# Acoustic Units Selection in Chinese-English Bilingual Speech Recognition

*Lin Yang. Jianping Zhang. Yonghong Yan*

ThinkIT Speech Lab
Chinese Academy of Sciences, Beijing, China
yanglin@h

## Abstract

We present an effective method to merge the acoustic units between Chinese and English to develop a language-independent speech recognition system. Chinese as a tonal language has large differences from English. An optimal Chinese phoneme inventory is set up in order to keep consistent with the representation of English acoustic units. Two different approaches for Chinese-English bilingual phoneme modeling are illustrated and compared. One is to combine the Chinese and English phonemes together based on International Phonetic Association (IPA). The other is a data-driven method on the basis of the confusion matrix. Experimental results show that all these methods are feasible and the data-driven method reduced the WER by 0.73% in Chinese and 3.76% in English relatively compared to the IPA-based method. As a by-product, the idea of data sharing across languages can obtain relative 8.7% error reduction under noise condition.

## 1. Introduction

With the increasing internationalization, research in multilingual speech recognition (MSR) has gained more and more interest in the last few years. But Chinese as one of the most important languages in the world was not considered as much as western languages in some present MSR systems [1][2][3] because of its own peculiarity. So a thorough research on the Chinese-English bilingual recognition is potentially needed, especially the groundwork of a MSR system: the selection of language-independent acoustic units.

So far, there are two main frameworks when solving the MSR problem. One is using a language identification (LID) module to identify the speech as a specific language. Then the specific monolingual utterance can be recognized by language-dependant speech recognizers. The other is using a universal framework including multi-lingual acoustic model, language model and decoder. In the first framework an apparent weak is that the upper bound of performance is limited by the accuracy of LID module. Thus we use the second framework to study how to construct a consistent Chinese-English phoneme inventory and acoustic model.

Linguistically speaking, there are many differences between Chinese and English since they come from two different language families [4]. Firstly, Chinese is a kind of tonal language, including 5 tones; Secondly it is monosyllabic mainly in CV structure where C is consonant and V is vowel, whereas English is a kind of atonal language whose structure is more complex than Chinese. As a result in most cases the basic acoustic units are different largely between the two languages. For example, the Chinese mono phoneme is often represented by initials and tonal finals rather than more subtle units as English. So it is essential to make the acoustic units of the two languages uniform. This paper gives some experiments about how to split the phonemes of Chinese to make it keep consistent with English representations.

Some phonemes across the two languages may be similar enough to be equated. Those resemblances must be merged together for decreasing the number of parameters in the acoustic model. At present there are two main methods of phoneme cluster. One is based on phonetic knowledge; the other is data-driven. Both of the methods are studied to find the most suitable universal phoneme inventory and build a language-independent acoustic model which keeps balance between the number of parameters and system performance in a real MSR framework.

The remainder of this paper is organized as follows. In section 2, an approach of Chinese phone splitting to subtle units and two methods of bilingual phoneme clustering are illustrated in detail. Some experimental results on language independent speech recognizer with different cluster technologies are compared in section 3. Conclusion is given in section 4.

## 2. Building Language-Independent Acoustic Model and MSR System

In order to define a universal phoneme sets for the two languages, we firstly split original Chinese initials and tonal finals into subtle units consistent with the representation of English. If the phonemes of the two languages are put together directly, the final number of acoustical model parameters would be much large and be a burden for decoder. The cluster of phonemes can be performed either manually or automatically. A kind of most common used method of data-driven cluster is based on the direct distance of mono phoneme models. However this method does not consider the context of the phonemes. In order to utilize the information of triphones a new data-driven method based on the concept of confusion matrix is given.

### 2.1. Setup of Chinese phoneme inventory

In our original system the acoustic units are initials and tonal finals, a total number of 213, which is a characteristic of Chinese. In order to be compatible with English atonal phonemes, we discard the tones at first, as a result with the total number of 69. There still exist some compound vowels in the atonal Chinese phoneme inventory although they have resembled to an extent with English units. So splitting according to IPA [5] at various levels is attempted with a number of comparative experiments. Detailed results are shown in table 1. Finally the best Chinese inventory is given with a total number of 49 (including silence and short pause), balanced in size with English phoneme set with the number of 42 (including silence and short pause).

