Recognition of Mixed Language Speech Without Language Identification

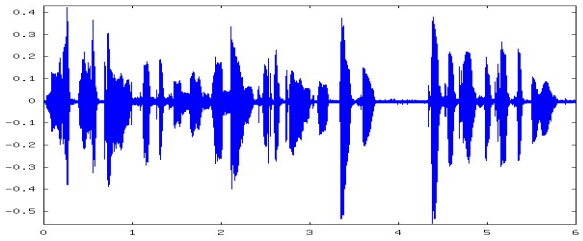
Kiran Kumar Bhuvanagiri, Sunil Kumar Kopparapu

***Abstract*— Use of mixed language in day to day spoken speech is becoming common and is being accepted as being syntactically correct. However machine recognition of mixed language spoken speech is a challenge to a speech recognition engine. There are studies on how to enable recognition of mixed language 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 language dependent speech recognition engines to do the recognition. Each of this has its own implications. In this paper, we approach the problem of mixed language recognition by using available resources and show that by suitably modifying the language model to use mixed language and construction of pronunciation dictionary one can achieve a good recognition of spoken mixed language.**

***Index Terms*— Speech Recognition, Mixed-language speech, language identification, phoneme set.**

I. INTRODUCTION

Mixed language arises through the fusion of two or more, usually distinct, Mixed 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 in many languages is becoming common. A very large population of people are using mixed language in everyday conversation without actually being aware of its usage. Use of mixed language is found to be very common in 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 is the mixed or the secondary language as shown in Figure 1. Primary language can be defined as that language in the mixed language which is spoken in majority. Meaning there are a majority of words from that language in a given conversation or sentence and a smaller number of words from the secondary language or the non-native language. One can observe that the words uttered in the secondary language are very often keywords or foreign words or phrases which are colloquially used. Subsequently, the rate of language change or shift is very frequent. Thus recognition of mixed language speech requires in our opinion an entirely different approach.



मैं अपने **account** से किसी दूसरे **bank** के **account** मे पैसा कैसे **Transfer** कर सकता हुँ ?

Fig 1 Mixed Language sentence

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 (at the least be able to 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 recognition based solution can be built. Having identified the language of the speaker, the speaker could be directed to a language specific recognition engine. Clearly, this kind of system cannot work in the scenario where people used mixed language speech even if one knew the mix of languages in use because the language shift is very frequent. Recently there has been increased interest in research communities on mixed language recognition (for example [2][3][19]) although the work has been 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 distinct frameworks to build mixed language automatic speech recognition (ML-ASR). One is the multi pass framework while the other is a one pass framework. In a multi pass ML-ASR, the exact instances in spoken speech where language switch happens is determined and the language of the speech identified. Once the slice of speech and its language is found, a corresponding language dependent (Automatic Speech Recognition) ASR is used to decode or

Manuscript received November 16, 2010.

Kiran Kumar Bhuvanagiri is with TCS Innovation Labs-Mumbai. Tata Consultancy Services, Yantra Park, Thane (West), Maharastra 400 60 (e-mail : kirankumar.bhuvanagiri@tcs.com)

Sunil Kumar Kopparapu is with TCS Innovation Labs-Mumbai, Tata Consultancy Services, Yantra Park, Thane (West), Maharastra 400 601 (e-mail : sunilkumar.kopparapu@tcs.com)

recognize the speech slice. On the other hand in the one pass approach, an acoustic model, pronunciation dictionary and language model are built to encompass both the languages in the mixed language. This enables ML-ASR on mixed language speech. This is relatively a simpler approach compared to 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 sufficient amount of mixed language speech corpus (and its text transcription) which can be used to build the mixed language acoustic and the mixed language language model required for ML-ASR In this paper, we hypothesize that one could use available resources (for example acoustic models for one of the languages that is one of the mixed language) and carefully construct the pronunciation lexicon and the language model to do a ML-ASR. We conducted 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 the approach is independent of the language mix in the sense that any other Indian language can take the place of Hindi. This will however require appropriate mapping of the phone in that Indian language to the English phones.

The rest of the paper is organized as follows. A short review on multi pass and one pass frameworks for multi lingual speech is discussed in Section II, followed by discussion on the mixed language database used in our experiments and highlighting our approach in performing mixed language ASR in Section III. In Section IV, we discuss experimental results and finally conclude in Section V.

