Using English Acoustic Models for Hindi Automatic Speech Recognition

Anik DEY[1] *Ying Li*[1] *Pascale FUNG*[1]
(1) Human Language Technology Center
Department of Engineering and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
adey@ust.hk, eewing@ust.hk, pascale@ee.ust.hk

ABSTRACT

Bilingual speakers of Hindi and English often mix English and Hindi together in their everyday conversations. This motivates us to build a mix language Hindi-English recognizer. For this purpose, we need well-trained English and Hindi recognizers. For training our English recognizer we have at our disposal many hours of annotated English speech data. For Hindi, however, we have very limited resources. Therefore, in this paper we are proposing methods for rapid development of a Hindi speech recognizer using (i) trained English acoustic models to replace Hindi acoustic models; and (ii) adapting Hindi acoustic models from English acoustic models using Maximum Likelihood Linear Regression. We propose using data-driven methods for both substitution and adaptation. Our proposed recognizer has an accuracy of 96% for recognizing isolated Hindi words.

KEYWORDS : English, Hindi, Recognizer, Maximum Likelihood Linear Regression, Adaptation, Substituiton, Data-driven

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 123–134,
COLING 2012, Mumbai, December 2012.

1. INTRODUCTION

Hindi is one of the most widely spoken languages in the world. It is the major language of India and linguistically speaking, in its everyday spoken form, it is identical to Urdu, the major language spoken in Pakistan. Approximately 405 million people speak Hindi and Urdu worldwide (Sil, 1999). This makes research on Hindi automatic speech recognition systems very interesting due to the high utility of the languages. Hindi is written left to right in a script called Devangari, which we will discuss more in detail in section 1.1.

The last two decades have a seen a gradual progression in the development and fine tuning of automatic speech recognition systems. A few commercial automatic speech recognition (ASR) systems in Hindi have been in use for the last couple of years. The most prevalent ASR systems among them are IBM Via voice and Microsoft SAPI.

In (Kumar and Agarwal, 2011) we see a Hindi ASR being tested and evaluated on a small vocabulary for isolated word recognition. Other recognition systems we have seen so far have been tailor made for certain domains. The Centre for Development of Advanced Computing has developed a speaker independent Hindi ASR which makes use of the Julius recognition engine (Mathur et al., 2010). We have also seen significant work to deal with different accents of Hindi in (Malhotra and Khosla, 2008).

So far the most comprehensive Hindi ASR system we have come across is from the IBM Research Laboratory of India. They have developed a Hindi ASR where the acoustic models are trained with training data that is composed of 40 hours of audio data, and their language model has been trained with 3 million words. The IBM Research group has also worked on large-vocabulary continuous Hindi speech recognition in (Neti, Rajput and Verma, 2004).

However, significant research work has not been done to build a mixed language Hindi-English recognizer. To build such a recognizer we face a low-resource problem, because annotated Hindi speech data is very sparse. Hence, we propose to use well-trained English acoustic models to represent Hindi acoustic models for Hindi speech recognition. In this paper, we have discussed the MLRR adaptation technique, which we have used to map English to Hindi acoustic models using a data-driven approach, in Section 3. We have evaluated the performance of our Hindi ASR system in Section 4.

2. THE DEVANGARI SCRIPT

The Devangari script employed by Hindi contains both vowels and consonants just like in English. However, in contrast to English, Hindi is a highly phonetic language. This means that the pronunciation of any word can be very accurately predicted from the written form of the word.

In comparison with English, Hindi has half as many vowels and twice as many consonants. This usually leads to pronunciation problems. This problem is also encountered while modelling of Hindi phones using English phones is performed. This is because some phones in Hindi may not

be present in English at all. For this reason, we propose the data-driven approach. As a result of this approach we can approximate the English phone/s that is most closely matched to such a Hindi phone. The result of this approach is elaborated in the following sections.

In Hindi, consonants can be classified depending on which place within the mouth that they are pronounced.