Table 1. Results of recognition in various phoneme sets of Chinese

| | 213 tonal set | 69 atonal set | 57 atonal set | 49 atonal set | 43 atonal set |
|---|---|---|---|---|---|
| Accuracy (%) | 94.9 | 94.1 | 94.2 | 94.4 | 94.0 |

In this group of comparative experiments, the training set includes 70 hours data and the test set is 863test set including 9042 read utterances. From the results we could conclude that the set of 49 phonemes provides the best performance comparative with other atonal sets, although there still exists a disparity from the original tonal phone set.

### 2.2. Experience-based phoneme cluster

After the subtle split of Chinese phonemes, we get two sets of similar phoneme inventories representing the two languages. Then the language-dependent phonemes should be combined into one set in order to realize the language-independent phoneme modeling. A manual and direct way of building bilingual inventory is according to the phonetic knowledge. Some language-dependent phonemes which are represented by the same IPA symbols can be merged into one unit. Table 2 is a list of IPA-based Chinese-English universal phoneme set consisting of 67 units (excluding silence, short pause and garbage model). In this table 19 pairs of phonemes sharing the same IPA symbols are merged, thus the total number of parameters is reduced by more than 21% comparative to the total number of the two sets, which is 89.

Table 2. Phoneme cluster based on IPA

| Lang. | Phonemes | number |
|---|---|---|
| Chinese | p_ch t_ch nn_ch k_ch z_ch c_ch sh_ch r_ch zh_ch ch_ch j_ch q_ch x_ch h_ch a_ch au_ch at_ch e_ch err_ch ix_ch iy_ch v_ch iaa_ch ioo_ch iee_ch iii_ch iuu_ch ivv_ch | 28 |
| English | b_en ch_en d_en dh_en g_en hh_en jh_en r_en sh_en th_en v_en w_en y_en z_en zh_en ah_en ao_en aw_en ay_en oy_en | 20 |
| Merged | b_ch/p_en f_ch/f_en n_ch/n_en g_ch/k_en ng_ch/ng_en d_ch/t_en m_ch/m_en l_ch/l_en s_ch/s_en aa_ch/aa_en ee_ch/eh_en ak_ch/ae_en o_ch/ow_en uu_ch/uh_en u_ch/uw_en ea_ch/ey_en ii_ch/ih_en er_ch/er_en i_ch/iy_en | 19 |

### 2.3. Data-driven phoneme cluster

The data-driven method of cluster is based on the statistical similarity or distance measurement rather than phonetic knowledge. So far the clustering algorithm is mostly applied Bhattacharyya distance [6] as a theoretical similar measure between two Gaussian distributions. However this approach considers only the mono phoneme's distance without any information of context. A novel method of distance measurement with the information of triphones, which is similar to [7][8], is used in this study. The phonemes' distances are calculated according to confusion matrix between English and Chinese phonemes. The confusion is measured by finding the best path for test data of one language on the other's triphone model. The optimization is realized by simple Viterbi decoding and a large number of training utterances are required.

Detailed steps are as follows:

(1) Based on the English triphone HMM model, find the most possible phoneme series for every Chinese utterance using Viterbi decoding. That is

$$\begin{aligned}\widehat{\Lambda}_{en} &= \arg\max_{\Lambda_{en}} P(\Lambda_{en} \mid O_{ch}, M_{trien}) \\ &= \arg\max_{\Lambda_{en}} P(O_{ch} \mid \Lambda_{en}, M_{trien})\end{aligned} \qquad (1)$$

where $\Lambda_{en}$ means the best possible English phoneme series, $M_{engtri}$ means the English triphone model, and $O_{ch}$ represents the Chinese training utterance. Then based on the time information to align $\Lambda_{en}$ and the real phoneme array $\Lambda_{ch}$ by DTW.

(2) A confusion matrix is build by a large number of training data. The degree of confusion is calculated by

$$C_{chi \to enj} = N(\lambda_{enj} \mid \lambda_{chi}) / N(\lambda_{chi}) \qquad (2)$$

where $C_{chi \to enj}$ means the distance of Chinese phoneme $i$ from English phoneme $j$, and $N(\lambda_{chi})$ and $N(\lambda_{enj})$ represent the number of Chinese phoneme $i$ and the number of English phoneme $j$ respectively. Thus a matrix denotes the distance of Chinese phonemes from English phonemes is achieved.