II. ML-ASR Literature Review

Recognition of mixed language speech is still in its initial stages of research. There are two approaches reported in literature. One being multi pass framework [4] and other is the one pass framework [3]. Multi lingual speech recognition, is an area of research which has close relationship with ML-ASR. In multi lingual speech recognition, the spoken speech is although in a single language the main challenge is that one does not know a priori the identity of the language. So the first task in multi lingual 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 do language identification 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 indetification. Later in 1995 Yan [9] applied acoustic, phonotactic and prosodic information for language identification. Naratil and others [10] successfully used phonotactic-accoustic features. 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 the multilingual approaches in mixed language speech recognition, one needs to find exact time instants at which switching from one language to another occurs and follow it up with language identification. The 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 Delta-MFCC features of each frame and group frames based on the scores. In another work, Chi-Jiun et al, [12] use statistical approach to segment and identify language boundaries and language identification. They use 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 2. The recognition performance of multi pass approach depends on (a) performance of the language boundary detection and (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 multi pass based ML-ASR system. The one pass framework [3] avoids the drawback of multi pass system by building a pronunciation dictionary, super acoustic models and the language model to encompass both the languages in the mixed language. Super acoustic model is acoustic model generated for the combined phoneme set of the languages in the mixed language. Advantage of this approach (shown in Fig. 3) 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.

In our approach, we used one pass framework however we used 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 corpus to (a) construct the language model to handle mixed language recognition [16] and (b) to test our approach. We describe the approach next.

III. Mixed Language ASR Our Approach

We have worked on a specific language mix, namely, Hindi-English whose usage is very common in Indian subcontinent. Specifically Hindi being the native language, is spoken majority of time and English is the secondary language. In our corpus, a little more than two thirds of the total spoken words in the corpus were spoken in Hindi and the rest, namely, one third, being either English words or 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, which had a mix of two languages, 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 213 unique spoken sentences and 1946 words . All our experimental results reported in this paper are based on these words. 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, specifically we used a Dialogic CTI card. The speech was recorded at 11 kHz and 8 bits per sample using a home grown data collecting application.

