JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 26, 1491-1507 (2010)

Development of a Mandarin-English Bilingual Speech Recognition System with Unified Acoustic Models*

QING-QING ZHANG, JIE-LIN PAN AND YONG-HONG YAN
ThinkIT Speech Laboratory
Institute of Acoustics
Chinese Academy of Sciences
Beijing, 100190 China

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 for real-world applications are tackled: 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. A unified bilingual acoustic model, which is derived by the novel Two-pass phone-clustering method based on the Confusion Matrix (TCM), is developed to solve the first problem. To deal with the second problem, several nonnative model modification approaches are investigated on the unified acoustic models. Compared to the existing log-likelihood phone-clustering method, the proposed TCM method with effective incorporation of limited amounts of nonnative adaptation data and adaptive modification, relatively reduces the Phrase Error Rate (PER) by 10.9% for nonnative English phrases and the PER on Mandarin phrases decreases favorably, and besides, the recognition rate for bilingual code-mixing phrases achieves an 8.9% relative PER reduction.

Keywords: bilingual speech recognition, two-pass phone clustering, confusion matrix, nonnative adaptation, model retraining

1. INTRODUCTION

With the globalization in modern society, bilingual or multilingual communication has become a common phenomenon. This presents a new challenge for 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 phrases spoken by corresponding native speakers. Although these bilingual systems achieved respectable performances for native monolingual speakers, their performances could degrade badly for nonnative utterances. For some applications, a bilingual system has to face the fact that many users who take their native language as the matrix language are not native to the embedded language[1]. Therefore, improving performance on the nonna-

Received September 26, 2008; revised June 2 & July 16 & September 8, 2009; accepted January 5, 2010.
Communicated by Suh-Yin Lee.
* This work was partially supported by the National Science and Technology Pillar Program (2008BAI50B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014), and has been presented in the ICASSP (International Conference on Acoustics, Speech, and Signal Processing), March 30-April 4, 2008, Las Vegas, Nevada, U.S.A.
[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; the other languages are considered as the embedded languages, according to Myers-Scotton's Matrix Language Frame model [1].

tive embedded language is needed before these systems can be put into practical use [5]. Studies on nonnative speech recognition have focused on two basic approaches, one based on pronunciation modeling, and the other based on acoustic modeling.

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

For the acoustic-modeling approach [8-10] 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, obtaining sufficient nonnative training data is very difficult. When training data is limited, the performances of these nonnative acoustic models can degrade significantly due to insufficient training.

Thus, how to improve system performance with only a limited amount of data becomes an important issue. Speaker-adaptation techniques such as Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) have been used to adapt acoustic models trained with native speech to handle nonnative utterances. [11] studied adaptation methods for nonnative speech and found that substantial gains could be obtained. [12] compared the effectiveness of several adaptation techniques on nonnative 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 becoming more closely connected to the world. Foreign languages (particularly English) are being used more frequently, especially among younger generations in cities. It is very common for English to be embedded in Chinese sentences during conversations. 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 real-world application. Our goal is to develop a system that can yield decent performance on nonnative English with only limited amounts of nonnative English data while maintaining the highest possible recognition rate in Mandarin. Although the task is specific, the methodologies presented in this paper can be applied to general recognition tasks as well.

The task domain of our MESRS is music retrieval[2]. The system enables a user to find a song by simply saying the singer's name or the title of the song through a phone call. Because many songs are English songs, the incoming calls may consist of songs and singers' names uttered either monolingually (Mandarin or English) or bilingually. Therefore, the system has to deal with input phrases with code switching or code mixing between two languages. The following are typical examples of these phrases: 在那遥远的地方, Backstreet Boys (code switching), "One Night in 北京, 阿 ben" (code mixing).

To process language switching and reduce computing resource requirements for practical reasons, only one set of Mandarin-English bilingual acoustic models is devel-

[2] This application has been used in Color Ring Back Tone services of China Mobile Co.

oped in our study. By merging and clustering the phone sets of these two languages, a new set of mono-phones covering both languages is determined. Different sizes of Mandarin-English bilingual phone sets are experimented with and compared. To deal with the accent issue with limited amount of nonnative training data, different types of Phone Clustering Information (PCI) and nonnative adaptation methods are investigated based on clustered acoustic models. Comparative studies were conducted, and only the most effective methods were selected. Experiment results show that encouraging advances are made compared to the baseline system.