To pronounce -

- *Velar* consonants: the back of the tongue touches the soft palate.
- *Palatal* consonants: the tongue touches the hard palate.
- *Retroflex* consonants: the tongue is curled slightly backward and touches the front portion of the hard palate. There are no retroflex consonants in English.
- *Dental* consonants: the tip of the tongue touches the back of the upper front teeth.
- *Labial* consonants: lips are used.

The consonants can also be classified according to their manner of articulation, as shown in Table 1 (Shapiro, 2008).

- *Unvoiced* consonants are when the vocal cords are not vibrated during their pronounciation.
- *Voiced* consonants are when the vocal cords are vibrated during pronounciation.
- *Unaspirated* consonants are when consonants are pronounced without a breath of air following the pronounciations. Example in English: “p” in “spit.
- *Aspirated* consonants are when a strong breath of air follows the consonant. Example in English: “p” in “pit”.
- *Nasal* consonants are pronounced when some air flows through the nose during pronounciation.

The vowels in Hindi are ordered in similar ways, as shown in Table 2 (Shapiro, 2008)

The manner of articulation of vowels can be classified into two particular categories:

- *Short* vowels are articulated for a comparatively shorter duration of time.
- *Long* vowels are articulated for a comparatively longer duration of time.

Monophthongs are vowels pronounced as a single sound, whereas *diphthongs* are vowels pronounced as a syllable comprising of two adjacent sounds glided together.

	STOPS				
	UNVOICED		VOICED		NASALS
	Unaspirated	Aspirated	Unaspirated	Aspirated	
Velar	क	ख	ग	घ	ङ
Palatal	च	छ	ज	झ	ञ
Retroflex	ट	ठ	ड (ड़)	ढ (ढ़)	ण
Dental	त	थ	द	ध	न
Labial	प	फ (फ़)	ब	भ	म

Table 1: Hindi Consonants

ARTICULATION	VOWELS		
	MONOPHTHONGS		DIPHTHONGS
	SHORT	LONG	
Guttural	अ	आ	
Palatal	इ	ई	
Labial	उ	ऊ	
Retroflex	ऋ	-	
Palato-Guttural		ए	ऐ
Labio-Guttural		ओ	औ

Table 2 : Hindi Vowels

4

The main difference between vowel pairs in Hindi (such as इ ई) is the vowel length. इ is pronounced like "i" in "bit" whereas ई is pronounced like "ee" in "feet", and उ is pronounced like "u" in "put" whereas ऊ is pronounced like "oo" in "boot".

The final consonants in the Devanagari script are organized into three categories: semivowels/approximants (य र ल व), sibilants (श ष स), and a glottal (ह).

Table 1 and Table 2 shows an in-depth categorization of Hindi alphabets. Since the distinction between manners of articulation is more prominent than from where within the mouth the alphabet is pronounced, we chose to classify the Hindi consonants into two classes, nasals and other consonants and Hindi vowels into two classes, monophthongs and diphthongs.

The distinction between voiced and unvoiced consonants is not as prominent as the difference between nasals and all other consonants, hence we chose to keep all voiced and unvoiced consonants within the same class.

For obtaining the phonetic transcription of English, we are using Arpabet where every English phoneme is represented by one or two capital letters.

In Arpabet, English vowels are classified into three classes : monophthongs, diphthongs and R-colored vowels (The CMU Pronounciation Dictionary, 2007).

Consonants are classified into 6 classes : stops, affricates, fricatives, nasals, liquids and semivowels (The CMU Pronounciation Dictionary, 2007).

By comparing English phonemes with Hindi alphabets, we notice that both languages have nasal consonants and monophthongs and diphthongs vowels. Hence we are also classifying the English phonemes into 4 classes : monophthongs (M), diphthongs (D), nasals (N) and all other consonants (C).

In English Arpabet we have an extra vowel which is neither a monophthong nor a diphthong. This extra vowel, ER, we label into a separate class V.

3. DATA DRIVEN PHONE MAPPING

One of the first steps to map Hindi phones to English phones is to obtain English phoneme transcriptions of Hindi characters. As an intermediate step all the Hindi characters can be transliterated to English using Google Transliteration which uses the International Alphabet of Sanskrit Transliteration (IAST) scheme.

One can use the English recognizer in free form to search through Hindi speech to obtain English phonemes to represent each Hindi characters.

