Development of a Mandarin-English Bilingual Speech Recognition System for Real World Music Retrieval

Qingqing ZHANG^{†a)}, Jielin PAN[†], Yang LIN[†], *Nonmembers*, Jian SHAO[†], *Student Member*, and Yonghong YAN^{†*b)}, *Nonmember*

SUMMARY In recent decades, there has been a great deal of research into the problem of bilingual speech recognition - to develop a recognizer that can handle inter- and intra-sentential language switching between two languages. 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. Two of the main difficult issues in handling the bilingual speech recognition systems for real world applications are tackled in this paper. One is to balance the performance and the complexity of the bilingual speech recognition system; the other is to effectively deal with the matrix language accents in embedded language^{**}. In order to process the intra-sentential language switching and reduce the amount of data required to robustly estimate statistical models, a compact single set of bilingual acoustic models derived by phone set merging and clustering is developed instead of using two separate monolingual models for each language. In our study, a novel Two-pass phone clustering method based on Confusion Matrix (TCM) is presented and compared with the log-likelihood measure method. Experiments testify that TCM can achieve better performance. Since potential system users' native language is Mandarin which is regarded as a matrix language in our application, their pronunciations of English as the embedded language usually contain Mandarin accents. In order to deal with the matrix language accents in embedded language, different non-native adaptation approaches are investigated. Experiments show that model retraining method outperforms the other common adaptation methods such as Maximum A Posteriori (MAP). With the effective incorporation of approaches on phone clustering and non-native adaptation, the Phrase Error Rate (PER) of MESRS for English utterances was reduced by 24.47% relatively compared to the baseline monolingual English system while the PER on Mandarin utterances was comparable to that of the baseline monolingual Mandarin system. The performance for bilingual utterances achieved 22.37% relative PER reduction.

key words: *bilingual speech recognition, phone clustering, TCM, log-likelihood measure, non-native adaptation, Mandarin-English*

1. Introduction

With the globalization in modern society, bilingual or multilingual communication becomes a common phenomenon. This 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 bilingual recognition and [3] described a Slovenian-Croatian weather forecast

system. Similarly, [4] investigated Chinese-English speech recognition. 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 some 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. Therefore, improving performance on the non-native embedded language is needed before these systems can be put into practical use [5]. Studies on the non-native speech recognition have focused on two basic approaches, one based on pronunciation modeling and the other based on acoustic modeling.

In pronunciation modeling approach, Livescu and Glass [6] proposed a lexical modeling technique to improve non-native speech recognition. Similarly, Amdal and Korkmazskiy [7] used joint pronunciation modeling to incorporate non-native pronunciations into the lexicon. Both methods used data driven techniques to derive a multiple pronunciation lexicon, and only achieved modest reduction in word error rate.

For acoustic modeling approach, Chan et al. [8], [9] and Marino et al. [10] developed Cantonese-English and Spanish-Catalan bilingual speech corpora respectively. In these corpora, utterances for the embedded language were collected in regions where the matrix language was spoken, and were directly used to train the acoustic models of the embedded language. Compared to the monolingual models originally trained with native speakers' utterances, these acoustic models greatly improved the recognition performances on the embedded language. However, to obtain sufficient non-native training data is very difficult. When training data is limited, the performances of these non-native acoustic models can degrade significantly due to insufficient training.

Thus, how to improve system performance with only limited amount of data becomes an important issue. Speaker adaptation techniques such as Maximum A Pos-

Manuscript received July 4, 2007.

Manuscript revised September 10, 2007.

[†]The authors are now with ThinkIT Speech Laboratory, Institute of Acoustics Chinese Academy of Sciences Beijing, China, 100080.

*Correspondence author

a) E-mail: zhangqingqing@hcll.ioa.ac.cn

b) E-mail: yyan@hcll.ioa.ac.cn

DOI: 10.1093/ietisy/e91-d.3.514

^{**}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].

teriori (MAP) and Maximum Likelihood Linear Regression (MLLR) have been used to adapt acoustic models trained with native speech to handle non-native utterances. Tomokiyo and Waibel [11] studied adaptation methods for non-native speech and found that substantial gains could be obtained. Wang et al. [12] 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.