This paper is structured as follows: The database is presented in section 2. In section 3, we describe the baseline system of our experiments. In sections 4 and 5, we document our efforts in improving the recognition performance via bilingual acoustic modeling and nonnative adaptation. Section 6 gives a brief conclusion of this paper.

2. BILINGUAL CORPUS

This section briefly describes the data resources and feature analysis that were used for all the experiments presented in this paper. All the speech data were recorded through telephone lines and digitized at an 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 are analyzed at a 10msec frame rate with a 25msec window size. Online Cepstral Mean Subtraction (CMS) is employed.

2.1 Training Corpora

Our training corpora are divided into three categories: the native Mandarin corpus, the native English corpus, and the Mandarin accented English corpus. The native Mandarin training corpus consists of the native Chinese speech corpus of National 863 Hi-Tech Project (DB863) [16]. It is a standard corpus published by governmental research program 863 for read speech in Mandarin. The English training corpus is WSJ, *etc*. These corpora were recorded in higher sampling rate and were band limited to 4 KHz by down sampling for our experiments. The Mandarin Accented English corpus was collected in house. It includes 24 hours of spontaneous English speech data from everyday conversations and 50 hours of English read speech data with texts from news Web pages [23], which are all spoken by native Mandarin speakers. Table 1 summarizes the main information about these three corpora.

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	Lab	74

2.2 Testing Corpora

The test data contains phrases of names of singers and songs spoken by hundreds of Mandarin residents, which were collected from eleven different provinces in China. There are 10,179 phrases in total, which consist of 8,183 mono-Mandarin phrases, 1,650 mono-English phrases, and 346 bilingual code-mixing phrases, respectively. The examples of these three types of test phrases can be found in Table 2. The grammar used to test these three types of sets is uniform, contains all of the phrases in these three types of sets, and has about 6,000 different items in total. In the paper, each phrase is considered as an item to recognize, and the Phrase Error Rate (PER) is used to refer to the percentage of incorrect recognitions for phrases in our systems.

Table 2. Summary of three test corpora.

Test Corpus	Language	No. of utterances	Example (song; singer)
TestM	monolingual	8183	花样年华; 张学友
TestE	monolingual	1650	Hey Jude; The Beatles
TestB	code-mixing	346	Hello 朋友; Newz 乐队

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

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 of the acoustic models used throughout our paper are state-clustered, crossword tri-phone HMMs with 32-component Gaussian mixture output densities per state. Model_M comprises 5,886 states, and Model_E comprises 5,829 states. The English phone set is supplied by the ARPABET, and the dictionary is based on the CMU pronunciation dictionary [13]. This dictionary consists of approximately 53,000 words with associated phonetic transcriptions, and all the words are selected from general conversations. 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 consists of 25,000 isolated Chinese characters. The English dictionary and Mandarin dictionary are used for both training and decoding.

Table 3 shows the performances of the baseline acoustic models. This yields the PER of 20.91% on the Mono-Mandarin corpus (TestM) and 45.33% on Mono-English corpus (TestE). It is noticeable that the performance on the Mono-English test corpus (TestE) is much worse than that of the Mono-Mandarin one (TestM). This is because in

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

	PER (%)		
Acoustic Model	Test M	Test E	Test B
Model_M	20.91	--	--
Model_E	--	45.33	--
Model_ME	25.17	47.68	16.76

our test, the test English phrases contain strong Mandarin accents, and the Mono-English acoustic model (Model_E) was trained by the native English corpus. There is a huge mismatch between the pronunciations in the dictionary and the actual acoustic realizations in those Mandarin-accented test phrases. As these two separate monolingual models cannot recognize the bilingual code-mixing phrases directly, only the results on their corresponding languages are presented.

3.2 Baseline Bilingual System

For comparison, a baseline bilingual acoustic model was trained on TrainM and TrainE together, and this model's 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 the "Model_ME" bilingual acoustic model. The performance is also compared with that of the baseline monolingual acoustic models in Table 3.

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. Even though Model_ME can deal with the code-mixing phrases (TestB), the performances on Mandarin and English phrases decrease drastically when compared to baseline monolingual models. On the other hand, because of the direct combination of Mandarin and English phone sets, 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 because the English test phrases are Mandarin accented, while the Mandarin ones are native. How to deal with these problems was the original motivation of the work presented in this paper.

