MANDARIN-ENGLISH BILINGUAL SPEECH RECOGNITION FOR REAL WORLD MUSIC RETRIEVAL

Qingqing Zhang,Jielin Pan and Yonghong Yan

ThinkIT Speech Laboratory
Institute of Acoustics Chinese Academy of Sciences
Beijing 100080, China

ABSTRACT

This paper presents our recent work on the development of a grammar-constrained, Mandarin-English bilingual Speech Recognition System (MESRS) for real world music retrieval. In order to balance the performance and the complexity of the bilingual SR system, an unified single set of bilingual acoustic models derived by phone clustering is developed. A novel Two-pass phone clustering method based on Confusion Matrix (TCM) is presented and compared with the log-likelihood measure method. In order to deal with the Mandarin accent in spoken English, different non-native adaptation approaches are investigated. With the effective incorporation of approaches on phone clustering and non-native adaptation, the Phrase Error Rate (PhrER) of MESRS for English utterances was reduced by 24.5% relatively compared to the baseline monolingual English system while the PhrER on Mandarin utterances was comparable to that of the baseline monolingual Mandarin system, and the performance for bilingual code-mixing utterances achieved 22.4% relative PhrER reduction.

Index Terms— Bilingual speech recognition, clustering methods, information retrieval

1. INTRODUCTION

For decades, bilingual communication becomes a common phenomenon as a result of globalization. It presents a new challenge to the real world applications of speech recognition technology. In recent years, research on bilingual speech recognition has made significant progress. [2] focused on English-German SR and [3] investigated Chinese-English SR. One of the commonalities among these studies is that those test corpora used in their experiments consist of monolingual utterances spoken by corresponding native speakers. Although these bilingual systems achieved respectable performances for native monolingual speakers, their performances could degrade badly on non-native utterances. For many applications, a bilingual system has to face the fact that many users who take their native language as matrix language are not native to the embedded language[1]. Therefore, improving performance on the non-native embedded language is needed before these systems can be put into practical use [4].

In order to deal with the non-native speech recognition, [5] proposed a lexical modeling technique to improve non-native speech recognition, but only achieved modest reduction in word error rate. [6] developed a Cantonese-English bilingual speech corpus. In this corpus, non-native utterances were collected from the corresponding non-native region, and were directly used to train the acoustic models. These models greatly improved the recognition performances on the non-native test sets. However, to obtain sufficient non-native training data is very difficult. When training data is limited, the performance of non-native acoustic models will degrade significantly.

Thus, how to improve system performance with only limited amount of non-native data becomes an important issue. Speaker adaptation techniques such as MAP and MLLR have been used to adapt acoustic models trained with native speech to handle non-native utterances. [7] compared the effectiveness of several adaptation techniques on non-native speech, and consistent improvements were confirmed. These activities mainly focused on solving the accent issue, and their improvements, however, were generally accompanied by the degradation of the recognition on native speech.

In this paper, we present our efforts in developing MESRS for real world application (Color Ring Back Tone services for China Mobile Co.). Our goal is to yield decent performance on non-native English with only limited amount of non-native English data while maintaining the highest possible recognition rate on Mandarin. In order to process language switching and reduce computing resource requirements for practical reason, only one set of Mandarin-English bilingual acoustic models is developed by clustering the phone sets of these two languages. In order to deal with the accent issue with limited amount of non-native training data, different approaches are explored as well. Experiment results show that encouraging advances are made compared to the baseline system.

The paper is structured as follows: The database is presented in section 2. In section 3, we describe the baseline system of our experiments and in section 4 we document how bilingual acoustic modeling and non-native adaptation can help to improve the recognizer performance. Section 5 gives a brief conclusion of this paper.

2. BILINGUAL CORPORA

This section briefly describes the data resources and feature analysis used for the development. All the speech data are recorded through telephone lines and digitized at 8 KHz sampling rate with 16-bit resolutions. The speech feature vector consists of 36 components (12 MFCC parameters, and their first and second order time derivatives), which is analyzed at a 10msec frame rate with a 25msec window size. Cepstral Mean Subtraction (CMS) is employed.

2.1. Training corpora

Our training corpora are divided into three categories: the native Mandarin corpus (labeled as TrainM), the native English corpus (labeled as TrainE) and the Mandarin accented English corpus (labeled as TrainB). TrainM consists of 865 hours' native Chinese speech from National 863 Hi-Tech Project. It is a standard corpus published by governmental research program 863 for read speech in Mandarin. The 232 hours' TrainE is Wall Street Journal etc. TrainB was collected in house. It includes 24 hours' English speech data from 60 male and 60 female Mandarin residents. Each speaker contributed same 200 English utterances from everyday conversations.