With globalization, China is more and more closely connected to the world. Foreign languages (particularly English) are being used more and more frequently, especially among young generations in cities. It is very common that English words are being embedded in Chinese sentences during conversations among the well-educated. This makes the Mandarin-English bilingual speech recognition a necessity for many speech recognition applications. In this paper, we will present our efforts in developing a Mandarin-English bilingual speech recognition system for a real world application.

Our goal is to develop a system that can yield decent performance on non-native English with only limited amount of non-native English data while maintaining highest possible recognition rate on Mandarin. The novelties of our work include: (1) a single set of acoustic models that can handle both inter- and intra-sentence language switching, (2) a two-pass cross languages phone clustering approach based on confusion matrix, and (3) optimal way of merging the phone sets of English and Mandarin via extensive experiments. Although the task is domain specific, the methodologies presented in this paper can be applied to general recognition task as well.

The task domain of our MESRS is music retrieval. The system enables a user to find a song by simply saying the singer's name and the title of the song through a phone call. Since many songs are English songs (and sometimes performed by Chinese singers), thus the incoming calls may consist of songs and singers' names uttered either monolingually (Mandarin or English) or bilingually. So the system has to deal with input with inter- and intra-sentential language switching. The following are typical examples of these phrases: 在那遥远的地方, backstreet boys (inter-sentential language switching); One night in 北京, 阿 ben (intra-sentential language switching).

In order to process language switching and reduce computing resource requirements for practical reasons, only one set of Mandarin-English bilingual acoustic models is developed in our study. By merging and clustering the phone sets of these two languages, a new set of monophones covering both languages are determined. Different sizes of Mandarin-English bilingual phone sets are experimented and compared. In order to deal with the accent issue with limited amount of non-native training data, different approaches are explored as well. Comparative studies were conducted, and only the most effective methods were se-

lected. Experiment results show that encouraging advances are made compared to the baseline system.

2. Bilingual Corpus

This section briefly describes the data resources and feature analysis that were used for all the experiments presented in this paper. These speech corpora are divided into three categories: the native Mandarin corpus, the native English corpus and the Mandarin accented English corpus. All the speech data were recorded through telephone lines and digitized at 8 KHz sampling rate with 16-bit resolutions. The speech feature vector used throughout this paper consists of 36 components (12 MFCC parameters, and their first and second order time derivatives), which is analyzed at a 10 msec frame rate with a 25 msec window size. Cepstral Mean Subtraction (CMS) was employed.

2.1 Training Corpus

The native Mandarin training corpus consists of native Chinese speech corpus of National 863 Hi-Tech Project (DB863). It is a standard corpus published by governmental research program 863 for read speech in Mandarin. The English training corpus is Wall Street Journal (WSJ). These corpora were recorded in higher sampling rate and were band limited to 4 KHz by down sampling. The Mandarin accented English corpus was collected in house (labeled as GTC (General Telephone Conversation) corpus). It includes speech data from 60 male and 60 female speakers. Each speaker contributed same 200 English utterances from everyday conversations. Table 1 summarizes the main information about these three corpora.

2.2 Testing Corpus

The test data contains utterances of names of singers and titles of songs. There are 10179 utterances in total, which consist of 8183 mono- Mandarin utterances, 1650 mono-English utterances and 346 bilingual utterances respectively. The examples of these three types of test utterances can be found in Table 2.

Compared with the quiet lab environment, the environment of real world applications can be extremely variable. The test utterances 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.

Table 1 Summary of three training corpora.

Training Corpus	Type	Source	Time (hour)
TrainM	Native Mandarin	DB863	865
TrainE	Native English	WSJ	232
TrainA	Mandarin accented English	GTC	24

Table 2 Summary of three test corpora.

