American Journal of Signal Processing 2012, 2(5): 92-97
DOI: 10.5923/j.ajsp.20120205.02

Mixed Language Speech Recognition without Explicit Identification of Language

Kiran Bhuvanagirir, Sunil Kumar Kopparapu*

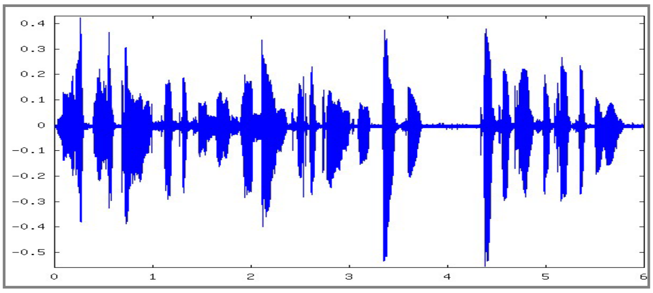
TCS Innovation Labs, Mumbai, Tata Consultancy Services, Thane (West), 400601, India

Abstract Use of mixed language in day to day spoken speech is becoming common and is accepted as being syntactically correct. However machine recognition of mixed language spoken speech is a challenge to a conventional speech recognition engine. There are studies on how to enable recognition of mixed language speech. At one end of the spectra is to use acoustic models of the complete phone set of the mixed language to enable recognition while on the other end of the spectra 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 speech recognition by using available resources and show that by suitably constructing an appropriate pronunciation dictionary and modifying the language model to use mixed language, one can achieve a good recognition accuracy of spoken mixed language.

Keywords Speech Recognition, Mixed-language Speech, Language Identification, Phoneme Set

1. Introduction

Mixed language, also termed as code switching in literature, 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 urbanisation and geography shift of people the ability to converse in many languages is becoming common. A very large number of people, especially urban youth, use mixed language in everyday conversation without actually being aware of it. Though mixed language is defined as a mixture of two distinct languages in equal proportion without giving away as to which language is mixed into which; at least in the Indian context, the non-native language (generally English words) is mixed into the native language. As shown in Fig. 1 the native language (Hindi) is the primary language and the non-native English language is the secondary language. Primary language can be defined as that language in the mixed language which is spoken in majority. 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 in mixed language. Thus recognition of mixed language speech requires, in our opinion, an entirely different approach.



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

Figure 1. Mixed Language sentence

Consider a call centre in a metropolitan city which has to cater to people speaking different languages. This requires all the agents in the call centre to be able to communicate in multiple languages which is very unlikely. A possible solution can be to ascertain the language of the caller and then, based on the language, direct the caller to an agent who can converse in that language expertly. In a similar vein, in a speech enabled application, having identified the language of the caller, a language specific speech recognition engine can be employed to cater to the caller. Clearly, this kind of system cannot work when people use 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 mixed language recognition (for example[3],[4],[19]) however the work has been restricted to a mix of Mandarin and Taiwanese. Mixed language speech recognition is in its nascent stages of research and to the best

* Corresponding author:
sunilkumar.kopparapu@tcs.com (Sunil Kumar Kopparapu)

Published online at http://journal.sapub.org/ajsp

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), namely multi pass and one pass framework. In a multi pass ML-ASR, the exact instance in spoken speech where language switch happens is determined and the language of the speech identified. Once the language of the speech segment is known, corresponding language dependent automatic speech recognition (ASR) is used to recognize the speech segment. Note that a typical ASR is language specific and uses acoustic model (AM), language model (LM) and a pronunciation lexicon (PL) built for that language to recognize spoken speech. The AM. LM and PL are constructed from language specific speech and text corpus through a training process. In the one pass approach, an ASR is built (namely, AM, LM and PL) which encompasses both the languages in the mixed language. This enables ML-ASR on mixed language speech. The one pass approach is simpler 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 (audio and the associated text transcription) which can be used to build the mixed language acoustic and the ML language model required for ML-ASR. In this paper, we hypothesize that one could use available resources (for example acoustic models of one of the languages in the mixed language) and carefully construct the LM and PL to do a ML-ASR. We conducted several 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 with appropriate mapping of the phone in that Indian language to the English phones.