This work is partially supported by MOST (973program,2004CB318106), National Natural Science Foundation of China (10574140, 60535030), The National High Technology Research and Development Program of China (863 program, 2006AA010102,2006AA01Z195).

[1]Matrix language can be identified as the main language of the speaker or the language in which the morphemes or words are more frequently used and the other languages are considered as the embedded languages, according to Myers-Scotton's Matrix Language Frame model[1].

2.2. Testing corpora

The task domain is music retrieval. Our system enables users to find a song by simply saying the singer's name or the title of the song, which are allowed to be either monolingual or bilingual. In our testing corpora, there are 10179 utterances of names of singers and titles of songs in total, which consist of 8183 mono- Mandarin utterances, 1650 mono- English utterances and 346 bilingual utterances respectively. The corpora were collected under realistic conditions such as in restaurants, streets and other noisy places, which cover variations in background noise, microphones, volumes, speaker fluency and accents. The examples of these three types of test utterances can be found in Table 1.

Table 1. Summary of three test corpora.

Test Corpus	Language	No. of utterances	Example (song; singer)
TestM	Mandarin	8183	甜蜜蜜; 张学友
TestE	English	1650	Hey jude; Madonna
TestB	Bilingual	346	Hello朋友; Newz乐团

3. BASELINE SYSTEM

3.1. Baseline monolingual system

The mono- Mandarin acoustic model (Model_M) and mono- English acoustic model (Model_E) trained with TrainM and TrainE respectively are used in our baseline system. All the acoustic models used in our paper are state clustered crossword triphone HMMs with 32-component Gaussian mixture output densities per state. Model_M comprises 5886 states and Model_E comprises 5829 states. The English phone set is supplied by the ARPABET and the dictionary is based on CMU pronunciation dictionary [8]. This dictionary consists of approximately 53000 words with associated phonetic transcriptions. As Mandarin is a tonal language, incorporating the tone markers into the acoustic models could improve the system performance, so 179 Initials and tonal Finals are selected to form the Mandarin phone set. The Mandarin dictionary contains 25000 words, which are formed by isolated Chinese characters. Table 2 shows the performances of the baseline acoustic models. Please note that these two separate monolingual models cannot recognize the bilingual utterances directly. Thus, only results on their corresponding languages are presented.

3.2. Baseline bilingual system

For the sake of comparison, a baseline bilingual acoustic model was established first, whose phone set was created by simply combining the 179 Mandarin tonal phones and 42 English toneless phones without clustering. We call it "Model_ME". Table 2 shows that the PhrER of the Model_ME bilingual acoustic model on TestB is 16.76%, which is taken as the baseline result for TestB.

Table 2. Performances on three corpora by the Baseline monolingual and bilingual acoustic models

PhrER(%)			
Acoustic Model	TestM	TestE	TestB
Model_M	20.9	–	–
Model_E	–	45.3	–
Model_ME	25.2	47.7	16.8

As shown above, even though Model_ME can deal with the bilingual utterances (TestB), the performances on Mandarin and English utterances decrease drastically. On the other hand, because of the direct combination of Mandarin and English phone set, the number of model parameters of Model_ME expands rapidly. This leads to a large, insufficiently trained acoustic model and slows down the recognition speed. Another noticeable point is the large performance gap between TestM and TestE (20.9% V.S. 45.3%)since all the test utterances are from Mandarin accented speakers. In order to deal with these problems, cross-lingual phone clusterings and non-native adaptations were investigated and will be reported in the following section.

4. PHONE CLUSTERING AND NON-NATIVE ADAPTATION

4.1. Phone clustering

For bilingual speech recognition, especially for intra-sentential switching, it is very important to determine a global phone set for different languages involved in the system. Since for inter-sentential switching, instead of using one recognizer, a system can first use language identification technology [11] to determine which language is being used, then uses recognizer for that language to conduct the recognition. For intra-sentential switching, using language identification first is at least computationally not feasible for real world applications. Adopting a global phone set can also reduce the amount of data required to robustly estimate statistical models. In our case, English words involved are Mandarin accented, which have non-native pronunciation variations. Flege et al. [9] argues that non-native speakers may produce speech sounds which are either part of their own native language or which are established via merging characteristics of a native sound with a non-native speech sound. Based on this, one can speculate that a suitable phone set resulting of merging and clustering of phones in these two languages may efficiently handle the Mandarin accents in English words, since the combination merges the characteristics of the two languages. In this section, different approaches to phone clustering are compared and evaluated, and the approach with the best performance on the development set was selected for our final system.