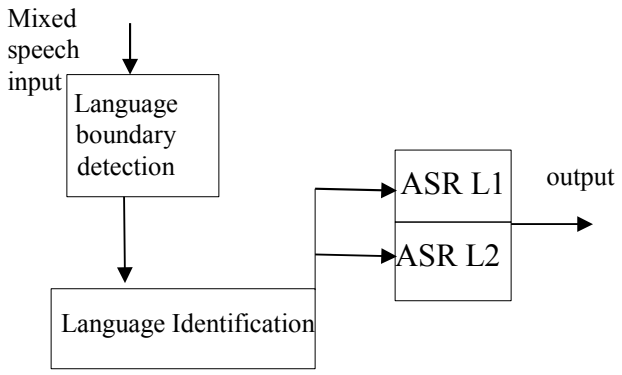


Fig 2. Multi pass Approach for mixed language ASR

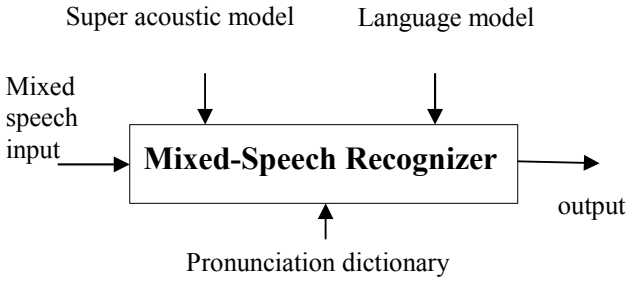


Fig 3. One pass Approach for mixed language ASR

Our approach is one pass, meaning there is no need to identify language segments unlike the multi pass approach.. It is apparent that multi pass approach require as many ASRs as the number of languages in mixed language speech, while one pass approach requires super phonetic model to be built. Our approach retains the framework of a one pass method with the use of modified approximate dictionaries. The use of modified lexicon enables us to avoid building of the expensive super acoustic models; further recognition can be performed with ASR of one of the languages. We used the public domain speech recognition engine, Sphinx [15], with the HUB4 (16 KHz) acoustic (English phones) model in one set of experiments and in another set of experiments we used the Hindi ASR [20] acoustic models (Hindi phones). The reason for using these acoustic models instead of acoustic models for mixed language was was (a) it was readily available for use and (b) building acoustic models for mixed language was too cumbersome requiring actual collection of large amount of data to which we did not have access. It should be noted that a Hindi ASR has 59 phonemes while English has only 39 phonemes. When using English acoustic models we approximate those phonemes (mainly occurring in Hindi words) which are not in English but in Hindi by replacing the phoneme in Hindi by a combination of two or more 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 are first transliterated into English and the pronunciation of this English word is obtained using [14] or approximate phoneme mapping. So we have all the words in mixed language expressed using only the English phone set This allows us to use the HUB4 English acoustic models[15] for ASR.

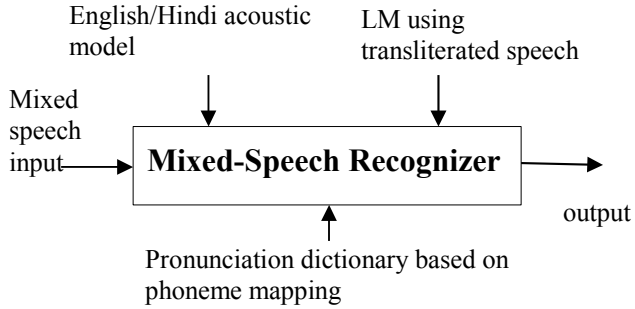


Fig 4 Proposed approach

IV. Experimental Results

We conducted three sets of experiments to evaluate the performance of our approach to ML-ASR using the English acoustic models and one experiment using the Hindi ASR. All the experiments used the Sphinx ASR [15] and the language model (mixed language model) generated from the mixed

language speech corpus that we collected[1]. In each of these experiments the manner of construction of the phonetic lexicon was different. The distribution of words in the corpus was 62% were Hindi words, and 28% were English words and the remaining 10% of the words were proper nouns. In the first set of experiments (Expt 1), the construction of the phonetic lexicon was done using the CMU tool kit [14] for all the words. While in the second experiment (Expt 2) the phonetic mapping of all the English words and the proper nouns was done using the CMU tool kit [14], while the Hindi words were mapped using an approximated English phoneme(s). By this we mean that, all the Hindi words are represented using only the English phoneme set. For example the word मत्स्यगंध (Matsyagandha) is represented using the CMU tool kit as *M AE TH S AY AH G AH N D* (see Fig 5(a)). While the equivalent pronunciation representation using Hindi phoneme set is *M A T A S Y A G A N DH A* (see Fig 5(b)). Using approximate phoneme mapping from Hindi phone set to English, the word मत्स्यगंध is represented as *M AH TH AH S Y AH G AH N DH HH AH* (see Fig 5(c)). *Note that i*n approximate phoneme mapping, each Hindi phoneme is replaced by an equivalent one of more English phones. For example the phone DH, occurring only in Hindi is substituted by the phones "DH HH" in English (see Fig 5) shows . In the third experiment (Expt 3), the phonetic lexicon of English words is created using CMU tool kit [14] while both the proper nouns and the Hindi words are constructed using the approximated phoneme set (Fig 5(c)).

(a)

```
ENG:BALANCE   B AE L AH N S
ENG:BANK   B AE NG K
ENG:IDENTIFICATION AY D  EH N T AH F AH K EY SH AH N
JANKAREE   JH AE NG K AH R IY IY
KRUPAYA   K R UW P EY AH
NAME:MATHSYAGANDH  M AE TH S AY AH G AH N D
```

(b)

```
ENG:BALANCE  B AA L A N: S A
ENG:BANK   B AA N K
ENG:IDENTIFICATION      EI  D  E N: T: I PH  I K EI SH A N A
JANKAREE       J AA N A K  AA R II
KRUPAYA       K  RX  P  A Y AA
NAME:MATHSYAGANDH       M A T A S Y A  G A N DH A
```