The rest of the paper is organised as follows. A short review on multi pass and one pass frameworks for multi lingual speech is discussed in Section 2, followed by discussion on the mixed language database used in our experiments and highlighting our approach in performing mixed language ASR in Section 3. In Section 4, we discuss experimental results and finally conclude in Section 5.

2. Existing Approaches

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]. However, multilingual speech recognition is another area of research which has close relationship with ML-ASR. In multilingual speech recognition, the spoken speech is not a mix of two languages unlike ML-ASR, however 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]. Language identification using LPC based acoustic features was proposed by Cimarusti et al[5]. They were able to identify eight different languages with reasonable success. In another work, Foil[7] used prosodic features for language identification and Naratil et. al.[10] successfully used phonotactic-acoustic features. Later Yan [9] applied a combination of acoustic, phonotactic and prosodic information for language identification. Nagawaka[8] compared four different methods to identify languages and concluded that continuous hidden Markov model (HMM) based method works best. Many recognizers like Gaussian Mixture Model (GMM), single language phone recognition followed by language modelling (PRLM), parallel PRLM (PPRLM), GMM tokenization[6] and Gaussian Mixture Bi-gram Model (GMBM)[11] have also been studied in literature for multi lingual speech recognition.

In order to use the multilingual approaches in mixed language speech recognition, one needs to identify the exact time instants at which switching from one language to another occurs and follow it up with language identification. Automatic segmentation of different languages within a speech utterance had been addressed by Wu et al.[4] who use Bayesian information criteria (BIC) on Delta-MFCC (Mel Frequency Cepstral Coefficients). In another related work, Chi-Jiun et al.[12] use statistical approach to segment and identify language in a speech utterance. They use maximum a posterior (MAP) estimate to find the boundary segments to do language identification.

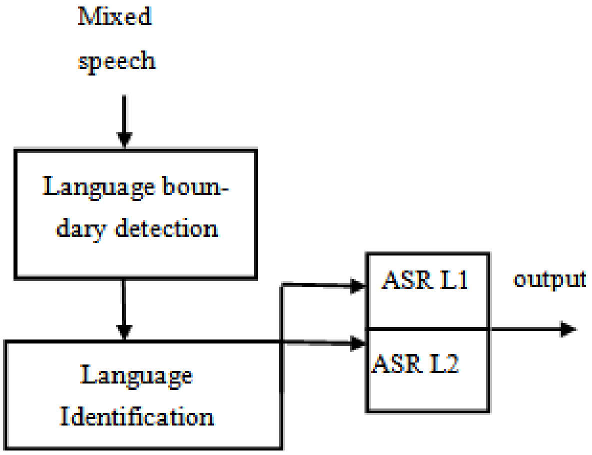


Figure 2. Multi pass approach for mixed language ASR

Mixed language speech recognition using multi pass framework can be realised using the following steps (see Fig. 2). The mixed language speech input is divided into segments based on identification of instant where language change occurs. Then the language of the segment is identified using a language identification module. Then a language dependent ASR is used to recognize that particular segment of speech. 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 specific ASR. Clearly 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 PL, AM and LM to encompass both the languages in the mixed language. The acoustic model for mixed language is an AM 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 dependent on the language boundary detection block or the language identification block. It is similar to a language specific ASR, except that the AM, LM and PL are built for the mixed language. Note that this approach needs mixed language speech and text corpus, which generally is not available. Clearly the existing approaches cannot be used for ML-ASR.

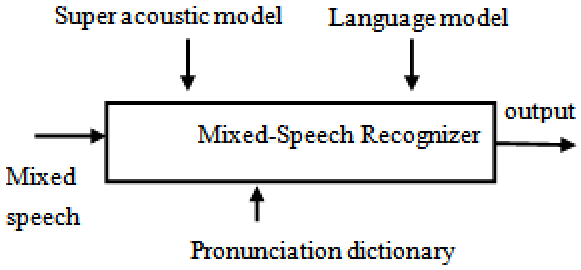


Figure 3. One pass approach for mixed language ASR

In our approach, we used the one pass framework however we used the AM 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 AM for the complete 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.