Recent approaches to phonetic clustering can be roughly divided into two categories: knowledge-based and data-driven. Since data-driven methods outperformed knowledge-based ones consistently [2] [3], only the data-driven approaches are explored in the paper. Several phone clustering algorithms based on data-driven approaches have been investigated[3] [12] [14]. In [3], the clustering approach based on log-likelihood measure is explored between Chinese and English phones. In our research, a novel phone clustering algorithm, which is a Two-pass process based on Confusion Matrix (TCM) is proposed. The two different approaches of phone clustering are investigated and compared.

4.1.1. Log-Likelihood measurement

For log-likelihood (LL) approach [3], a similarity between two phone models has to be defined. The distance between two phone models λ_i and λ_j is:

$$L(\lambda_i, \lambda_j) = f(\overrightarrow{X_i}|\lambda_j)^\alpha / \sum_{k=1}^{n} f(\overrightarrow{X_i}|\lambda_k)^\alpha \quad (1)$$

Where $\overrightarrow{X_i}$ denotes a sequence of observations labeled as phone i . $f(\overrightarrow{X_i}|\lambda_j)$ 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. Since the distances are not symmetric, the average distance can be calculated as follows:

$$L = \frac{1}{2}(L(\lambda_i, \lambda_j) + L(\lambda_j, \lambda_i)) \quad (2)$$

Distances between phones are repeatedly calculated based on this measure and phones with the minimum distance are merged until the desired number of phone classes is reached. Finally, the single bilingual phone set is obtained.

4.1.2. TCM

TCM is a phone clustering approach similar to automatic phone mapping method using confusion matrix [13] usually used in fast acoustic modeling for a new target language. We proposed a new bilingual phone clustering algorithm based on this method. TCM is a two-pass approach. In each pass, Mandarin and English take turns as the source language and the target language. The detailed algorithm is described as follow:

1. Target reference: Force-align target language speech utterances using target language acoustic model to get the time-label information. The resulting time-aligned phone strings are considered as the target phone references.
2. Source hypothesis: The source language phone recognizer is applied to these utterances to obtain the phonetic transcriptions. This yields parallel phonetic segmentations of the target language acoustic data in the source language phone inventories. This source phonetic representation is considered as the source phone hypothesis.
3. Co-occurrence criterion: Define a criterion for co-occurrence between two phonetic labels of the reference and hypothesis. In our implementation, when the number of overlapping frames between the reference and hypothesis is more than half of the reference phone duration, we arrange the phones of the target and source language into a matrix that contains the counts of co-occurrences between the i^{th} and j^{th} phones of the source and target languages. This matrix of co-occurrences is the confusion matrix [12]. Figure 1 shows an example of the co-occurrence between phone "au_ch" and phone "ay_en" when Mandarin is taken as the target language. (Note: the Mandarin phones and the English phones are labeled by tag "_ch" and "_en" respectively)

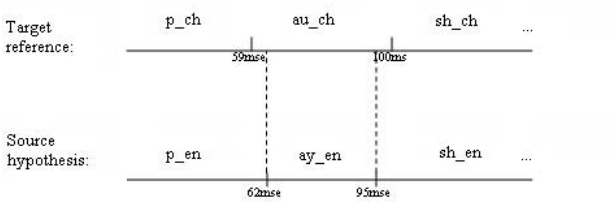

Fig. 1. Example of the co-occurrence between phone "au_ch" and phone "ay_en" when Mandarin is taken as the target language.

4. Calculation of confusion probability: Let M, N be the numbers of phones in source and target language respectively. Let $\mathrm{A_{S,T}(M,N)}$ be the confusion matrix and $\mathrm{A_{i,j}}$ be the i^{th} row and j^{th} column element of this matrix. Given the target language phoneme $\mathrm{t_j}$ and the source language phoneme $\mathrm{s_i}$, the confusion probability can be computed as:

$$\mathrm{A_{i,j}}=\frac{\mathrm{count(t_j|s_i)}}{\sum_{\mathrm{n=1}}^{\mathrm{N}}\mathrm{count(t_n|s_i)}} \tag{3}$$

where

$$\mathrm{A_{i,j}} \in \mathrm{A_{S,T}(M,N)}, i = 1...M, j = 1...N.$$

5. The final confusion matrix: How to obtain a confusion matrix given that the source language (Mandarin or English) has been introduced already. We exchange the target and source languages, which means the old target language would become the new source language and the old source language would become the new target one, then go back to step 1 to calculate the second confusion matrix. After the two-pass process, we have two matrixes $(\mathrm{A_{man,eng}}, \mathrm{A_{eng,man}})$. The final confusion matrix after two-pass process is calculated as:

$$\mathrm{A_{TCM}}=\frac{1}{2}(\mathrm{A_{man,eng}}+\mathrm{A_{eng,man}}^{T}) \tag{4}$$

