CHINESE-ENGLISH BILINGUAL PHONE MODELING FOR CROSS-LANGUAGE SPEECH RECOGNITION

Shengmin Yu Shuwu Zhang Bo Xu

Hi-Tech Innovation Center, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100080
{smyu, swzhang, xubo}@hitic.ia.ac.cn

ABSTRACT

In this paper, three different approaches of Chinese-English bilingual phone modeling are investigated and compared. The first approach is to simply combine Chinese and English phone inventories together without phone shared across the languages. The second one is to map language-dependent phones to the inventory of the International Phonetic Association (IPA) based on phonetic knowledge to construct the bilingual phone inventory. The third one is to merge the language-dependent phone models by hierarchical phone clustering algorithm to get a compact bilingual inventory. In the third approach, two distance measures are used to perform the bottom-up clustering. One is Bhattacharyya distance. The other is acoustic likelihood distance. Experimental results show that phone clustering approach outperforms IPA-based phone mapping approach, and it can also achieve comparable performance to the simple combination of language-dependent phone inventories with less model parameters, especially when using acoustic likelihood distance measurement.

1. INTRODUCTION

In recent years, the significant progress achieved in automatic speech recognition (ASR) has led to a variety of successful demos and potential commercial applications. Most of these successes are currently language-specific (monolingual) and are limited to be remarkable only for native speakers. To increase the usability of a prototype system, the problems of multilingual and non-native speech should be addressed efficiently. If a language independent ASR system could be shared for the user with different nationalities, it would be very challenging to stride over language boundary for international communication. The development of multilingual or cross-lingual speech recognizer (MSR) is thus an important research topic for opening a large spectrum of potential ASR applications. Striding over the language boundary towards multilingual interoperability is the prospective goal for the progress of ASR technology.

Some studies on MSR have been reported during recent years [1][2][3][4][5]. Most of them are focused on language independent acoustic modeling and fast adaptation for new languages. But the results are still not very significant yet compared to language dependent monolingual ASR systems.

Actually, present monolingual ASR systems are potentially multilingual, as the main infrastructures and algorithms for developing recognizers in a large variety of languages are in the same family. Simply speaking, what we need to further study for multilingual speech is to combine the universal acoustic phone inventory and models, and to improve its portability under a unified ASR architecture. However, it should be clear, when porting monolingual speech recognizer to bilingual or multilingual recognizer, certain system parameters or components will have to be changed, e.g., those language-dependent knowledge sources such as phone inventory, the recognition lexicon and phonological rules.

Focused on Chinese and English languages, the work of this paper is try to find a suitable and robust phone inventory for building a real Chinese-English bilingual speech recognition system. Some issues we encountered in porting our Chinese speech recognition technology to bilingual construction in the framework of LVCSR are addressed and discussed too. Experimental results show that phone clustering approach is a promising way for the determination of global phone inventory in multilingual acoustic modeling.

The remainder of the paper is organized as follows. In section 2, three approaches and corresponding issues in training are described in detail. Some experimental results on Chinese-English bilingual phone modeling are compared in section 4. Conclusion is given in section 5.

2. THREE DIFFERENT BILINGUAL PHONE MODELING APPROACHES

In multilingual speech recognition, it is very important to determine a global phone inventory for different languages involved in the system. Some approaches for building the phone inventory can mainly be classified as: 1) to simply combine the language-dependent phones into one set; 2) to share the phones with acoustic similarity by mapping language-dependent phone into IPA set [8]; and 3) to merge monolingual's phone models by data-driven clustering. In this section, we will give a comparison of above three approaches based on the Chinese-English bilingual phone modeling.

2.1. Combination of Language-dependent Phone Inventories

A natural way of building multilingual phone inventory is to combine language-dependent phone inventories into one set. But it could cause a sharp increase of parameter size. Table 1 shows the comparative Chinese-English monolingual phone inventories. It includes 35 Chinese phones [9] and 40 English phones

0-7803-8484-9/04/$20.00 ©2004 IEEE

respectively. Following this approach, the phones within two languages are pooled together to a unified bilingual phone inventory. Based on the phone inventory, we can train a bilingual acoustic model. No any acoustic parameters are shared across the two languages in the model.