3. Proposed 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 majority of time compared to the non-native language English. 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 corpus consisted of 46 different speakers (with sufficient gender and age variability) from different metros in India. Each of the speakers uttered three to five different sentences, which had a mix of Hindi-English, 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 consisting of 1946 words. All the experimental results reported in this paper are based on these word utterances. 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.

Our approach retains the framework of a one pass method with the use of appropriate PL. The use of a modified PL enables us (a) avoid building an AM for the mixed language (note that mixed language speech corpus is difficult to collect) and (b) 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 (English phones) AM in one set of experiments and in another set of experiments we used the readily available Hindi ASR[20] AM (Hindi phones). The reason for using these AM instead of AM for mixed language was (a) these AMs were readily available for use and (b) building acoustic models for mixed language was too cumbersome requiring actual on the field collection of a large amount of speech corpus 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 by replacing the phoneme in Hindi by a combination of two or more English phonemes[13]. The PL that supports the ASR is constructed in the usual way by using the CMU language toolkit[14] for all the English words in the corpus. However, all the Hindi words are first transliterated into English and the pronunciation of this English word is obtained using[14] or approximate phoneme mapping (APM).

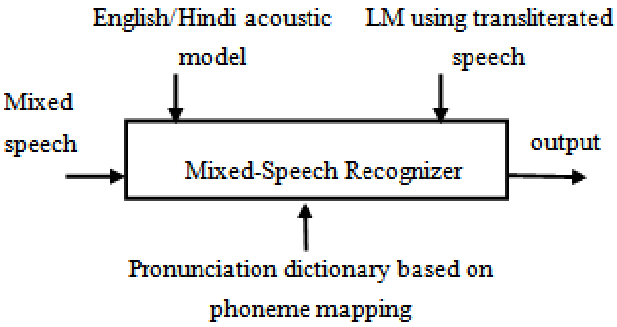

English/Hindi acoustic model
LM using transliterated speech
Mixed
speech
Mixed-Speech Recognizer
output
Pronunciation dictionary based on phoneme mapping

Figure 4. Proposed approach

4. Results and Discussion

We conducted, in all, a set of nine different experiments to evaluate the performance of our approach for ML-ASR. In the first set of experiments we used the English AM's while in the second set of experiment we used the Hindi language AM's.

In all our experiments we used the Sphinx ASR[15] and the well-known n-gram LM created using the mixed language speech corpus that we collected (Section 2). In each of these experiments the manner of construction of PL was different. The distribution of the Hindi, English and proper noun words in the corpus was 62%, 28% and 10% respectively. For the first set of eight experiments done using English AMs, we used two different methods of PL construction for the three different types of words, namely, English words, Hindi words and proper nouns. The first method of PL creation is based on the CMU toolkit[14] and the second method is based on approximate phoneme mapping (APM). In APM method of lexicon creation, a word is first transliterated and the equivalent Hindi phonemes are generated; each of these Hindi phonemes is then replaced by one or more equivalent English phonemes. For example the Hindi word मत्स्यगंध (Matsyagandha) is represented using the CMU tool kit as M AE TH S AYAH GAHN 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 APM the same word मत्स्यगंध is represented as म (M AH) त् (TH AH) स् (S) य (Y AH) गं (G AH N)ध (DH HH AH) (see Fig 5(c)). Note that in APM, a Hindi phoneme is replaced by one or more equivalent English phonemes. For example the phone DH, occurring only in Hindi is substituted by the phones "DH HH" in English (see Fig 5). For example, the English word "Identification" (आइडेंटीफिकशन) can be transliterated similarly as "aidentiphikation" and equivalent pronunciation using Hindi phoneme set is EI D E N: T: I PH I K EI SH A N A (see Fig 5(b)). Using APM, it is represented as AY D EH N T IY F IY K EY SH AH N (see Fig 5(c)). In the ninth experiment, we represented every word in the PL using both the alternative phonetic representations (see Fig 5(d)), namely using CMU and APM. Table 1 shows experiment number and the method used to construct the PL. For example, in "Expt 6" APM was used to construct the English and the Hindi words however CMU was used to construct the proper nouns. Pronunciation using CMU toolkit is denoted as CMU while, approximate phoneme mapping is denoted as APM

