Development of IIITH Hindi English Code Mixed Speech Database

by

Rambabu B, Suryakanth V Gangashetty

in

6th international workshop on spoken language technologies for under-resourced languages(SLTU'18)
(*SLTU-2018*)

Gurugram, India

Report No: IIIT/TR/2018/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2018

The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Development of IIITH Hindi English Code Mixed Speech Database

Banothu Rambabu and Suryakanth V Gangashetty

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad - 500032, India
rambabu.b@research.iiit.ac.in, svg@iiit.ac.in

Abstract

This paper presents the design and development of IIITH Hindi-English code mixed (IIITH-HE-CM) text and corresponding speech corpus. The corpus is collected from several Hindi native speakers from different geographical parts of India. The IIITH-HE-CM corpus has phonetically balanced code mixed sentences with all the phoneme coverage of Hindi and English languages. We used triphone frequency of word internal triphone sequence, consists the language specific information, which helps in code mixed speech recognition and language modelling. The code mixed sentences are written in Devanagari script. Since computers can recognize Roman symbols, we used Indian Language Speech Sound Label (ILSL) transcription. An acoustic model is built for Hindi-English mixed language instead of language-dependent models. A large vocabulary code-mixing speech recognition system is developed based on a deep neural network (DNN) architecture. The proposed code-mixed speech recognition system attains low word error rate (WER) compared to conventional system.

Index Terms: Code-mixing, Speech Recognition, Deep Neural Network.

1. Introduction

Code mixing also termed as mixed language is a word level embedding of one language into the matrix of another [1]. It arises by the fusion of two or more mixed source languages, normally in a situation where it is not possible to classify the resulting language as it belongs to either of the languages. In the multilingual and bilingual communities [2], code-mixing is a predominant phenomena in normal communication. People embed secondary language with primary language because of certain reasons as they want to feel the prestige of using english words [3] to exhibit a particular emotion [4] when they are short of words in one language, and to express in more appropriate manner. Different combinations of languages are found in code mixing. The ability of people to converse in many languages is becoming common due to geographical shift and urbanization. In India, there are 22 official languages [5]. Each state in India has a regional language and English being the medium of instruction at school or college level. All the official languages are spoken by millions of people. Different combinations of languages are found in code-mixing in India. For example, Telugu-English in Andhra Pradesh and Telangana, Tamil-English in Tamil Nadu, and Hindi-English in northern parts of India. In this paper, Hindi is the primary language and English is the secondary language also known as the embedded language. Although mixed language is defined as a mixture of equal proportions of two languages, in Indian reference the secondary language (English) is mixed with the primary language (Hindi).

In a mixed language, the language which is spoken more often can be considered as a primary language. The rate in shift of language among the mixed languages is frequent as the secondary language majorly consists of the foreign words or keywords [6]. As English is common among all the states, Indian bilingual speakers show extreme usage of mixing and switching between regional language and English [7]. Due to this reason, there has been a shift in tradition from mono-lingual Automatic Speech Recognition (ASR) [8] [9] [10] studies to the code-mixed speech recognition [11] [12]. In this paper, the ASR is trained on code mixed Hindi-English speech. The development of robust ASR has a setback from a sincere disadvantage, i.e., lack of data. The paper comes up with a large code-mixed phonetically balanced speech corpus and corresponding speech from 142 bilingual Hindi native speakers covered from different states of India as shown in (tabulated format) Figure 1. Since this corpus is a large one, it can be used for vivid applications such as speaker identification [11], emotional speech analysis [4], synthesizing mono-lingual sentences [13] from the multilingual speech recognition system [14].

The organization of the paper is as follows: Section 2 describes the scripts and sounds of Indian languages. Section 3 describes design of text data corpus from a large data that are pooled from public domain, i.e., internet. Section 4 describes the development of speech database from several Hindi native speakers and the other techniques used for data preprocessing. Section 5 introduces code-mixed ASR system using DNN models. Section 6 presents the results and discussion. Finally, Section 7 concludes the paper.