This method yields very poor results when the Hindi acoustic data is limited, and is comparable to randomly searching the Hindi data with an English recognizer with no constraints.

This free form phoneme network of the recognizer allows every phoneme to be followed by every other phoneme including itself just as shown in figure 1.

```
$phone = all consonants and vowels
( sil <$phone> sil )
```

Figure 1: Free Form phonetic network

To improve English-phoneme labeling of Hindi speech, we propose to use the linguistic knowledge of Hindi and English as discussed in section 2 to classify all Hindi syllables and English phonemes into four different classes based on their articulation properties.

The four classes we selected are monophthongs (class M), dipthongs (class D), nasals (class N) and consonants (class C).

Each Hindi syllable and English phoneme is labeled to be one of these classes. The classification is shown in table on page.

By using linguistic knowledge of Hindi, we then modify our recognizer into a constrained form network where one phone from one class of the target language, Hindi, is mapped to one phone from the same class of the source language, English.

```
$phone = class M or class D or class N or class C
( sil <$phone> sil )
```

Figure 2: Constrained Form phonetic network

For adaptation, we have made use of the Maximum Likelihood Linear Regression (MLRR) technique which is a popular Expectation-Maximisation technique used for speech adaptations.

MLRR adaptation is performed to minimize the mismatch between the English acoustic models and the Hindi acoustic data which is used as the adaptation data. MLLR will compute a set of transformations which will alter the means and variances of Gaussian mixture HMM English acoustic models so that each state of the HMM model is more likely to generate the Hindi adaptation data.

The transformation matrix used to give a new estimate of the adapted mean is given by

$$\hat{\mu} = W \xi,$$

where W is the $n \times (n + 1)$ transformation matrix (where n is the dimensionality of the data) and ξ is the extended mean vector,

$$\xi=[w\mu 1\mu 2 \ldots \mu n]^T$$

where w represents a bias offset whose value is fixed (within HTK, the Hidden Markov Model Toolkit) at 1.

Hence W can be decomposed into

$$W=[bA]$$

where A represents a $n \times n$ transformation matrix and b represents a bias vector.

After adaptation we can use the Hindi-English phonème mapping (shown in table on page) to construct a pronunciation dictionary for Hindi syllables. Adding linguistic knowledge to enhance the recognizer improves the Hindi ASR.

4. EXPERIMENT

We collected 1 hour of Hindi acoustic data from 9 native Hindi speakers. We asked each speaker a set of questions regarding their university life, likes, dislikes, hobbies and about their career ambitions.

The complete set of Hindi data collected was divided into development and test sets. The test set of data consists of 50 Hindi phrases from one of the 9 speakers. The development set consists of 45 minutes of Hindi acoustic data from 8 different speakers.

After collecting the data, we hired a native speaker of Hindi to transcribe the data for us. Most of the speakers used both English and Hindi while answering the questions on the questionairre. Hence, the transcribed data was a mix of English and Hindi written using the Devangari script.

We used the Carnegie Mellon University (CMU) Pronouncing Dictionary to obtain the phone level transcriptions of all English words in the above transcription. The CMU dictionary uses a phoneme set that consists of 39 phonemes. Each phoneme is represented by one or two capital ASCII letters (ARPAbet).

For all the words written using the Devangari script, we made use of Google's Phonetic Typing service to obtain phone level transcription for all Hindi words. The list of phone level transcription for each Hindi alphabet is shown in table 3 on page 10.

After obtaining the transcriptions, we labelled each phoneme in the transcriptions as one of the 4 classes, discussed in section 3.

For training the English acoustic models 65 hours of native English speech was used, which was kindly shared to us by the guys at the Wall Street Journal.

Using adaptation by reconstruction, we can now obtain the mapping of Hindi phonemes to English. This is shown in table 4 on page 11.

By using English acoustic models, the recognition accuracy to recognize Hindi phrases in the test set, discussed above, is 96%.