Phone Classes	Chinese*	English
Voiced plosives	b d g	b d dx g
Unvoiced plosives	p t k	p t k
Fricatives	f h s sh x	f dh hh th v
Affricatives	c ch j q z zh	ch
Sibilants		s sh z
Nasals	m n ng	m n ng
Lateral	l	
Glides	r	er l r
Front vowels	Ci i v	ae eh ey ih ix iy y
Central vowels	e eI eN Ie	ax axr ah uh
Back rounded vowels	o oU	ao
Back unrounded vowels	a aI Chi u	aa ow uw w
Diphthongs		aw ay

**Note: In our bilingual applications, phones in Chinese are fully labeled by a tag ("_Ch") to make a distinction with English.*

Table 1. Bilingual Global Phone Inventory

2.2. Direct IPA Mapping

The second set of bilingual phone inventory is defined based on phonetic knowledge. Some language-dependent phones, which are represented by the same IPA symbol, share one common phone category. This approach is performed based on phonetic knowledge rather than some of statistically based similarity or distance measurement. The rule of phone mapping is similar to [6], which can be expressed by

$$Ph_{l,i}^{LDP} \rightarrow Ph_{j}^{IPA} . \qquad (1)$$

where $Ph_{l,i}^{LDP}$ denotes the *i*th phone of language *l*, Ph_{j}^{IPA} denotes *j*th symbol of IPA.

Compared with the combination of two monolingual's phone inventories, some phones are shared across the languages and this will lead to reduction of acoustic model parameters. The advantage of this approach is that the multilingual phone symbols have clear representation in the context. On the other hand, direct IPA mapping does not consider the spectral properties of phone models and is not consistent with the statistical similarities of the final search space. Hence, acoustic model parameters probably can't describe the distribution of real training data precisely. When more languages are considered in the multilingual system, the problem will become more serious.

Table 2 is a list of IPA-based Chinese-English Bilingual phone inventory including total of 57 phones.

2.3. Automatic Phone Model Clustering

In hierarchical clustering algorithm, the definition of distance measure among phone models is an issue that has been widely addressed. The divergence of two gaussian distributions as a function of their mean and variance values was defined in [11]. However, this approach only applies to one state Markov models with a single gaussian distribution. The Bhattacharyya distance is a theoretical similar measure between two Gaussian distributions as it is equivalent to an upper bound on the optimal Bayesian classification error probability [12]. The Bhattacharyya distance is calculated using phone model's parameters directly as the following formula:

$$D_{bha} = \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\left| \frac{\Sigma_1 + \Sigma_2}{2} \right|}{\sqrt{|\Sigma_1||\Sigma_2|}} , \qquad (2)$$

where μ and $\sum$ denote mean vector and variance of each phone model, T means matrix transpose. In this paper, it is also performed as a comparison with acoustic likelihood measurement.

Phone Classes	IPA-based bilingual phones
Voiced plosives	b d dx g
Unvoiced plosives	p t k
Fricatives	Sh f dh hh th v
Affricatives	C zh Ch ch
Sibilants	s sh z
Nasals	m n ng
Glides	er l r
Front vowels	Ci i V ae eh ey ih ix iy y
Central vowels	E eI eN Ie ax axr ah uh
Back rounded vowels	o oU ao
Back unrounded vowels	a aI Chi u aa ow uw w
Diphthongs	aw ay

Table 2. IPA-based Bilingual Phone Inventory

A novel acoustic likelihood measurement, which be similar to [6] and [10], is also used to measure distance between two phone models in this work. Based on the mutual information, likelihood between two phone models λ_i and λ_j can be defined as

$$L(\lambda_i, \lambda_j) = f(\vec{X}_i \mid \lambda_j)^{\alpha} / \sum_{k=1}^{n} f(\vec{X}_i \mid \lambda_k)^{\alpha} , \qquad (3)$$

where $\vec{X}_i$ denotes a sequence of observations labeled with phone *i*. $f(\vec{X}_i \mid \lambda_j)$ is the probability density function (pdf) of the observations, and *n* is the number of phone models. The coefficient α is introduced to compensate the hypothesis of independence between phone models and is fixed to 0.5 in an empirical way. Since this measure is asymmetrical, we calculate the average distance as follows

$$L = \frac{1}{2}(L(\lambda_i, \lambda_j) + L(\lambda_j, \lambda_i)) . \qquad (4)$$

Because it is difficult to estimate the new phone model's parameters of the merged class, the distance is always calculated with the language-dependent models. To avoid getting too large classes that including many phone models, the furthest neighbor criterion is also used in each iteration step, as it is

$$L_{ij} = \max_{k\in C_i, l\in C_j} L(k,l)\,. \qquad (5)$$

Detailed algorithm can be described as follows:

1) *Initialization:* Assign each language-dependent phone model to a class.
2) *Loop:* Calculate the distance matrix and merge the two classes with the minimum distance.
3) *Termination:* Check if the desired number of phone classes is reached. If yes, then terminate loop of the procedure, else go back to step 2).

