Using English Phoneme Models for Chinese Speech Recognition

MA Chi Yuen and Pascale FUNG
The Human Language Technology Center
Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology (HKUST), Hong Kong
Tel. (852) 2358 8537, FAX: (852) 2358 1485, Email:{eecyma,pascale}@ee.ust.hk

ABSTRACT

To build a speech recognizer, database design, collection and transcription is the most time consuming and tedious job. This paper proposes some fast and easy methods to use English phoneme models for Mandarin and Cantonese speech recognition with little to no training data in Mandarin and Cantonese. While a recognizer built with such transformed models might not perform as ideally as one that is trained on a large database, we demonstrate that its performance is good in constrained applications such as speech-based Web browsing and searching. The web link recognition rate is 83% for Mandarin and 92.5% for Cantonese.

1. INTRODUCTION

As the World Wide Web has became a major platform of information repository and dissemination, it can be used for many applications that previously require stand-alone user interface design. We have successfully implemented a speech-based web browser for the purpose of campus information access for staff and students[1]. This system can be easily extended to any other information access domain, such as tourist information at the airport, by adding web pages containing all the information and provide query selections for the user in the form of page links. In Hong Kong, it is essential for public information to be available in both its official languages---Chinese and English. For speech input, it is also necessary to recognize three languages---Cantonese, English and Mandarin. Multilingual systems are also useful in many countries such as in the U.S., where Spanish is fast becoming the nation's second language, and in Europe, where there is a high rate of information exchange between the nations.

The ultimate solution for supporting multilinguality is a universal phoneme set including all phonemes from all languages. The most common method for recognizing the speech of a new speaker is to use speaker-independent models trained from a large group of speakers. Similarly, large amount of multilingual data is needed to train such a universal phoneme set. In many cases, however, we are not able to obtain such a large database of different languages to begin with, especially one which encompasses European as well as Asian languages. If we use the data form different database, the different recording conditions can lead to considerable inconsistency in our training set. In a constrained domain such as Web-based information access, we suggest that there is a faster and more economical way to achieve multilinguality---phoneme set mapping.

We propose to map English phoneme models to Mandarin and Cantonese phonetic units, and use these models to recognize Mandarin and Cantonese speech. English models can be trained from the large amount of English speech databases that are readily available. Even though some Mandarin database has also become available in recent years, they are inadequate in comparison. In addition, there is not yet any large Cantonese database available. Mapping of English phoneme models to Mandarin and Cantonese allows us to add multilingual capability to our speech-based Web browser in a very short amount of time.

2. EXPERIMENT

We need to map English phonemes to Chinese phonemes with the most similar acoustic pattern or pronunciation. The distance between two HMMs could be used to measure the relationship between two phonemes[2]. In this paper, we propose and compare different ways of mapping: (1) the Free-Form Data Driven method; (2) the Fixed-Form Data Driven method; (3) the Knowledge-based method; and (4) a Hybrid method.

2.1 Experiment Setup

We use context independent English phoneme HMM with one Gaussian mixture as the baseline model for cross-lingual phoneme mapping. Features used in the HMM are: 12 MFCC, 12 delta MFCC, 12 delta delta MFCC, energy, delta energy, and delta delta energy. Each phoneme has 3 states. The training database is Wall Street Journal.

We evaluate the performance of the mapped models on our speech based web browser system. On the average, there are 40 links in a page. Some pages have both English and Chinese links.

2.2. Data Driven – Free Form

One convenient way of mapping phoneme sets is to use a recognizer based on English models to decode the acoustic data of Mandarin or Cantonese speech. We collect some utterances in Mandarin and use this method to obtain English phoneme transcriptions of Chinese characters in Mandarin. We feed a small set of isolated utterances of monosyllabic Mandarin data into a well-trained English phoneme recognizer. The phoneme network of the recognizer allows every phoneme to be followed by every other phoneme, including itself (fig 1).

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

Fig. 1 Free Form phonetic network

Using the English decoder output, we construct a pronunciation dictionary for Mandarin syllables using English phoneme set. However, such a free form phoneme network allows for too many insertion, deletion and replacement in the phoneme transcription of the same syllable. It is very difficult to reconcile the different phonetic transcriptions of the same syllable except with more accurate statistical modeling. This cannot be achieved without more acoustic data in Mandarin and Cantonese. We will study the effectiveness of this method in the future when more data becomes available.

2.3. Data Driven – Fixed Form

For quick mapping of English phoneme set to Mandarin and Cantonese phoneme sets, we decide to make use of a special property of Chinese characters. In Chinese, each word is made up of one to several Chinese characters, and each Chinese character consists of only one syllable[3,4]. Conventionally, each syllable is considered to be consisting of only two phonetic units---initial and final. So, instead of using a free form phoneme network, we use a fixed form phoneme network for the English phoneme recognizer (fig 2), to recognize isolated Chinese syllables. By the output confusion matrix, we can find out which Chinese phonetic unit (initial or final) is most similar to which set of English phonemes.

```
( sil [ $consonant ] $vowel [$nasal] sil )
```

Fig. 2 Fixed Form phonetic network

Man	Eng	Man	Eng	Man	Eng
a	aa	iang	aa ng	t	hh
ai	ae	iao	aw	u	uw
an	ae ng	ie	ey	ua	aa
ang	aa ng	in	iy ng	uai	ay
ao	aw	ing	iy ng	uan	ay ng
b	b	iong	uw ng	uang	aa ng
c	th	iou	ow	uei	ey
ch	ch	j	y	ue	er n
d	b	k	k	uen	er n
e	aa	l	y	ueng	aa ng
ei	ey	m	m	uo	ao
en	ae n	n	y	ü	iy
eng	aa ng	o	ao	üan	ae ng
er	aa	ong	ow ng	üe	ey
f	f	ou	ow	ün	ey ng
g	g	p	p	w	w
g	hh	q	ch	x	s
i	iy	r	y	y	y
ia	aa	s	s	z	th
ian	ae ng	sh	sh	zh	jh

Fig 3. Mapping of Mandarin to English phonemes found by Fixed Form Data Driven approach

In the confusion matrix, most of the Chinese phonemes have one-to-one mapping with English phonemes. But some other Chinese phonemes are mapped to several English phonemes in different samples. In such case the relationship between the English phoneme and Chinese phoneme is not so consistent, and the mapping is not certain.