Our application assumed the Mandarin and English models have the equal importance, so the weight 1/2 in Eq.4 has been selected to balance. After the final confusion matrix $\mathrm{A_{TCM}}$ is obtained, the clustering information can be derived from this matrix. If the i^{th} row and j^{th} column element of $\mathrm{A_{TCM}}$ has the largest value among all the elements, it means that the i^{th} phone and the j^{th} phone from corresponding languages have the maximum similarity, thus the i^{th} phone and the j^{th} phone from two languages will be clustered into one class. Then the i^{th} row and j^{th} column are removed from $\mathrm{A_{TCM}}$, the entry with the largest value among the rest elements is found and the corresponding phones will be clustered. This clustering procedure continues until the desired number of phone classes is reached.

With different clustering methods mentioned above, we can use obtained phone classes to map the language-dependent phone models to the corresponding bilingual inventory. The bilingual dictionary, question list for decision tree and transcriptions are also processed with the clustering information. Then the bilingual acoustic models will be retrained based on these bilingual training files.

Since English phone set is toneless, the tone markers of Mandarin phones were removed before clustering. Considering the phonetic inventory of IPA [10] is toneless, the Mandarin 179 tonal phones are split and mapped into this inventory. The information reported below is based on the mapped 49 toneless Mandarin phones and 42 toneless English phones including Short Pause (sp), Silence (sil) and garbage.

Table 3. Performances of different phone clustering approaches.

PhrER(%)			
Acoustic Model	TestM	TestE	TestB
Model_ME	25.2	47.7	16.8
Model_LL70	21.6	44.5	16.5
Model_TCM70	21.5	42.1	14.5

Table 3 shows the performances of acoustic models with LL and TCM approaches. Model_LL70 and Model_TCM70 refer to the acoustic models whose phone sets are clustered into 70 classes with LL measure and TCM respectively. As shown, phone clustering approaches (LL, TCM) achieved a significant improvement compared to the direct combination of monolingual phone inventories (Model_ME). Further more, phone clustering by TCM reaches even lower PhrERs on all the three corpora when compared to LL measure. Results of comparative experiment indicate that the proposed TCM outperforms LL algorithm favorably.

4.2. Non-native Adaptation

By clustering Mandarin and English phones with TCM into 70 classes[2], a bilingual acoustic model (Model_TCM70) was trained with native

[2]Three different numbers of phones (50, 70 and 89) in the clustered phone sets have been investigated in our experiment,which stand for three representative degree of clustering. Results show that instead of other numbers of

speech from each language. The non-native speakers' pronunciations, however, different from those native speakers' pronunciation observed during system training, drastically decreases the recognition performance(TestE). With the native bilingual acoustic model in hand, the challenge for non-native speech recognition is to maximize the recognition performance with the available small amount of non-native data.We investigate speaker adaptation such as MAP and model retraining method, and compare their impacts on the performances of native Mandarin and accented English test corpora respectively.

In MAP adaptation, the native model parameters are re-estimated individually, using held-out non-native adaptation data. An updated mean is then formed by shifting the original native value toward the non-native sample value. If there is insufficient adaptation data to reliably estimate the sample mean of a phone, no adaptation is performed. Model retraining method is a compromise settlement. Since the bilingual acoustic models consist of clustered phones which are language-independent, sharpening the acoustic models on non-native training data may move further away from the native speakers. Therefore, compared to MAP, appending the pool of non-native adaptation data in training process is implemented as a compromise.

Table 4. Recognition results of Non-native adaptation using the TrainB data set.