(c)

```
ENG:BALANCE   B AE L AH N S
ENG:BANK   B AE NG K
ENG:IDENTIFICATION AY D  EH N T AH F AH K EY SH AH N
JANKAREE   JH AA N  K  AA R IY
KRUPAYA   K R UH P AH Y AA
NAME:MATHSYAGANDH  M AH TH  S Y AH G AH N DH HH  A
```

Fig 5 Different dictionary constructions. (a) using CMU tool kit. (b) using Hindi phoneme set (c) using approximated phoneme mapping (from Hindi to English)

As mentioned earlier for all the experiments we used Sphinx [15] in the same configuration. We have presented word error rates (WER) on Train dataset (Table 1) and Test dataset (cross validation on three sets) in Table 2 separately. In case of the Train dataset the data used for constructing the LM is same as used for recognition while in case of the Test dataset the data used for constructing the LM was not part of the recognition experiments, namely the data used for LM construction and that used for recognition were 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-51.31%, Test-53.64%) compared to pronunciation lexicon built using CMU tool kit (Expt 1: Train-64.80%, Test-67.93%) and (Expt 2: Train-56.78%, Test-51.31%). This suggests that representing non-English (in this experiment Hindi words and proper names) words using approximate English phonemes decreases WER. Also note that the performance in Expt 1 even for English words is far poorer than the performance in Expt 2 or Expt 3, compare (51.69% to 45.51% and 41.58% on Train dataset), (53.31% to 47.45% and 44.61% on Test dataset). This is essentially because of the non perfect representation of Hindi (or proper nouns) words in Expt 1 which results in mis-recognition of English words preceding or succeeding Hindi words (we used 3-gram representation of the mixed language in the LM).

In the last experiment, we used Hindi ASR (16 Khz) [20]. The acoustic models consists of 59 phonemes. Pronunciation dictionary was constructed using Hindi phone set. In case of English words, dictionary is created on transliterated words. As Hindi phoneme (#59) set is a super set of the English phone set (#39) and primary language being Hindi, it can be observed that there is a decrease in WER. The percentage of the number of words correctly recognized is more using Hindi-ASR (60.23%) than Expt 3 using HUB4 acoustic models (55.96%). Further we observe an increase in Hindi words and proper names recognition when Hindi acoustic models are used.

TABLE I
Word Error Rate (Train dataset)

Experiments	Accuracies			
	English words (100-%correct)	***Hindi words(100-%correct)***	***Proper nouns(100-%correct)***	***Overall accuracy(WER)***
Expt 1	51.69%	51.94%	78.79%	64.80%
Expt 2	45.51%	47.99%	79.55%	56.78%
Expt 3	41.58%	44.69%	48.49%	51.31%
Hindi-ASR	45.89%	38.59%	34.10%	49.64%

[1] For the purpose of compatibility we re-sampled out speech recordings to 16KHz

TABLE II
Word Error Rate (Test dataset)

Experiments	Accuracies			
	English words (100-%correct)	***Hindi words(100-%correct)***	***Proper nouns(100-%correct)***	***Overall accuracy(WER)***
Expt 1	53.31%	53.19%	86.36%	67.93%
Expt 2	47.45%	46.71%	84.09%	57.29%
Expt 3	44.61%	46.29%	60.61%	53.64%
Hindi-ASR	49.50%	40.62%	45.31%	53.39%

V. Conclusion

Mixed language automatic speech recognition (ML-ASR) is gaining increasing popularity because of its wide spread use and more importantly its usage acceptance in the society. In this paper we have shown an 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 (LM) from a small mixed language text corpus. The advantage of our approach is that (a) there is no actual need to segment speech and identify the language which is conversational speech is very difficult because in mixed speech the switch from one language to another is very fast, (b) it does not require one to collect extensive speech data to construct the acoustic models to enable mixed language recognition. 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.

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)

[18] Kiran Kumar Bhuvanagiri, Sunil Kopparapu, "An approach to mixed language automatic speech recognition" , Oriental COCOSDA 2010,Nepal

[19] Imseng, David, Bourlard, Herve, Magimai-Doss and Matthew " Towards mixed language speech recognition systems", proceedings of Interspeech , sept 2010, Pages 278-281, Japan.

[20] www.Sourceforge.net/projects/hindiasr.