Test Corpus	Language	No. of utterances	Example (song; singer)
TestM	Mandarin	8183	在那遥远的地方; 张学友
TestE	English	1650	Hey jude; Backstreetboys
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 the Mandarin corpus (TrainM) and the English corpus (TrainE) respectively are used in our baseline system. All the acoustic models used throughout our paper are state clustered cross-word 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 [13]. This dictionary consists of approximately 53,000 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 25,000 words, which are formed by isolated Chinese characters.

Table 3 shows the performances of the baseline acoustic models. This yields the Phrase Error Rate (PER) of 20.91% on Mono- Mandarin corpus (TestM) and 45.33% on Mono- English corpus (TestE). It is noticeable that the performance on Mono- English test corpus (TestE) is much worse than that of the Mono- Mandarin one (TestM). This is because in our test, the English contains strong Mandarin accent, and the Mono- English acoustic model (Model_E) was trained by native English corpus. There is a huge mismatch between the pronunciations in dictionary and the actual acoustic realizations in those Mandarin accented test utterances. (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 (language-dependent phones) into one set. We call it "Model_ME" bilingual acoustic model. The performance is also compared with those of baseline monolingual acoustic models in Table 3.

Table 3 Performances on three corpora by the baseline monolingual and bilingual acoustic models.

PER (%)			
Acoustic Model	TestM	TestE	TestB
Model_M	20.91	—	—
Model_E	—	45.33	—
Model_ME	25.17	47.68	16.76

Table 3 shows that the PER of the Model_ME bilingual acoustic model on TestB is 16.76%, which is taken as the baseline result for test set TestB. However on the monolingual test corpora, the performance of this bilingual baseline system degrades significantly, especially on the Mandarin test corpus, which gives a 20% relative increase in PER over the baseline Mandarin acoustic model (Model_M). How to reduce this large performance gap is one of the original motivations of the work presented in this paper.

4. Attempts Made to Improve the Baseline System

As shown above, even though the baseline bilingual acoustic model (Model_ME) can deal with the bilingual utterances (TestB), the performances on Mandarin and English utterances decrease drastically when compared to baseline monolingual models. 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 since the English test utterances are Mandarin accented while the Mandarin ones are native. In order to deal with these problems, cross-lingual phone clusterings and adaptations were investigated. The findings will be discussed in the following sections.

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 [17] 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. [14] 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.

4.1.1 Clustering Algorithm

Recent approaches to phonetic clustering can be roughly divided into two categories: knowledge-based approaches and data-driven approaches. Since knowledge-based approaches do not consider the statistical similarities between phones while the data-driven ones do, [2] and [4] demonstrated that data-driven methods outperformed knowledge-based ones consistently. Thus only the data-driven approaches are explored in our study.

Several phone clustering algorithms based on data-driven approaches have been investigated [16], [18], [20]. In [4], the clustering approach based on log-likelihood measure (LL) 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. Previous work has used this method for cross language phone mapping, i.e., to map a phone (or an HMM state) in one language to a phone (or an HMM state) in another language. In our work, the confusion matrix has been used as a basis for cross language phone clustering (instead of simply mapping). These two different approaches of phone clustering are investigated and compared in this paper.

(1) Log-Likelihood measurement

For log-likelihood (LL) approach [4], a similarity between two phone models has to be defined. The distance between two phone models λ_i and λ_j is:

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

Where $\vec{X}_i$ denotes a sequence of observations labeled as phone i . $f(\vec{X}_i | \lambda_j)$ 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

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

When the measure between phones is determined, the clustering algorithm can be described as follows:

1. Initialization: Assign each language-dependent phone model to a class.

2. Loop: Calculate the distance based on the selected measure and cluster 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, otherwise go back to step 2.

After the clustering, the single bilingual phone set is obtained.