4. UNIFIED BILINGUAL ACOUSTIC MODELS

For bilingual speech recognition, especially for code mixing, it is very important to determine a global phone set for different languages involved in the system. For code switching, instead of using one recognizer, a system can first use language identification technology [18] to determine which language is being used, and then it uses a recognizer for that language to conduct the recognition. For code mixing, using language identification first is computationally not feasible for real-world applications. Adopting a global phone set can also potentially reduce the amount of data required to robustly estimate statistical models. In our case, English words involved are Mandarin accented, which have nonnative pronunciation variations. [14] argue that nonnative speakers may produce speech sounds that are either part of their own native language or that are established via

merging characteristics of a native sound with a nonnative speech sound. Based on this, one can speculate that a suitable phone set resulting from merging and clustering phones in these two languages may efficiently handle the Mandarin accents in English words because 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 testing set was selected for our final system.

4.1 Phone-Clustering Algorithms

Recent approaches to phonetic clustering can be roughly divided into two categories: knowledge-based approaches and data-driven approaches. Because knowledge-based approaches do not consider the statistical similarities between phones, but the data-driven ones do [2, 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 [17, 19, 22]. 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 a Confusion Matrix (TCM) is proposed. Previous work has used this method for cross-language phone mapping, that is, to map a phone (or an HMM state) in one language to a phone (or an HMM state) in another language. In our application, both Mandarin and English words are possible to present in the decoding utterances. When clustering phones, we should consider how the phone clustering approach affects the recognition accuracies of Mandarin utterances and English utterances respectively. Instead of simply mapping, the confusion matrix has been used as a basis for cross-language phone clustering in our work. These two different approaches of phone clustering are investigated and compared as follows.

4.1.1 Log-likelihood measurement

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

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

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

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

Similarities between phones are iteratively calculated based on this measure, and phones with the maximum similarities are merged until the desired number of phone

classes is reached. Thus, the single bilingual phone set is obtained.

4.1.2 TCM

In the bilingual speech recognition, a suitable phone clustering approach should capture the phone similarities of different languages in the decoding results, and tries to improve the speech recognition performances in these different languages. In our work, a phone clustering approach named TCM is proposed, which is similar to the automatic phone-mapping method using confusion matrix [20, 21]. To construct a conversion method of the confusion matrix into a symmetric similarity matrix, the Houtgast algorithm is used in [21]. In our research, a bilingual phone-clustering algorithm based on a two-pass approach called TCM is proposed to get the symmetric similarity matrix. In each pass, Mandarin and English take turns as the source language and the target language to calculate the corresponding confusion matrixes. These two confusion matrixes capture the phone similarities based on the utterances of Mandarin and English in the decoding respectively. The detailed algorithm is described as follows:

1. Target reference: Force-align target language speech utterances using the target-language acoustic model to get the time-label information. The resulting time-aligned phone strings are considered the target phone references (*e.g.*, Model_M was used for forced alignment on Mandarin utterances).
2. Source hypothesis: The source language phone recognizer is applied to these utterances to obtain the phonetic transcriptions. During the recognizing process, no language model is used. 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 (*e.g.*, Model_E was used in the phone recognizer on Mandarin speech data).
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 [19]. Fig. 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.)

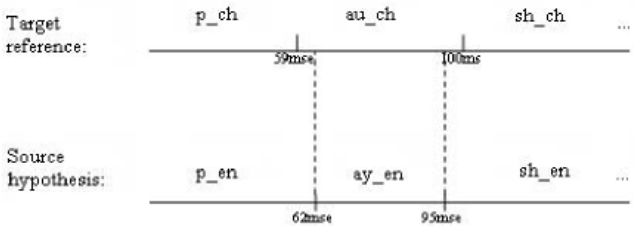

Fig. 1. Example of the co-occurrence between phone "au_ch" and phone "ay_en" when Mandarin is taken as the target language.

4. Calculation of confusion probability: Let M, N be the numbers of phones in source and target language, respectively. Let $A_{ST}(M, N)$ be the confusion matrix and $A_{i,j}$ be the ith row and jth column element of this matrix. Given the target language phone t_j and the source language phone S_i, the confusion probability can be computed as

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

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

5. The final confusion matrix: A confusion matrix is obtained given that the source language (Mandarin or English) has been introduced already. We switch 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, and then we go back to step 1 to calculate the second confusion matrix. After this two-pass process, we have two matrixes: ($A_{man,eng}$, $A_{eng,man}$). The overall confusion probability matrix after two-pass process is calculated as

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

After the final confusion matrix A_{TCM} is obtained, the clustering information can be derived from this matrix. If the ith row and jth column element of A_{TCM} has the largest value among all the elements, it means that the ith phone and the jth phone from corresponding languages have the maximum similarity, thus the ith phone and the jth phone from two languages will be clustered into one class. Then the ith row and jth column are removed from A_{TCM}, the entry with the largest value among the rest elements will be found, and the corresponding phones will be clustered if needed. 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-labeled files.