CONCLUSION AND DISCUSSION

In this paper, we have proposed steps to rapidly develop a Hindi speech recognizer: (1) by substituting Hindi acoustic models with trained English acoustic models; and (2) by adapting these models using MLRR. We have shown how data-driven methods and linguistic knowledge can be used to map English phonemes to Hindi syllables. With the pronunciation dictionary we constructed we can easily find the phone level transcriptions of any new Hindi word given in written form.

Given a small set of training data, our proposed Hindi constrained-form recognizer has shown promising results. However, there is a lot of room for improvements. Provided that we can collect more Hindi acoustic data, to increase the size of training data drastically, we will be able to better model the Hindi syllables with more than one phoneme transcription.

We also plant to further study multilingual speech recognition since Hindi and English are spoken together by virtually all bilingual speakers of English and Hindi. Also we think better modeling is needed for Hindi phonetic units that do not exist in English.

We are collecting more Hindi acoustic data every month and fine-tuning our Hindi acoustic models. We hope that this can enhance our Hindi acoustic models and improve recognition accuracy.

We are also exploring asymmetric acoustic modelling using selective decision tree merging between a bilingual model and an accented embedded speech model for Hindi and English multilingual speech recognition since this method has shown to improve recognition results for mixed language speech consisting of English and Chinese (Ying et al., 2011). For English and Chinese this method works because English phrases are generally pronounced by a Chinese speaker with varying degrees of accent. The same is true for English and Hindi.

Acknowledgments

We will like to thank Abhilash Veeragouni of IIT Bombay for helping us collect and transcribe Hindi acoustic data which we used as training and test data for our experiments.

References

K. Kumar and R. K. Agarwal (2011). Hindi Speech Recognition System Using HTK. *International Journal of Computing and Business Research*, Vol. 2, No. 2, 2011, ISSN (On- line): 2229-6166.

R. Mathur, Babita and A. Kansal (2010). Domain Specific Speaker Independent Continuous Speech Recognition Using Julius. ASCNT 2010.

K. Malhotra and A. Khosla (2008). Automatic Identification of Gender & Accent in Spoken Hindi Utterances with Re- gional Indian Accents. *IEEE Spoken Language Tech- nology Workshop*, Goa, 15-19 December 2008, pp. 309- 312.

C. Neti, N. Rajput and A. Verma (2004). A Large Vocabulary Continuous Speech Recognition System for Hindi. *IBM Research and Development Journal*, September 2004.

L. Ying, P. Fung, P. Xu, Y. Liu (2011). Asymmetric Acoustic Modeling of Mixed Language Speech. ICASSP 2011.

Michael C. Shapiro (2008). A Primer of Modern Standard Hindi. *Motilal Banarsidass Publishers Private Limited*, 2008 Reprint

Carnegie Mellon University (2007). The CMU Pronounciation Dictionary. < http://www.speech.cs.cmu.edu/cgi-bin/cmudict > (visited 23, October, 2012)

H. Sil (1999). Ethnologue: Languages of the World. < http://www.ethnologue.com/web.asp > (visited 23, October, 2012)

APPENDIX

अ	A
आ	Ā
इ	I
ई	Ī
उ	U
ऊ	Ū
ऋ	R̥
ए	Ē
ऐ	Ai
ओ	Ō
औ	Au
क	Ka

ख	Kha
ग	Ga
घ	Gha
च	Ca
छ	Cha
ज	Ja
झ	Jha
ञ	Ña
ट	Ṭa
ठ	Ṭha
ड	Ḍa
ढ	Ḍha

ण	Ṇa
त	Ta
थ	Tha
द	Da
ध	Dha
न	Na
प	Pa
फ	Pha
ब	Ba
भ	Bha
म	Ma
य	Ya

र	Ra
ल	La
व	Va
श	Śa
ष	Ṣa
स	Sa
ह	Ha

Table 3: IAST Transliteration of Hindi alphabets

kh	k
ai	ey
ch	t
au	ow ey
gh	l
th	l
ph	f
bh	l r
jh	d
a	ah
c	t

b	b
e	ih
d	v
g	l
i	iy
h	l
k	k
j	d
m	m
l	l
o	ah

n	n
p	p
s	s
r	d
u	ah
t	t
v	l
y	hh

Table 4: Hindi to English Phoneme Mapping