014, October 1992.

[9]Yan, Y., "Development of an Approach to Language Identification Based on Language dependent phone Recognition." PhD thesis, Oregon Graduate Institute of Science and Technology, October 1995.

[10]Navrátil, J. Spoken Language Recognition–A step Toward Multilinguality in Speech Processing, IEEE Trans. Speech Audio Processing, vol. 9, pp. 678-685, September 2001.
[11]W.-H. Tsai and W.-W. Chang,, "Discriminative training of Gaussian mixture bi-gram models with application to Chinese dialect identification", Speech Commun., vol. 36, pp. 317–326, 2002.
[12]Chi Jiun shia, Yu-Hien Chiu, Jia-Hin Hieh, Chung-Hsien Wu, "Language boundary detection and identification of mixed language speech based on MAP estimation", ICASSP 04, vol 1, 381-384.
[13]Niloy Mukherjee, Nitendra Rajput, L V Subramaniam, Asish verma, On deriving A phoneme model for new language, proc ICSLP, 2000, pages 850-852.
[14]http://www.speech.cs.cmu.edu/cgi-bin/cmudict (last accessed Aug 2010)
[15]http://cmusphinx.sourceforge.net/ (last accessed Aug 2010)
[16]Sunil Kumar Kopparapu, "Voice Based Self Help System: User Experience Vs Accuracy", International Conference on Systems, Computing Sciences and Software Engineering: pages 101-105, 2008.
[17]http://en.wikipedia.org/wiki/Mixed_language (last accessed Aug 2010)