We then transcribe Chinese characters into Mandarin initial-finals by using the above mapping. English phoneme HMMs are used in lieu of Mandarin initial-final HMMs. For simplicity, we use the highest score's English-to-Mandarin phoneme mapping pairs, no matter the mapping is close or not. The result accuracy of Mandarin web link recognition is **83%**[1].

2.4. Knowledge-based

In the above approach, English-to-Mandarin phoneme mapping is obtained by Viterbi alignment between English phonemes and Mandarin acoustic data. For Cantonese, such a mapping is available in knowledge

[1] Since the speech recognizer works for mix languages, it searches through both Mandarin and Cantonese phonetic lattice for recognizing Chinese web links. Hence the performances of both Mandarin and Cantonese link recognition are not independent of each other. We plan to perform more independent evaluations.

form. From a Cantonese pronunciation textbook[5] designed for English speakers, we extract the rules shown in (fig.4). When the English recognizer is applied to Cantonese input speech using this knowledge-based mapping, we obtain a link recognition accuracy of **90.5%**.

2.5. Hybrid Approach

The results from the data-driven approach and the knowledge-driven approach are comparable. For Cantonese speech, linguistic knowledge eliminates the need to collect speech data. However, most recognition errors for Cantonese speech are caused by the fact that the mapping between Cantonese and English phonemes are inadequate. Some finals in Cantonese simply do not exist in the English phoneme set. To handle such cases, we apply Free-form Data-Driven method to the Cantonese syllable and obtain another transcription. As an example, /au/ in Cantonese is found to be best matched to /aw/ /ao/ in concatenation instead of to /aw/ only in English. By apply this hybrid method to amend some of the Cantonese initial-final mappings to English phonemes, the web link recognition rate is improved to **92.5%**.

Can	Eng	Can	Eng	Can	Eng
aa	aa	eoi	uw	oe	er
aai	ay	eon	uh n	oei	uh
aak	aa k	eot	uh t	oek	er ng
aam	aa m	ep	ea p	oeng	oy
aan	aa n	eu	uw	oi	ay
aang	aa ng	f	f	ok	ao k
aap	aa p	g	g	on	ao n
aat	aa t	gw	g w	ong	ao ng
aau	aw	h	hh	ot	ao t
ai	ay	i	iy	ou	ow
ak	ah k	ik	ih k	p	p
am	ah m	im	iy m	s	s
an	an	in	iy n	t	t
ang	ah ng	ing	ih ng	u	uw
ap	ah p	ip	ih p	ue	uw
at	ah t	it	ih t	uen	uw n
au	aw	iu	iy	uet	uw t
b	b	j	y	ui	uh
c	ch	k	k	uk	uh k
d	d	kw	kw	un	uw n
e	ea	l	l	ue	uw
ei	ey	m	m	ung	uh ng
ek	ea k	n	n	ut	uw t
em	ea m	ng	ng	z	jh
eng	ea ng	o	ao		

Fig 4 Mapping of Cantonese to English phonemes obtained from linguistic knowledge

3. CONCLUSION AND DISCUSSION

We have proposed some fast and easy way for Chinese speech recognition using English phoneme models. Fixed Form Data-Driven method is used when a small amount of Mandarin data is available. The web link recognition rate is **83%** for Mandarin in this case. A hybrid approach combining both linguistic knowledge and the Fixed Form Data-Driven method is found to be effective for Cantonese web link recognition **(92.5%)**. We suggest that these methods can be applied to any constrained-domain recognition task, such as telephone directory assistance where only a limited number of names need to be recognized.

Although the results in constrained-domain application are found to be good, we plan to further study issues related to full-fledged multilingual speech recognition. For example, better models can be designed for the Free-Form Data Driven method; better modeling is also needed for Chinese phonetic units that do not exist in English phoneme set; similarly for tonal information. We are currently experimenting with online adaptation of English phoneme models to Mandarin and Cantonese by continuously using user input to the speech-based Web browser. We hope that it can improve the recognition rate even further.

4. REFERENCES

[1] Pascale Fung, Cheung Chi Shun, Lam Kwok Leung, Liu Wai Kat, Lo Yuen Yee: "Salsa, A Hong Kong English Speech-based Web Browser", International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia, 1998.

[2] B.Wheatley, K. Kondo, W. Anderson, and Y. Muhtusamy. An evaluation of cross-lanugage adaptation for rapid hmm developmetn in a new language. In ICASSP, volume1, pages 237-240, 1994.

[3] Pascale Fung and Lam Kwok Leung: "Are Initial/Final Units Acoustically Accurate?", Symposium on Image, Speech, Signal Processing and Robotics (ISSPR 98), Hong Kong, 1998.

[4] Zhang Jialu: "Phonetic and linguistic features of spoken Chinese", 1994 International Symposium on Speech, Image Processing and Neural Networks, pages 117-121, 1994

[5] Huang, Hsi-ling: "Yueh yin yun hui", pages 98-106, Hong Kong, 1991.

[6] Steve Young: "The HTK Book", 1997