2. Scripts and sounds of Indian languages

One of the ancient scripts of Indian language origin is Brahmi script. Brahmi script is the source of scripts for some major languages in India, the basic sound units of Indian languages are refered to Aksharas [13]. The properties of Aksharas are as follows: An Akshara is a written form of basic speech sound labels for Indian languages. These akshara's are syllabic in nature with all allophonemic variations. The typical form of syllables are combination of vowels and consonants with many forms V, CV, CVC and CVCC, and thus have a generalized form of C*V. Here V denotes a vowel and C denotes a consonant. Languages spoken in India belong to several language families, north and eastern part of India follow Indo-Aryan language family and the major language in this family is Hindi. Hindi writing system uses Devanagari script. The southern part of India follow Dravidian language family, regionally Telugu, Tamil, Kannada and Malayalam are the major languages and all these languages have their own writing system. Allthough the wiriting system is different but these Indo-Aryan and Dravidian language family share common phonetic base, i.e., they share common speech sounds. In order to build any speech system like speech recognition and speech synthesis system, we need a common

10.21437/SLTU.2018-23

Table 1: *LHE and IIITH-HE-CM Corpus statistics*

Section	Corpus	Sentences	Words	Phones
Life style	LHE	25620	33220	67
	IIITH-HE-CM	2625	19760	52
Sports	LHE	25327	42180	65
	IIITH-HE-CM	2555	19920	52
Gadgets	LHE	24342	33015	65
	IIITH-HE-CM	2430	15020	57
Health	LHE	12645	15525	67
	IIITH-HE-CM	1194	10600	57

State names	No of speakers
Andhra Pradesh	1
Bihar	18
Chhatisgarh	1
Delhi	9
Gujarat	2
Haryana	8
Himachal Pradesh	2
Jammu and Kashmir	1
Jharkhand	3
Karnataka	2
Kolkata	4
Madhya Pradesh	18
Maharastra	7
Odisha	1
Punjab	6
Rajasthan	21
Telangana	3
Uttarakhand	5
Uttar Pradesh	30

Figure 1: *Speakers covered from different states in India*

phonetic label corresponding to all phoneme and allophonemic variations. In this paper we use Indian Language Speech Sound Labels (ILSL), for all Hindi and English sentences which contains 15 vowels and 42 consonants to cover all phones atleast once.

2.1. Indian Language Speech Sound Label

An ILSL is the one which covers all the variations and unique sound labels for speech sounds (developed by ASR/TTS consortia of TDIL, DIT, GoI). In order to assign a sound label for a particular phoneme, the ILSL involves the following steps 1) Similar sounds in different laguages are given same label. 2) IPA symbols refers to an example for Hindi, Telugu, Tamil and other languages but this is not an IPA chart of sounds of Indian languages. 3) The label set is designed such that the native script is largely recoverable from the transliteration. In this work, we are mostly concentrating on Hindi language and its script Devanagari, because the IIITH-HE-CM corpus contains most of the sentences which are written in Devanagari. Along side the English words are also written using Devanagari. All labels are in lower case even though the labels are case sensitive. Since the number of speech sounds are larger than the Roman alphabet, a system comprising suffixes, letter combination are used for labels. The Figures 2 and 3 show the common and unique ILSL sound labels corresponding to their respective Devanagari script. Some note on adding suffixes: 1) If speech sound is aspiration then use suffix 'h' to denote aspiration. 2) If speech sound is retroflex consonant, then use suffix 'x' to denote retroflex. 3) If vowel or consonant with Nukta / Bindu then use suffix 'q' to denote a nukta or bindu. 4) If the speech sound is nasalized vowel then use suffix 'n' to denote nasalization of a vowel. 5) For geminated sounds, use the label of geminated consonant as the label of the corresponding single consonant. 6) Mantras, diphthongs and halants exist in Hindi and Marathi languages and for these category of phonemes we use vowels and geminated vowels to represent their respective sound labels.

3. Design method of text corpus

In general communication between bilingual and multilingual speakers, they communicate through multiple code mixed language. However, there are print and public media currently exihibiting recurrent patterns of code mixing and code switching in their report. The print and social media content have the major code mixed sentences. These sentences from sports, electronic gadgets, fashion trends, technology, business columns, and political discussions show frequent code mixing and code switching at phrases, words, and morphemes of one language into another language. Hindi sentences have word-level mixing of English words. The lexical diversity in phoneme and allophone coverage of these languages are major concern in corpora design. The contents collected from different domains should cover all phonemes of Hindi and English. These phoneme frequency of occurence in intra-word and inter-word level is very important in building acoustic models for speech recognition system. We have proposed a design method of text corpus, in this proposed method we use a phonotactic [11] approach. The phonotactic approach allow us to compute permissable combinations of phones that can co-occur in Hindi language. The IIITH-HE-CM corpus contains the phonetically balanced sentences [7] with 8804 sentences and more than one million unique Hindi-English words. The phonological system in Hindi-English code mixed text corpus uses the ILSL labels. Details are given in Figures 2 and 3. The rules of suffixes, nukta or bindu, matras, diphthongs, and halant are described in Section 2.1.