PhrER(%)			
Acoustic Model	TestM	TestE	TestB
Model_TCM70	21.5	42.1	14.5
Model_TCM70_MAP	24.6	29.9	11.0
Model_TCM70_Retrain	21.6	34.2	13.0

Table 4 presents the recognition results of different adaptation methods. The clustered baseline bilingual model used for comparison is Model_TCM70 which is trained by native training data (TrainM and TrainE). With 24000 adaptation utterances of TrainB, MAP transforms were conducted. The results show that with a pool of non-native adaptation data collected in advance, MAP can substantially improve the performance on new non-native speakers (TestE). However, the performance on native Mandarin test set (TestM) degrades seriously, which gives a 14.4% relative increase in PhrER compared to Model_TCM70. As a whole, the model retraining method achieved 18.6% and 10.0% relative reduction in PhrER on TestE and TestB respectively while little degradation was observed for TestM corpus.

Finally, the optimal approaches mentioned above are integrated to form our final MESRS. Figure 2 presents the PhrERs of MESRS are reduced by 24.5% and 22.4% relatively for TestE and TestB compared to the corresponding baseline systems, and the performance on TestM is comparable to that of the baseline Mandarin only acoustic model (Model_M).

5. CONCLUSION

This paper presents our recent work in developing a Mandarin-English bilingual speech recognition prototype system for real world music retrieval. In addition to the requirement of handling inter- and intra-sentential language switching at the same time, the challenge also includes that only limited amount of out-of-task-domain accented English data is available. Since this type of recognition task has not been reported before, various cross language phone clustering methods and adaptation techniques have been investigated. A novel method named TCM is proposed. Experiment results showed that the proposed TCM outperforms existing approaches favorably. It is also shown that with limited availability of non-native adaptation data, model retraining method outperforms MAP adaptation, since it improves the recognition of non-native English while maintaining the performance on Mandarin. Although the task is domain specific, we believe the research findings presented in this paper can be applicable to other multi-lingual recognition task as well.

classes, clustering phone into 70 classes gives the best performances on the three test corpora.

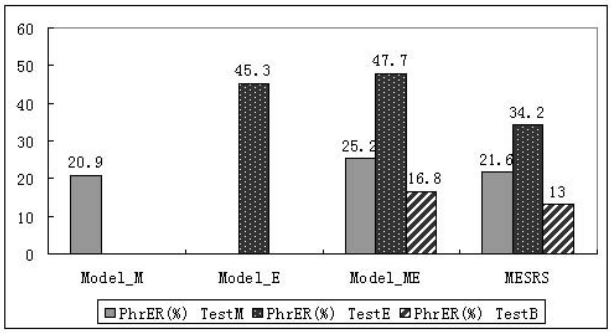

Fig. 2. Comparison of acoustic modeling performances.

6. REFERENCES

[1] C. Myers-Scotton, "Duelling languages: Grammatical structure in codeswitching", 1993. Oxford: Clarendon Press.

[2] Z. Wang, U. Topkara, T. Schultz, A. Waibel, "Towards Universal Speech Recognition", Proc. ICMI 2002.

[3] S. Yu, S. Zhang, B. Xu, "Chinese-English bilingual phone modeling for cross-language speech recognition", International Conference on Natural Language Processing and Knowledge Engineering, pp. 603-609, 2003.

[4] H. Ye and S. Young,"Improving the Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning",Interspeech 2005, Lisbon, Portugal

[5] K. Livescu, J. Glass, "Lexical modeling of non-native speech for automatic speech recognition", Proc. ICASSP 2000.

[6] Y.C. Chan, P.C. Ching, T. Lee and H. Cao "Automatic speech recognition of Cantonese-English Code-Mixing utterances", 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), pp. 113 - 116, Pennsylvania, USA, September 17-21, 2006.

[7] Z. Wang, T. Schultz, A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech", Proc. ICASSP 2003.

[8] The CMU Pronuncing Dictionary v0.6, The Carnegia Mellon University, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[9] O.-S. Bohn, J.E. Flege, "The production of new and similar vowels by adult German learners of English". Stud. Second Lang. Acquis. 14, 131-158, 1992.

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

[11] Y. Yan, E. Barnard, R.A. Cole. "Development of an Approach to Automatic Language Identification based on Phone Recognition". In Computer, Speech Language, Volume 10(1), Pages 37-54, January, 1996

[12] P. Beyerlein et al., "Towards language independent acoustic modeling", ASRU'99, Keystone, CO, USA, December 1999.

[13] V. B. Le, L. Besacier, "First steps in fast acoustic modeling for a new target language: application to Vietnamese", ICASSP'05, vol. 1, pp. 821-824, Philadelphia, PA, USA, March 2005.

[14] R. Bayeh et al., "Towards multilingual speech recognition. using data driven source/target acoustical units association",. ICASSP'04, vol. I, pp. 521-524, Montreal, Canada, May 2004.