Table 4 presents the language tagged clustered phone information based on LL measure and our proposed work TCM[3]. Since English phone set is toneless, the tone markers of Mandarin phones were removed before clustering. As the phonetic inventory of the International Phonetic Association (IPA) [15] is toneless, the Mandarin 179 tonal phones are split and mapped into this inventory[4]. 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.

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

[4] Some diphthongs and tripthongs 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.

Table 4. Clustered phone-pairs with different clustering algorithms.

Mandarin	English
a_ch	aa_en
aa_ch	aw_en
ak_ch	ae_en
o_ch	ao_en
at_ch	ow_en
au_ch	ay_en
b_ch	b_en
q_ch	ch_en
ea_ch	ey_en
er_ch	er_en
f_ch	f_en
i_ch	iy_en
uu_ch	l_en
m_ch	m_en
x_ch	s_en
s_ch	z_en
sh_ch	sh_en
u_ch	w_en
v_ch	uw_en

Clustered phones by LL

Mandarin	English
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

Clustered phones by TCM

Table 5. Performances of different phone clustering approaches.

	PER (%)		
Acoustic Model	Test M	Test E	Test B
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

Table 5 shows the performances of acoustic models with LL and TCM approaches. All the clustering acoustic models in this section are trained based on TrainM and TrainE, and their model sizes are almost the same as the baseline monolingual acoustic models. 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). Furthermore, phone clustering by TCM reaches even lower PERs on all the three corpora when compared to LL measure. Results of comparative experiments indicate that the proposed TCM out-performs LL algorithm favorably.

The major reason of the performance improvement using TCM comes from its PCI generated from the decoding. In TCM, the PCI provides the direct feedback from the decoding results. This information effectively captures the phone similarities in the

decoding process. At the same time, the two-pass process integrates the decoding results from Mandarin and English utterances, which makes the final PCI balanced between the two languages. From these reasons, the phone clustering algorithm based on TCM performs well and it is selected in our following experiments.

4.2 The Stopping Criterion of the Clustering

For phone clustering, another aspect concerning model optimality 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".

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 set. These numbers stand for three representative degrees of clustering. Experiments show that appropriate stopping criterion of phone clustering can improve the performances for monolingual and bilingual phrases with the same clustering approach.

Table 6 presents the experimental results of three different numbers of terminal phone classes based on TCM clustering approach. Tags TCM89, TCM70, and TCM50 refer to the three phone-class numbers. Considering that 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[5], while setting it as 0 means mapping all the 42 English phones into corresponding Mandarin phones because the English phone set is smaller than the Mandarin one in our study; a 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 TestM and TestE, even though the other two models have a slight improvement over 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.

Table 6. Performances of different numbers of terminal phones classes based on the TCM clustering approach.

		PER (%)		
Acoustic Model	Threshold (0~1)	Test M	Test E	Test B
Model_TCM 89	> 1	21.69	44.67	13.87
Model_TCM 70	> 0.3	21.52	42.07	14.45
Model_TCM 50	> 0	22.45	44.48	12.72

5. NONNATIVE ADAPTIVE MODIFICATION

By clustering phones from Mandarin and English into 70 classes, a unified bilingual acoustic model (Model_TCM70) was trained with native speech from each language. The nonnative speakers' pronunciations, however, different from those native speakers' pronunciations observed during system training drastically decrease the recognition per-

formance. This difference can be distinctly found when comparing the performances of TestM and TestE in Tables 3 and 5. How to transform the unified native model set into a model set tuned for nonnative speakers is another important problem that has to be taken into account. With the native bilingual acoustic model in hand, the challenge for nonnative speech recognition is to maximize the recognition performance based on bilingual acoustic models with the available small amount of nonnative data. In the following sections, nonnative adaptation methods based on Model_TCM70 are investigated to improve the performance for nonnative English test corpus, while keeping the performance on native Mandarin test corpus comparable to that of the baseline Mandarin-only acoustic model (Model_M).

5.1 Nonnative Information of the PCI

For our application, compared with native English acoustic models, nonnative ones do have a more significant role because most users are Mandarin accented. Thus, Phone Clustering Information (PCI) based on nonnative English and native Mandarin will be more accordant to actual needs. Model_TCM70 is an acoustic model trained based on the PCI of native English and native Mandarin, which is not suitable enough for the application. Thus, the PCI based on nonnative English and native Mandarin needs to be developed.