(3) Vice verse. Change the places of English and Chinese and repeat the step 1 and 2 to calculate the distance of English phonemes from Chinese phonemes.

(4) Since the measure is asymmetrical, the average distance is given as follows:

$$C = (C_{chi \to enj} + C_{enj \to chi}) / 2 \qquad (3)$$

In order to compare with IPA-based method, we limited the size of data-driven cluster to 67 mono phonemes. The English training data come from TIMIT corpora including 6300 utterance, and transcriptions with the time information are available. The Chinese data come from 863test sets, including 9042 sentences with the time information transcriptions. Experimental results show that when the training data exceeds 2000 utterances the distance of phonemes in the confusion matrix is stable. Thus the size of our training data can gain the reliable statistical measurement. Table 3 shows the result of bilingual phoneme inventory using data-driven method.

Table 3. Phoneme cluster based on Confusion Matrix Data-Driven

| Lang. | Phonemes | number |
|---|---|---|
| Chinese | ng_ch l_ch t_ch nn_ch z_ch c_ch r_ch j_ch q_ch x_ch h_ch aa_ch ee_ch ak_ch o_ch ii_ch er_ch at_ch e_ch err_ch ix_ch iy_ch v_ch iaa_ch ioo_ch iee_ch iuu_ch ivv_ch | 28 |
| English | ng_en t_en l_en dh_en hh_en r_en th_en v_en z_en zh_en eh_en ae_en uh_en uw_en ih_en er_en ah_en ao_en aw_en oy_en | 20 |
| Merged | a_ch/aa_en au_ch/ay_en b_ch/b_en ch_ch/ch_en d_ch/d_en ea_ch/ey_en f_ch/f_en g_ch/g_en i_ch/iy_en iii_ch/y_en k_ch/k_en m_ch/m_en n_ch/n_en p_ch/p_en s_ch/s_en sh_ch/sh_en u_ch/u_en uu_ch/ow_en zh_ch/jh_en | 19 |

## 3. Experiments and Discussion

The goal of the experiment is to evaluate the performance of our language-independent LVCSR system and compare the cluster method based on IPA with our proposed approach based on confusion matrix data-driven.

### 3.1. Corpora and experiment setup

The training data has about 340 hours Chinese speech data (including various dialects) and 160 hours English data (including TIMIT, WSJ). These data are recorded under relatively clean acoustic conditions. The recognizer makes use of continuous density HMM with Gaussian mixture for acoustic model. Each triphone model is a 3-state left-to-right with Gaussian mixture observation densities (typically 32 components). The acoustic feature of speech is MFCC with 39 dimensions.

The size of vocabulary is 43K for Chinese and 50K for English. The bilingual dictionary is composed of the pooled monolingual dictionaries and consists of 93K entries. The training corpora of bilingual language model (BILM) are the combination of Chinese and English data. In order to test the performance of language-independent acoustic model we also build monolingual LM, labeled as MONOLM.

We selected two sets of Chinese testing data in order to evaluate the performance under various situations. One is recorded under clean acoustic condition, including 341 Chinese sentences labeled as TestCH1, while the other is recorded under noise condition consisting of 200 sentences labeled as TestCH2. The English testing data is standard WSJ0 testing set labeled as TestEN, including 330 English utterances.

### 3.2. Experiments and discussion

For the purpose of comparing the language-independent system with the language-dependent system, we conducted the experiments on the monolingual LVCSR system. Table 4 shows the word error rates of the baseline on different testing sets.

Table 4. The monolingual results on different testing sets

| | TestCH1 | TestCH2 | TestEN |
|---|---|---|---|
| WER (%) | 25.8 | 36.8 | 9.7 |

In order to compare the two methods of cluster we trained the acoustic model on the two bilingual phoneme sets, labeled as IPA and CMDD. They are combined with different language models MONOLM and BILM. The comparative results are shown in Table 5.