Table 1. Experimental Setup

Experiment	English words	Hindi words	Proper Nouns
Expt 1	CMU	CMU	CMU
Expt 2	CMU	APM	CMU
Expt 3	CMU	APM	APM
Expt 4	CMU	CMU	APM
Expt 5	APM	APM	APM
Expt 6	APM	APM	CMU
Expt 7	APM	CMU	APM
Expt 8	APM	CMU	CMU
Expt 9	APM+CMU	APM+CMU	APM+CMU

The configuration of the speech recognition platform Sphinx[15] was the same during all the experiments. We have presented word error rates (WER) on Train dataset (Table 2) and Test dataset (averaged over three rounds of cross validation) in Table 3 separately. In case of the Train dataset the textual data used for constructing the LM is same as the corresponding speech data used for recognition, while in the case of the Test dataset the text data used for constructing the LM was not part of the speech data used for recognition, in that sense the data used for LM construction and that used for recognition were complementary sets. It can be seen that the word accuracies for Hindi and proper nouns of the ML-ASR is higher when the PL for the Hindi words and proper nouns is built using the approximated English phones (Expt 3, Expt 5 and Expt 9) compared to when the PL is built using CMU toolkit (Expt 1, Expt 2, Expt 4, Expt 6, Expt 7 and Expt 8). We can conclude that representing non-English (in this case Hindi and proper nouns) words using approximate English phonemes decrease the WER. Overall WER is less when English words are represented by CMU toolkit and the Hindi words and proper nouns are represented using APM (Expt 3 and Expt 9) in the PL. Also note that the performance in Expt 1 and Expt 8 for English words is far poorer compared to performance in all other experiments, this can be attributed to the imperfect representation of Hindi (or proper nouns) words in Expt 1 and Expt 8 resulting in misrecognition of English words preceding or succeeding Hindi words or proper nouns (we used 3-gram representation of the mixed language in the LM).

Table 2. Word Error Rate (Train dataset)

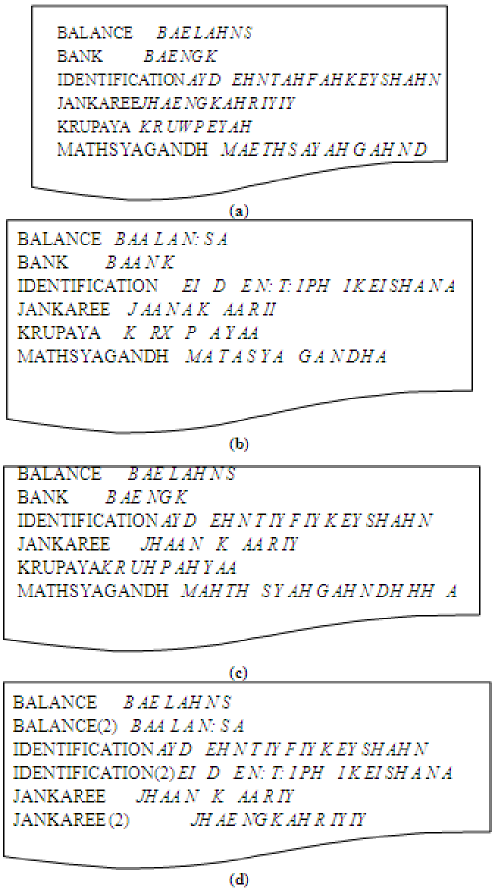
Experiments	Accuracies			
	English words (100-% correct)	Hindi words (100-% correct)	Proper nouns (100-% correct)	Overall accuracy(WER)
Expt 1	51.69%	51.94%	78.79%	64.80%
Expt 2	45.51%	47.99%	79.55%	56.78%
Expt 3	41.58%	44.69%	48.49%	51.31%
Expt 4	51.69%	51.94%	78.79%	64.80%
Expt 5	44.95%	45.43%	48.49%	51.75%
Expt 6	51.50%	47.90%	78.79%	58.43%
Expt 7	54.31%	52.27%	57.58%	62.90%
Expt 8	58.62%	52.35%	81.82%	66.86%
Expt 9	36.15%	27.38%	51.52%	40.24%
Hindi-ASR	45.89%	38.59%	34.10%	49.64%