After clustering we can use obtained phone classes to map the language-dependent phone models to the bilingual inventory. The bilingual dictionary and label files are also processed based on the clustering information.

The advantage of the approach is that it uses the training data of phone models for the distance calculation. Therefore, it is consistent with the final recognition phase which also uses statistical measurement based on HMM technology. In order to achieve a robust measurement, quantity of each phone model's training acoustic data is set to 2000 frames, which is larger than [6]. Table 3 shows the result of bilingual phone clustering.

Phone Classes	**CLU_B**	**CLU_L**
Voiced plosives	b d g dx	b d g
Unvoiced plosives	p t k	p t k
Fricatives	F h f dh hh th v S Sh x	F f dh hh th v Sh x
Affricatives	c Ch j q Z zh ch	c Ch j q Z zh ch
Sibilants	s sh z	s sh z
Nasals	m n ng	m n ng
Glides	er l r	R er l r
Front vowels	Ci i ae ey ih iy y	Ci i ae eh ey ih ix iy y
Central vowels	e eI Ie ax ah	e eI eN Ie ax ah
Back rounded vowels	o oU ao	oU ao
Back unrounded vowels	a aI Chi u aa uw w	aI Chi u aa ow uw w
Diphthongs	aw ay	aw ay

**Note: CLU_B means clustering by Bhattacharyya distance, and CLU_L by acoustic likelihood distance. Num. of terminate classes is 57 for both distance measurements.*

Table 3. Bilingual Phone Inventory by phone clustering

3. EXPERIMENTS AND DISCUSSION

3.1. Corpora and Experimental Setup

The training corpus consists of a Chinese speech database of DB863 and an English speech database of Wall Street Journal (WSJ0). DB863 is a continuous speech recognition corpus including 54 hours male speeches and 57 hours female speeches with total of 166 speakers. Each speaker uttered around 520~650 utterances. WSJ0 is a subset of general English speech database released by Linguistic Data Consortium (LDC) [7]. In this work, si-tr-s set of WSJ0 is pooled together with Chinese DB863 as new training set. The test set consists of 240 Chinese sentences, which be recorded by 6 male speaker (40 sentences by each one), and 330 sentences of standard WSJ0 test set.

3.2. Experiments and Discussion

In our experiments of building Chinese-English bilingual speech recognition system, the decision-tree based clustering algorithm is employed for training context-dependent triphone acoustic models. The same left-to-right 3-state topology with no skip transition is employed for both Chinese and English languages. In the recognition phase, one pass search algorithm integrated with tri-gram language model (LM) look-ahead technology is applied to decode input speech signals. For the consistency, we chose 35 phones as basic Chinese phone inventory for Chinese acoustic modeling.

For the sake of comparison, we firstly conducted the experiments on language-dependent monolingual speech recognition systems. Table 4 shows the result of baseline monolingual recognition accuracy. The average accuracy of two languages is 88.25%.

Table 4 also shows that Chinese side has relatively lower recognition accuracy. It is mainly due to the compromised selection of the phone inventory [13]. With consonant/vowel decomposition, our Chinese monolingual system can achieve much higher performance [14].

Monolingual	Accuracy (%)
Chinese	88.9
English	89.6

Table 4. Two monolingual's experimental results

Based on the different definitions of Chinese-English bilingual phone inventory, we further conducted a series of experiments on bilingual acoustic modeling. We also trained two sets of language models for the comparison. One set is two of monolingual language models. Another one is a combined bilingual language model. The bilingual dictionary has 47,307 words, which are combined by 39,969 Chinese words and 7338 English words.

Figure 1 and Figure 2 draw the accuracy scales with regard to the definitions of bilingual phone inventories by using different type of language models respectively.

We can see that all bilingual models caused a clear degradation of word accuracy compared to monolingual models (Mono). Under bilingual environment (as shown in Figure 1), the combination of language-dependent phone inventories (Comb) achieves an acceptable performance with an average accuracy of 81.9%. The IPA mapping approach (IPA) caused a moderate degradation with 78.6% compared to previous one. Phone clustering approach (CLU_B, CLU_L) achieved a significant improvement compared to IPA mapping and a comparable performance to direct combination of monolingual phone inventories. Especially, phone clustering by acoustic likelihood distance (CLU_L) can reach even a slight higher accuracy of 82.2% compared to the Comb approach.

On the other hand, considering the number of parameters in phone modeling (number of pdfs), monolingual models have the number of parameters about 53k and 36k corresponding to Chinese and English respectively. In the first set of bilingual acoustic model (Comb), the number of parameters is 89k, which is the sum of above two. In IPA mapping and phone clustering, since we set the stop threshold of the number of phone classes in clustering to the same as IPA set (57 classes), the number of parameters are 63k for both sets. It has a parameter reduction of 29.21% compared to Comb set. We can, thus, say that phone clustering approach can achieve comparable performance to the simple combination of language-dependent phone inventories with less model parameters, especially when using acoustic likelihood distance measurement.