(2) TCM

TCM is a phone clustering approach similar to automatic phone mapping method using confusion matrix [19] 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 [18]. Figure 1 shows an example of the co-occurrence between phone

Fig. 1 Example of the co-occurrence between phone “aa.ch” and phone “ae.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)

“aa_ch” and phone “ae_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 $A_{S,T}(M, N)$ be the confusion matrix and A_{ij} be the i^{th} row and j^{th} column element of this matrix. Given the target language phoneme t_j and the source language phoneme s_i , the confusion probability can be computed as:

$$A_{ij} = \frac{\text{count}(t_j|s_i)}{\sum_{n=1}^N \text{count}(t_n|s_i)} \quad (3)$$

where

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

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 confusion matrixes ($A_{\text{man,eng}}, A_{\text{eng,man}}$). The final confusion matrix after two-pass process is calculated as:

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

After the final confusion matrix A_{TCM} is obtained, the clustering information can be derived from this matrix. If the i^{th} row and j^{th} column element of 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 among all the phones, thus the i^{th} phone and the j^{th} phone from two languages will be clustered into one class. Then remove the i^{th} row and j^{th} column from A_{TCM} , and find the entry with the largest value among the rest elements and cluster the corresponding phones. This clustering procedure continues until the desired number of phone classes is reached.

By clustering all the phones from both languages with LL or TCM criteria, a clustered bilingual phone set is obtained. The original language-dependent phone models are mapped to this bilingual phone set. The dictionary, question list for decision tree and phonetic transcriptions of training data are also processed accordingly (i.e. to rewrite with the clustered bilingual phone set). Thus bilingual acoustic models can be retrained with these new labelled files.

Table 4 presents the language tagged clustered phone information based on LL measure and our proposed work TCM[†]. Since English phone set is toneless, the tone markers of Mandarin phones were removed before clustering. Considering the phonetic inventory of the International Phonetic

Table 4 The clustered phones information.

Mandarin	English	Mandarin	English
a_ch	aa_en	a_ch	aa_en
aa_ch	aw_en	au_ch	ay_en
ak_ch	ae_en	b_ch	b_en
o_ch	ao_en	ch_ch	ch_en
at_ch	ow_en	d_ch	d_en
au_ch	ay_en	ea_ch	ey_en
b_ch	b_en	f_ch	f_en
q_ch	ch_en	g_ch	g_en
ea_ch	ey_en	i_ch	iy_en
er_ch	er_en	iii_ch	y_en
f_ch	f_en	k_ch	k_en
i_ch	iy_en	m_ch	m_en
uu_ch	l_en	n_ch	n_en
m_ch	m_en	p_ch	p_en
x_ch	s_en	s_ch	s_en
s_ch	z_en	sh_ch	sh_en
sh_ch	sh_en	u_ch	u_en
u_ch	w_en	uu_ch	ow_en
v_ch	uw_en	zh_ch	jh_en
clustered Phones by LL Method		clustered Phones by TCM Method	

Table 5 Performances of different phone clustering approaches.

Acoustic Model	PER (%)		
	TestM	TestE	TestB
Model_M	20.91	—	—
Model_E	—	45.33	—
Model_ME	25.17	47.68	16.76
Model_LL70	21.59	44.54	16.47
Model_TCM70	21.52	42.07	14.45

Association (IPA) [15] 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 5 shows the performances of acoustic models with different clustering 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 PERs on all the three corpora when compared to LL measure. Results of comparative experiment indicate that the proposed TCM algorithm outperforms LL algorithm favorably.

4.1.2 The Stopping Criterion of the Clustering Process

For phone clustering, another aspect concerning model op-

[†]The Mandarin and English speech data decoded for phone clustering are chosen from the training corpora of DB863 and WSJ respectively.

^{††}Some diphthongs and triphthongs with tone markers in 179 tonal phones will be split into corresponding sequences of toneless mono phones in IPA. After this mapping, the 179 tonal phones are replaced by the 49 toneless IPA phones.

timality and complexity that we have investigated is the terminal number of clustered phones. Appropriate number of clustered phones can improve the recognition accuracy with more economical model size. The criterion of this aspect requires looking at the distribution of the distances within clusters for a modeling unit to find what can be determined as “close enough”. Based on this criterion, we merge those units which are so common that there is no significant distinction, and keep the disparities among the obviously different candidate phones across different languages.

In our system, clustering approach with TCM is selected. Based on this measurement, we set the “close enough” stop thresholds experimentally, resulting in three different numbers of phones (50, 70 and 89) in the clustered phone sets. These numbers stand for three representative degree of clustering. Experiment shows that appropriate stopping criterion of phone clustering can improve the performances for monolingual and bilingual utterances with the same clustering approach.