In the last experiment (Hindi-ASR), we used Hindi AM (16 kHz)[20]. The AM consist of 59 phonemes and the PL was constructed using the Hindi phone set. English words in the lexicon are constructed by transliterating the English words. As Hindi phoneme (#59) set is a super set of the English phone set (#39) and the majority of the words spoken in the mixed language is Hindi, it can be observed that there is a decrease in WER in experiment Hindi-ASR as compared to experiments Expt 1 to Expt 8. The accuracy (100 – WER) of the number of words correctly recognized is more using

Hindi-ASR (60.23 %) than all the other experiments except for Expt 9 (68.43 %). Further we observe an increase in Hindi words and proper noun recognition when Hindi AM is used. As expected, the WER is better for the training set (Table 3) compared to the test set (Table 3) in all the experiments.

Table 3. 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%
Expt 4	53.08%	53.08%	86.47%	67.93%
Expt 5	49.53%	46.67%	63.30%	56.16%
Expt 6	52.71%	49.22%	86.47%	60.48%
Expt 7	55.33%	53.41%	60.90%	65.16%
Expt 8	60.18%	54.97%	88.72%	70.40%
Expt 9	40.74%	29.25%	64.66%	44.96%
Hindi-ASR	49.50%	40.62%	45.31%	53.39%



BALANCE *B AE L AH N S*
BANK *B AE NG K*
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*
MATHSYAGANDH *M AE TH S AY AH G AH N D*

(a)

BALANCE *B AA L A N: S A*
BANK *B AA N K*
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*
MATHSYAGANDH *M A T A S Y A G A N DH A*

(b)

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

(c)

BALANCE *B AE L AH N S*
BALANCE(2) *B AA L A N: S A*
IDENTIFICATION *AY D EH N T IY F IY K EY SH AH N*
IDENTIFICATION(2) *EI D E N: T: I PH I K EI SH A N A*
JANKAREE *JH AA N K AA R IY*
JANKAREE (2) *JH AE NG K AH R IY IY*

(d)

Figure 5. Sample lexicon constructions. (a) using CMU tool kit. (b) using Hindi phoneme set (c) using APM (from Hindi to English) (d) Both (CMU and APM) phonetic representations in same lexicon

5. Conclusions

Mixed language automatic speech recognition (ML-ASR) is gaining increasing popularity because of its wide spread use in everyday conversations and more importantly because of its acceptance in the society. While the best approach to build a ASR to recognize mixed language is to treat the mixed language as a language in itself and build AM, LM and PL as is done for a language specific ASR. This would involve an expensive and time consuming task of collecting a large amount of mixed language speech and text corpus and using this corpus to build AM, LM and PL for mixed language. Note that separate speech and text corpus has to be collected for each mixed language pair. In this paper we have shown an usable novel approach to enable mixed language speech recognition by making use of the available resources (English acoustic models, Hindi acoustic models but not the English-Hindi mixed acoustic models) and (a) carefully constructing a PL for the mixed language words and (b) constructing a 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 in most 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 corpus or data to construct the acoustic models to enable mixed language recognition. It should be noted that this approach can be applied to any other Indian language taking the place of Hindi; this would only require an appropriate mapping of the phones in that language to English phoneset.

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 Comm., 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 2012)

[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 2012)

[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 Matthew, "Towards mixed language speech recognition systems", Proceedings of Interspeech, Sept 2010, Pages 278-281, Japan.