**Table 5.** The comparative results of language-independent models

| | TestCH1 (%) | TestCH2 (%) | TestEN (%) |
|---|---|---|---|
| IPA-MONOLM | 27.1 | 36.0 | 10.8 |
| CMDD-MONOLM | 26.8 | 36.0 | 11.1 |
| IPA-BILM | 27.4 | 36.4 | 13.3 |
| CMDD-BILM | 27.2 | 36.2 | 12.8 |

From the table 4 and table 5, we can see that the bilingual system can achieve comparable performance to the monolingual system whether in English or Chinese testing sets. In the worst cases the WER increased 1.6% in Chinese and 3.6% in English respectively. This may be due to the dramatic increase of the size of bilingual dictionary, with nearly 10K entries.

By comparing the IPA cluster method with CMDD method in table 5, we can see that the CMDD method accepts moderate improvements in various testing conditions except combined with the monolingual language model in TestEN set. This loss may result both from the unmatchable acoustic and language model and from the asymmetric size of training data between English and Chinese. But on English testing set, using bilingual LM a significant improvement can be observed, with WER relative reduction by 3.76% than IPA-based method. Although the improvement is not dramatic as a whole, the advantage of CMDD is evident whether for adjusting the size of phoneme inventory or from the theoretical foundations. As a result, the method of confusion matrix data-driven gives us a promising belief that a great improvement can be achieved by conditioning the size of universal phoneme inventory and bilingual dictionary.

It is interesting that in the universal system the performance is improved greatly under noise conditions TestCH2, which WER is from 36.8% to 36%. For confirming the contribution of bilingual acoustic models which share data across languages we put them into monolingual systems, avoiding the influences of merged dictionary and language model. The results are listed in Table 6. These results demonstrated that the shared data across language can deduce the WER by %8.7 relatively, making the recognizer more robust under noise condition.

**Table 6:** Acoustic model performance comparison under noise conditions

| | Monolingual baseline | IPA bilingual model | CMDD bilingual model |
|---|---|---|---|
| WER(%) | 36.8 | 33.6 | 34.2 |

## 4. Summary and Future Work

In this paper, we presented the work of setting up Chinese phoneme inventory and two methods of building language-independent MSR system. By comparing the two cluster sets on different testing data and language models, the CMDD cluster method outperforms the IPA-based approach as a whole. As a by-product, the sharing data across languages provides us a new idea to improve the performance of recognizer under noise condition. In future experiments, we will try various methods to build bilingual language model and select the optimal size of universal phoneme inventory. It is also a challenging task about how to combine the experience-based method and data-driven method. This approach proposed in this paper could be generalized to other languages.

## 5. Acknowledgements

This work is supported by Chinese 973 program (2004CB318106), National Natural Science Foundation of China (10574140, 60535030)

## 6. References

[1] Zhirong Wang, Umut Tokpara, Tanja Schultz, Alex Waibel: Towards Universal Speech Recognition, Proceedings of the Fourth International Conference on Multimodal Interface (ICMI), IEEE Computer Society, (2002), 247-252

[2] F. Weng, H. Bratt, L. Neumeyer, A Stolcke: A Study of Multilingual speech recognition. In Proc. Eurospeech’97, Rhodos, Greece (1997)

[3] Ulla Uebler: Multilingual speech recognition in seven languages. Speech Communication. 35(2001), 53-69

[4] D. Lyu, R.Lyu, Y.Chiang, C.Hsu: Speech Recognition on Code-switching Among Chinese Dialects, ICASSP, I(1105-1108), (2006)

[5] IPA: The International Phonetic Association (revised to 1993) IPA Chart. Journal of the International Phonetic Association 23, (1993).

[6] M. Brian, B.Etienne: Phone Clustering Using the Bhattacharyya Distance, In Proc. ICSLP, 2005-2008, (1996)

[7] R.Bayeh, S. Lin, G. Chollet, C. Mokbel: Towards Multilingual Speech Recognition Using Data Driven Source/Target Acoustical Units Association, ICASSP’04, 521-524, (2004)

[8] E. Wong, T. Martin, T. Svendsen, S. Sridharan: Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modeling Techniques, EUROSPEECH’03, Geneva, 3133-3136, (2003)