Table 6 presents the experimental results of three different numbers of phone classes based on TCM clustering approach. Tags “_TCM89”, “_TCM70” and “_TCM50” refer to the three phone class number. Considering the similarity based on TCM clustering approach ranges from 0 to 1, setting the stop threshold as 1 means the direct combination of all the 89 phones without clustering while setting it as 0 means mapping all the 42 English phones into corresponding Mandarin phones since the English phone set is smaller than the Mandarin one in our study, and phone set of 70 clustered classes is a compromise of the above two. As can be seen, Model_TCM70 outperforms the other two models on both of TestM and TestE, even though the other two models have slight improvement on TestB. Especially on TestE, the Model_TCM70 yields a 2.6% absolute PER reduction compared to Model_TCM89 and almost 2.4% absolute reduction compared to Model_TCM50. These results highlight the effectiveness of our proposed work.

4.2 Adaptation for Non-native Speech

By clustering phones from Mandarin and English into 70 classes, a bilingual acoustic model (Model_TCM70) was trained with native 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. This difference can be distinctly found in our experimental results in Table 3 and Table 7. How to transform the native model set into a model set tuned for non-native speak-

ers is another important problem that has to be taken into account. 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. For this study, we investigate how speaker adaptation, such as MAP and model retraining method (by appending the non-native data to the training corpus to form the new training set), affects the recognition performance with the limited non-native (TrainA) adaptation data. Even though adaptation method such as MAP is commonly used in most non-native adaptation system, experiment shows that model retraining method is more appropriate for our application.

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. We compare their impacts on the performances of native Mandarin and accented English test corpora respectively in Table 7. Results from these experiments justified our speculation.

Table 7 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 24,000 adaptation utterances of TrainA, 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 14.36% relative increase in PER compared to Model_TCM70. Finally, the model retraining method achieved 18.61% and 10% relative reduction in PER on TestE and TestB respectively while little degradation was observed for TestM corpus.

Finally, the optimal approaches mentioned in each section are integrated to form our final MESRS. Figure 2 presents the performances of our MESRS compared to those of the baseline monolingual and bilingual systems.

Table 6 Performances of different numbers of terminate phones classes Based on TCM clustering approach.

PER (%)				
Acoustic Model	Threshold (0~1)	TestM	TestE	TestB
Model_TCM89	>1	21.69	44.67	13.87
Model_TCM70	>0.3	21.52	42.07	14.45
Model_TCM50	>0	22.45	44.48	12.72

Table 7 Recognition results of non-native adaptation using the GTC data set.

PER (%)			
Acoustic Model	TestM	TestE	TestB
Model_TCM70	21.52	42.07	14.45
Model_TCM70_MAP	24.61	29.90	10.98
Model_TCM70_Retrain	21.64	34.24	13.01

Fig. 2 Comparison of acoustic modeling performances.

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-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 re-training method outperforms MAP adaptation method in our study.

The final phrase error rates of our MESRS are reduced by 24.47% and 22.37% relatively for non-native English test set (TestE) and bilingual test set (TestB) compared to the corresponding baseline single language only system, and the performance on native Mandarin test set (TestM) is comparable to the performance of the baseline Mandarin only acoustic model (Model_M). These results show the effectiveness of these proposed methods in improving bilingual recognition performance. Although the task is specific for music retrieval, we believe the research findings presented in this paper can be applicable to other multi-lingual recognition tasks as well.

Acknowledgement

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).

References

- [1] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*, Clarendon Press, Oxford, 1993. (1997 edition with a new Afterword.)