TrainA is the only nonnative English training corpus, which only has limited amounts of data compared to the native English training corpus TrainE. Because TrainA lacks the sufficient tri-phone coverage as TrainE does, it cannot be used to train the bilingual acoustic models directly. However, this corpus has the ability to capture the main characteristics of nonnative pronunciation variations. More precise nonnative in-formation for phone clustering can be obtained from these nonnative training data. A small nonnative English acoustic model was trained using only TrainA, which is used to provide the nonnative information during phone clustering. We call this model "Model_ NE," which means nonnative English acoustic model.

Table 7 presents the language-tagged clustered phones based on different PCIs from (Model_M, Model_NE) and (Model_M, Model_E). The left two columns of phones refer to the pairs of clustered phones from (Model_M, Model_NE), which means that using Model_M and Model_NE for forced alignment/phone recognition, and using TrainA/ TrainM data for computing the numbers of co-occurrences. The right two columns refer to the ones from (Model_M, Model_E), which means that using Model_M and Model_E for forced alignment/phone recognition and using TrainE/TrainM data for computing the numbers of co-occurrences. Even though both of these use the same phone-clustering method (TCM), the clustered phone sets are not the same. For example, "au_ch" in Mandarin is close to the native English phone "ay_en," while it is more similar to the nonnative English one "ae_en." These result from the variations between native and nonnative pronunciations. Based on the clustered information of (Model_M, Model_NE), new bilingual acoustic models (Model_NE_TCM70) were trained using the same native training data as those of Model_TCM70 from TrainM and TrainE. Table 8 shows the performances of these two different clustered acoustic models. The perfor-mances of Model_TCM70 are better than those of Model_NE TCM70 on TestM and TestE, even though there is a little drop on TestB. On most of the testing sets, Model_TCM70 outperforms Model_NE_TCM70.

Table 7. Clustered phone-pair lists based on different PCI.

Mandarin	English
au_ch	ae_en
aa_ch	aw_en
b_ch	b_en
d_ch	d_en
ea_ch	ey_en
f_ch	f_en
g_ch	g_en
h_ch	hh_en
k_ch	k_en
l_ch	l_en
m_ch	m_en
nn_ch	n_en
ng_ch	ng_en
uu_ch	ow_en
p_ch	p_en
s_ch	s_en
sh_ch	sh_en
u_ch	w_en
i_ch	y_en

Clustered phones (Model_M, Model_NE)

Mandarin	English
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

Clustered phones (Model_M, Model_E)

Table 8. Performances of clustered acoustic models based on (Model_M, Model_E) and (Model_M, Model_NE) separately.

	PER (%)		
Acoustic Model	Test M	Test E	Test B
Model_TCM70	21.52	42.07	14.45
Model_NE_TCM70	21.62	43.04	12.72

5.2 Nonnative Modification Based on Clustered Models

So far the unified bilingual acoustic models were trained with data from the corresponding native corpora, so the mismatch between the native English training corpus and the nonnative English test corpus still exists. The conclusions in [11, 12] proved that with a certain amount of nonnative speech data in training process, the original native acoustic models could be more attuned to nonnative pronunciations. According to this, we then investigated how the limited amount of nonnative training data TrainA affects the recognition performances based on different PCI for the three test corpora. The results show that when the type of PCI is consistent with that of training data, which covers the characteristics of nonnative English's and native Mandarin's pronunciations, the large gap between the performances for TestM and TestE can be reduced significantly.

5.2.1 Nonnative adaptation

How to add these nonnative speech data into the training set is the first problem we

should deal with. We compared speaker adaptations such as MAP and the model retraining method (by appending the nonnative data to the training corpus to form the new training set) in our paper.

In MAP adaptation, the native model parameters are re-estimated individually, using held-out nonnative adaptation data. An updated mean is then formed by shifting the original native value toward the nonnative sample value. If there is insufficient adaptation data to reliably estimate the sample mean of a phone, no adaptation is performed.

The model retraining method is a compromise settlement. Because the bilingual acoustic models consist of clustered phones, which are language-independent, sharpening the acoustic models on nonnative training data may move further away from the native speakers. Therefore, compared to MAP, appending the pool of nonnative 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.

Table 9. Recognition results of nonnative adaptation with TrainA.