3.1. Public domain text

Extraction of relevant information from large web data storage, that is online news papers of Hindi language. The large Hindi-English corpus named LHE corpus contains 9 million sentences. These sentences consists many symbols, html tags

and url to other links. We have mined the data and removed all unwanted symbols, html tags and url links. The numerals are converted into its equivalant Devanagari symbols. The LHE corpus and the IIITH-Hindi-English Code-Mixed named as IIITH-HE-CM phonetically balanced corpus with number of phones details given in Table 1.

3.2. Optimal Text selection

The text mining from the LHE corpus and phoneme sequences for unique code mixed sentences of IIITH-HE-CM corpus were generated using a grapheme to phoneme (G2P) converter trained on the ILSL sound labels of a bilingual lexical dictionary. The Hindi and English grapheme to phoneme sequence-to-sequence sound label characters were transformed into their respective ILSL sound labels. Phoneme sequences for unique Hindi-English words in the IIITH-HE-CM corpus were generated by converting them to their corresponding pronounciation sound labels. The total number of unique phones in the corpus, derived from the combination of ARPAbet and Hindi phoneme sequences of bilingual lexical dictionary. Intra-word and Word-internal triphones are collected for unique sentences and arranged based on the descending order of their frequency of occurrence.

IPA	ARPABET	ILSL	Devanagari
a	AH	a	अ
a:	AA	aa	आ
I, i	IH	i	इ
i:	IY	ii	ई
ʊ, u	UH	u	उ
u:	UW	uu	ऊ
-	-	rq	ऋ
-	-	rxq	ॠ
e:	EY	ee	ए
ɛ:	AE	ei	ऐ
o	OW	o	ओ
aʊ	AO	au	औ
k	K	k	क
kʰ	KH	kh	ख
g	G	g	ग
gʰ	GH	gh	घ
ŋ	MG	ng	ङ
ʧ	CH	c	च
ʧʰ	CHH	ch	छ
ʤ	JH	j	ज
ʤʰ	JHH	jh	झ
ɲ	NG	nj	ञ
ʈ	T	tx	ट
ʈʰ	TH	txh	ठ
ɖ	D	dx	ड
ɖʰ	DH	dxh	ढ
ɳ	-	nx	ण
t̪	T	t	त
t̪ʰ	TH	th	थ
d̪	D	d	द
d̪ʰ	DH	dh	ध

Figure 2: *Indian Language Speech Label ILSL set1.*

4. Development of speech database

Phonetically balanced sentences contain phonetic events based on frequency of occurences. After several experiments on developing speech recognition system , the need of phonetically balanced speech database is crucial. The design of speech database for implementing a speech recognition system became a challenging issue. Many attempts have been made to collect the speech databases in different languages Mandi database, Phonetic engine database, IIITH-Indic database. The Hindi-English database present in this paper consists set of unique sen-

n	N	n	ऩ
n	-	nq	ऩ़
p	P	p	प
pʰ	F	ph	फ
b	B	b	ब
bʰ	BH	bh	भ
m	M	m	म
j	Y	y	य
j	-	yq	य़
r	R	r	र
r	-	rq	ऱ
l	L	l	ल
ɭ	LH	lx	ळ
ɭ	-	lxq	ऴ
ʋ	V	w	व
ʃ	SH	sh	श
ʂ	SHH	sx	ष
s	S	s	स
ɦ	HH	h	ह
q	-	kq	क़
x	-	khq	ख़
ɣ	-	gq	ग़
z	-	jq	ज़
ɽ	-	dxq	ड़
ɽʰ	-	dzhq	ढ़
f	-	f	फ़
-	-	q	◌ँ
-	-	hq	◌ः
-	-	mq	◌ं
x	-	x	_

Figure 3: *Indian Language Speech Label ILSL set2.*

tences ranging from minimum 5 to maximum 15 words. These phonetically balanced sentences cover all recognition units for building speech recognition system. For a set of sentences the phonetic representation contains all phonemes in all possible contexts.