- [2] Z. Wang, U. Topkara, T. Schultz, and A. Waibel, "Towards universal speech recognition," *Proc. ICMI 2002*, 2002.
- [3] S. Martinčič-Ipšić, J. Žibert, I. Ipsic, F. Mihelič, and N. Pavešić, "Bilingual speech recognition for a weather information retrieval dialog system," *EUROCON 2003*, Ljubljana, Slovenia, The IEEE Region 8, 2003.
- [4] S. Yu, S. Zhang, and 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.
- [5] H. Ye and S. Young, "Improving the speech recognition performance of beginners in spoken conversational interaction for language learning," *Interspeech 2005*, Lisbon, Portugal, 2005.
- [6] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," *Proc. ICASSP 2000*, 2000.
- [7] I. Amdal, F. Korkmazskiy, and A.C. Surendran, "Joint pronunciation modeling of non-native speakers using datadriven methods," *Proc. ICSLP00*, Beijing, China, 2000.
- [8] J.Y.C. Chan, P.C. Ching, and T. Lee, "Development of Cantonese-English code-mixing speech corpus," *Proc. 9th European Conference on Speech Communication and Technology*, pp.1533–1536, Lisboa, Portugal, Sept. 2005.
- [9] 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, Sept. 2006.
- [10] J.B. Mariño, J. Padrell, A. Moreno, and C. Nadeu, "Monolingual and bilingual Spanish-Catalan speech recognizers developed from SpeechDat databases," *Proc. International Workshop on Very Large Telephone Speech Databases*, pp.57–61, Athens, May 2000.
- [11] L.-M. Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," *Proc. Multilinguality in Spoken Language Processing*, 2001.
- [12] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," *Proc. ICASSP 2003*, 2003.
- [13] The CMU Pronouncing Dictionary v0.6, The Carnegie Mellon University, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- [14] O.-S. Bohn and J.E. Flege, "The production of new and similar vowels by adult German learners of English," *Stud. Second Lang. Acquis.*, vol.14, pp.131–158, 1992.
- [15] IPA, (1993). *The International Phonetic Association* (revised to 1993) IPA Chart. *Journal of the International Phonetic Association* 23, 1993.
- [16] J. Köhler, "Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks," *MIST*, Leusden, The Netherlands, Sept. 1999.
- [17] Y. Yan, E. Barnard, and R.A. Cole, "Development of an approach to automatic language identification based on phone recognition," *Comput. Speech Lang.*, vol.10, no.1, pp.37–54, Jan. 1996.
- [18] W. Byrne, P. Beyerlein, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," *ASRU'99*, Keystone, CO, USA, Dec. 1999.
- [19] V.B. Le and 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.
- [20] R. Bayeh, S. Lin, G. Chollet, and C. Mokbel, "Towards multilingual speech recognition. using data driven source/target acoustical units association," *ICASSP'04*, vol.I, pp.521–524, Montreal, Canada, May 2004.

Qingqing Zhang graduated from Beijing University of Posts and Telecommunications (BUPT) in July 2005 with Bachelor's degree from School of Telecommunication Engineering. Currently she is a Ph.D. candidate of ThinkIT Speech Laboratory, Institute of Acoustics (IOA), Chinese Academy of Science (CAS). Her research interests include speech signal processing and speech recognition.

Jieli Pan received his B.S. from Peking University in 1986 and his Master degree in Electronic Engineering from Tsinghua University in 1989. From Jan. 2000 to July 2001 he was working in Intel China Research Center as a senior researcher and speech recognition group manager. In Dec. 2002 he joined ThinkIT laboratory as a professor. His current research includes: LVCSR, speech analysis, acoustic model and search algorithm.

Yang Lin received her B.S. degree and M.S. degree from Institute of Information Science and Technology, Beijing University of Chemical Technology in 2002 and 2005 respectively. She is currently a Ph.D. candidate in ThinkIT Speech Lab, IOA, CAS. Her research interests are speech recognition, speech signal processing, speech and auditory perception.

Jian Shao received the B.S. degree in department of Electrical Science and Engineering from Nanjing University in 2003. He now is a Ph.D. candidate of ThinkIT Speech Lab, IOA, CAS. His research is focused on spontaneous speech recognition and spoken document retrieval.

Yonghong Yan received his B.E. from Tsinghua University in 1990 and his Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Now he is a professor and director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition and human computer interface. He has published more than 100 papers and holds 40 patents.