It is also shown that the system using separated monolingual language models (Figure 2) outperforms the systems with combined bilingual language model shown in Figure 1.

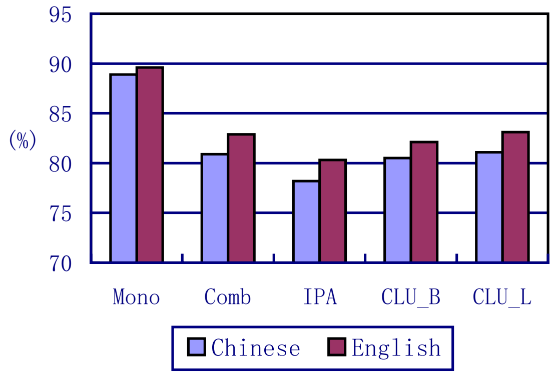

Figure 1. Bilingual recognition results with bilingual LMs

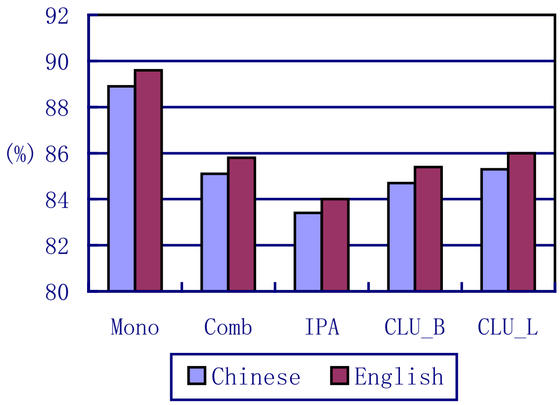

Figure 2. Bilingual recognition results with monolingual LMs

4. CONCLUSION

We have investigated the phone modeling approaches in Chinese-English bilingual speech recognition. Experimental results have testified that phone clustering by acoustic likelihood distance measurement can achieve the better recognition performance with less model parameters in LVCSR. By now, phone clustering algorithms adopted in the experiments are based on unsupervised learning. Based on the IPA definition, we will continue to investigate supervised phone clustering algorithms to pursue advanced improvement of multilingual phone modeling.

5. ACKNOWLEDGEMENT

This work is supported by High Technology Research & Development (863) Program of China (2001AA114015, 2001AA114070) and Olympic Program of Beijing Committee of Science and Technology (H030130050430).

6. REFERENCES

[1] M. Adda-Decker, "Towards Multilingual Interoperability in Automatic Speech Recognition," Speech Communication, Vol. 35(1-2), pp. 5-20, 2001.
[2] T. Schultz and A. Waibel, "Language-independent and language-adaptative acoustic modeling for speech recognition," Speech Communication, Vol. 35(1-2), pp. 31-51, 2001.
[3] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," In Proc. Eurospeech, pp. 371-374, 1997.
[4] U. Uebler, "Multilingual speech recognition in seven languages," Speech Communication, Vol. 35(1-2), pp. 53-69, 2001.
[5] F. Weng et al., “A study of multilingual speech recognition,” In Proc. Eurospeech'97, 1997.
[6] J. Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," Speech Communication, Vol. 35(1-2), pp. 21-30, 2001.
[7] Linguistic Data Consortium (LDC), University of Pennsylvania, http://www.ldc.upenn.edu.
[8] IPA, “The International Phonetic Association (revised to 1993) - IPA Chat.” J. Int. Phonetic Assoc., 23, 1993.
[9] Bin Ma and Qiang Huo, "Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 Putonghua corpora," In Proc. ISCSLP, pp. 359-362, 2000.
[10] B. H. Juang, L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," Bell System Technical Journal, 64(2), pp. 391-408, 1985.
[11] S. J. Young, P. C. Woodland, "The use of state tying in continuous speech recognition," Eurospeech'93, 1993.
[12] M. Brian, B. Etienne, "Phone clustering using the Bhattacharyya distance," In Proc. ICSLP, pp. 2005- 2008, 1996.
[13] S. M. Yu, S. Hu, S. W. Zhang and B. Xu, “Chinese-English Bilingual Speech Recognition,” NLP-KE’03, 2003. (To appear)
[14] L. Jia and B. Xu, “Include Detailed Information Feature in MFCC for Large Vocabulary Continuous Speech Recognition,” In Proc. ICASSP’02, pp. 805-808, 2002.