4.1. Speaker selection

The IIITH-HE-CM speech corpus is collected from 142 bilingual speakers (71 male and 71 female). These speakers belong to different parts of India (details are given in Figure 1), whose mother tongue is Hindi and the age group ranges between 18 to 35 years. The unique sentences (5-15 words) are 8804, these sentences are equally distributed among the speakers.

4.2. Speech recording

The code mixed sentences of IIITH-HE-CM corpus are recorded in a sound proof voice recording studio. The wave files are recorded with noise free and high fedility microphone, using audacity open source software. The recording wave files are sampled at 48 kHz, 24 bit resolution with PCM (Pulse Code Modulation) mono channel setup. Each speaker was recorded in three sessions, and instructed to maintain a distance of 4-6 inches from the microphone. Each session contains 20 sentences to speak with a 5 minutes vocal rest between the sessions.

4.3. Audio file segmentation

There are different non verbal sounds like breathy, caugh, sneeze, and paper sounds present in the recordings. These unwanted sound are manually chopped from wave files and for segmenting the wave files, ZFF (Zero Frequency Filtering) technique [15] is used. The wave files are down sampled to 16 kHz, with 16 bit PCM mono. To ease the use of IIITH-HE-CM speech corpus in building speech recognition system, each wave file is given a unique ID that represents speaker name,

gender of the speaker and wave file. One second of silence is appended to each wave file before and after utterance.

5. Baseline speech recognition system

The IIITH-HE-CM phonetically balanced speech corpus contains 142 speakers with 71 male and 71 female speakers. There are 8804 total number of unique sentences.

5.1. Training and testing dataset

The IIITH-HE-CM corpus is divided into training and testing datasets. The trainig dataset contains 110 speakers (55 male and 55 female). The training dataset consists 6820 unique utterances. Each unique utterance should consist 5-15 words and each word in the utterance should be among the Hindi-English unique lexicon. These training datasets are unique and independent from the test data set. Similarly the test dataset contains 1984 unique utterances. These test utterances consist 5-15 words in each utterence. Each of the 62 sentences consists of code-mixed utterances. Each speaker utters 62 unique sentences. The acoustic feature extraction and acoustic modelling is performed on training dataset. The total duration of training speech corpus is approximately 8 hours. The remaining 32 speakers speech data is used for testing, with 16 male and 16 female. The total duration of testing speech corpus is approximately 3 hours. The text set consists of 8804 sentences. Out of these sentences, there are 12850 unique Hindi words and 4390 unique English words.

6. ASR experiments

The recognition experiments are performed on IIITH-HE-CM speech corpus, using the Kaldi ASR (Automatic Speech Recognition) toolkit. We trained a conventional GMM-HMM (Gaussian Mixture Model-Hidden Markov Model) system [9] with 18 k Gaussians using 39 dimensional MFCC (Mel Frequency Cepstral Coefficient) features including the deltas and delta-deltas to obtain the alignments for DNN (Deep Neural Network) [16] training. A standard feature extraction scheme [17] [18] is used by applying Hamming windowing with a frame length of 25 ms and frameshift of 10 ms. The language-dependent monolingual and language-independent bilingual DNNs [19] [20] with 3 hidden layers and 2048 sigmoid hidden units at each hidden layer are trained on the 40-dimensional log-mel filterbank features with the deltas and delta-deltas. The DNN training is done by mini-batch Stochastic Gradient Descent [20] with an initial learning rate of 0.002 and a minibatch size of 256. The time context size is 11 frames achieved by concatenating 4 frames. We have performed two experiments, in experiment 1 (henceforth exp1) : IIITH-HE-CM speech corpus contains training and testing data sets. The speech transcripts of test corpus were covered in the language model training.

Similarly the experiment 2 (henceforth exp2) : contains the training and testing datasets. The speech transcripts of test corpus were excluded and only unique word from test transcripts were covered in the language model training. The Figure 4 (tabular format) show the results of exp1 and exp2. We have observed that the text corpus includes the training and testing text corpus gives better results. But the exp2 has been conducted over several models and finally we have got better results using SGMM and DNNs.

	MONO	TRI1	TRI2	TRI3	TRI3.SI	TRI4_NNET	SGMM2_4
EXP1	14.13	8.14	8.54	8.23	9.08	8.02	7.08
EXP2	38.79	22.38	21.08	18.30	21.27	16.99	15.32

Figure 4: *Results from different combination of acoustic feature sets*

7. Summary

In this paper, we present the development of a phonetically balanced speech corpus of Hindi-English code-mixed speech. The large data has been extracted from selected sections of a popular news paper, Dainik Bhaskar. With the help of the large corpus, we have collected 8804 utterances of read code-mixed data amounting to 11 hours. We presented the development of automatic speech recognition system by constructing acoustic model and language model for the mixed languages using DNN architecture. In this case, Hindi-Egnlish code-mixed languages are used, and in general any other Indian language can be used in place of Hindi, with the help of mapping appropriate phones in that language to English phoneset.

8. Acknowledgement

The authors would like to thank the internship students namely Rohit kumar, Sai Teja, Sumedh, Bhavana, Sai, Nikhil and students of IIITH, for their help in speech recording and spending their time in correcting the phonetic transcriptions and verification of text corpus sentences.

9. References

[1] A. Pandey, B. M. L. Srivastava, R. Kumar, B. T. Nellore, K. S. Teja, S. V. Gangashetty, "Pho-netically Balanced Code-Mixed Speech Corpus for Hindi-English Automatic Speech Recognition," in *in Proc. of the Eleventh In-ternational Conference on Language Resources and Evaluation, LREC, Miyazaki, Japan, May*, 2018.

[2] E. Ylmaz, H. V. D. Heuvel, D. V. Leeuwen, "Inves-tigating Bilingual Deep Neural Networks for Automatic Recog-nition of Code-switching Frisian Speech," *in Procedia Computer Science*, vol. 81, pp. 159–166, 2016.

[3] P. F. Y. Li, "Code-switch language model with inversion constraints for mixed language speech recognition," in *in Proc. of COLING, Mumbai, December*, 2012, pp. 1671–1680.

[4] G. Paidi, S. R. Kadiri, B Yegnanarayana, "Analysis of Emotional Speech A Review," *Springer International Publishing*, vol. 11, pp. 205–238, March, 2016.

[5] https://en.wikipedia.org/wiki/Languages_of_India, 2008.

[6] K. Bhuvanagiri, K. Kopparapu, S., "An Approach to Mixed Language Automatic Speech Recognition," vol. 1, p. 3, January, 2018.

[7] Habib, M., A. Firoj, S. Rabia, Chowdhury S., Khan M., "Phonetically balanced Bangla speech corpus," in *in Proc. of the Human Language Technology for Development, Alexandria, Egypt*, May, 2011, pp. 87–93.

[8] Joyce Y. C, H. Cao, P. C. Ching, T. Lee, "Automatic recog-nition of Cantonese-English code-mixing speech," *in Proc. of the IJCLCLP*, vol. 14, no. 3, pp. 281–304, 2009.

[9] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *in Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] A. W. T. Schultz, "Language independent and language-adaptive acoustic modeling for speech recognition," vol. 35, pp. 31–51, 2011.

[11] K. Bhuvanagirir, S. K. Kopparapu, "Mixed language speech recognition without explicit identification of language," *in Proc. of the AJSP*, vol. 2, no. 2(5), pp. 92–97, 2012.

[12] U. Uebler, "Multilingual speech recognition in seven languages," vol. 35, pp. 53–69, 2011.

[13] K. Prahallad, N. Kumar E, V. Keri, R. Suyambu, A. W. Black, "The IIIT-H Indic Speech Databases," pp. 2546–2549, November, 2012.

[14] J. Kohler, "Multilingual phone models for vocabulary-independent speech recognition tasks," vol. 35, pp. 21–30, 2011.

[15] S. R. Murty K, B. Yegnanarayana, A. J. Xavier M, "Characterization of glottal activity from speech signals," *in Proc. of the IEEE signal processing letters*, vol. 16, no. 8, pp. 469–472, 2009.

[16] D. Povey, Z. Xiaohui, S. Khudanpur, "Parallel training of DNNs with Natural Gradient and Parameter Averag-ing," October, 2014.

[17] D. P. S. K. Xiaohui Z., J. Trmal, "Improving deep neural network acoustic models using generalized maxout networks," in *in Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014, pp. 215–219.

[18] C. Uraga, E. Gamboa, "Voxmex speech database: design of a phonetically balanced corpus," pp. 25–30, June, 2018.

[19] D. Povey, Xiaohui Z., S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," *CoRR*, vol. abs/1410.7455, 2014.

[20] S. P. Rath, D. Povey, K. Vesely, J. H. Cernocky, "Improved feature processing for deep neural networks," in *in Proc. of the INTERSPEECH, Lyon, France, August 25-29*, 2013, pp. 109–113.