	PER (%)		
Acoustic Model	Test M	Test E	Test B
Model_TCM70_ReT	21.56	30.02	11.85
Model_NE_TCM70	21.62	43.04	12.72
Model_NE_TCM70_MAP	25.09	26.64	9.83
Model_NE_TCM70_ReT (MESRS)	21.65	29.72	8.96
Model_NE_LL70_ReT	22.07	33.34	9.83

Table 9 presents the recognition results of different adaptation methods. The clustered baseline bilingual model used for comparison is Model_NE_TCM70, whose PCI is obtained from Model_M and Model_NE. With adaptation utterances of TrainA, MAP transforms were conducted. Model_NE_TCM70_MAP refers to the corresponding adapted acoustic model. The results show that with a pool of nonnative adaptation data collected in advance, MAP can substantially improve the performance on new nonnative speakers (TestE). However, the performance on the native Mandarin test set (TestM) degrades seriously, which gives a 16.05% relative increase in PER compared to Model_NE_TCM70. Finally, the acoustic models with model retraining method (Model_NE_TCM70_ReT) achieved a 30.95% and a 29.56% relative reduction in PER on TestE and TestB, respectively, while little degradation was observed for TestM corpus when compared with Model_NE_TCM70. Thus, the model-retraining method is more suitable for our application, and is selected for our final system.

5.2.2 Consistency between PCI and training data

As shown above, appending the nonnative training data in the training process does achieve great improvement on the recognition of nonnative test data. It reduces the large performance gap between TestM and TestE. Considering that there are two different kinds of PCI, which stands for different clustering information between native Mandarin and native or nonnative English, nonnative adaptive methods should be applied based on these separately to discover the interaction between the influence of PCI and nonnative adaptation. The comparative experimental results are presented in Table 9. Model_

TCM70_ReT refers to the final clustered acoustic model based on the PCI of (Model_M, Model_E), and Model_NE_TCM70_ReT refers to the corresponding acoustic model based on the PCI of (Model_M, Model_NE). Both of the two acoustic models were retrained by adding the same nonnative training data (TrainA). As shown, even though the PER of Model_NE_TCM70_ReT on Mandarin phrases (TestM) has a tiny increase (0.42% relative) compared to that of Model_TCM70_ReT, the PER for English phrases (TestE) was reduced by 1.0%, relatively, compared to Model_TCM70_ReT, and the performance for bilingual code-mixing phrases (TestB) achieved 2.44% relative PER reduction. Model_NE_TCM70_ReT gives better performances on the whole.

This is a completely opposite conclusion to that in section 5.1. In section 5.1, when the training data for clustered bilingual acoustic models are all from the corresponding native training corpora, the PCI from (Model_M, Model_E) outperforms. When the nonnative training data are appended, however, the PCI from (Model_M, Model_NE) gives a more gratifying result. These results illustrate that when the type of PCI is consistent with that of the clustered training data for bilingual acoustic models, the unified bilingual acoustic models will have the preferable performances. In our application, the English parts in monolingual and code-mixing corpora are Mandarin accented, and so Model_NE_TCM70_ReT, which was retrained with the nonnative training data based on the PCI of nonnative English and native Mandarin, achieves the best performances.

Finally, instead of TCM, the LL approach and all the optimal approaches mentioned in these sections above are integrated to get the acoustic model Model_NE_LL70_ReT[6]. This model is compared with Model_NE_TCM70_ReT to find out which approach for phone clustering performs better in the final bilingual speech recognition system. As shown in Table 9, the Model_NE_TCM70_ReT outperforms Model_NE_LL70_ReT on all of the three testing sets. From the results we can see that the TCM approach has the capability to capture the similarities of phones between different languages to build a better bilingual phone set, and in the research the Model_NE_TCM70_ReT is selected to form our final MESRS.

6. CONCLUSIONS

This paper presents our recent work in developing a Mandarin-English bilingual speech recognition prototype system via the unified bilingual acoustic models for real world application. In addition to the requirement of handling inter- and intra-sentential language switching at the same time, the challenge also includes the fact that only a limited amount of out-of-task-domain accented English data is available. A novel method named TCM is proposed in the paper. Experiment results show that the proposed TCM outperforms existing LL approach favorably. It is also shown that, with limited availability of nonnative adaptation data, the model retraining method outperforms the MAP adaptation method based on a unified set of bilingual acoustic models, and when the type of phone clustering information is consistent with that of the training data for bilingual acoustic models, which covers the characteristics of nonnative English's and native Mandarin's pronunciations, the performances can be improved further. Besides, the results of comparative experiments in the initial and final bilingual speech recognition sys-

[6] The phone set of Model_NE_LL70_ReT is rebuilt with LL approach, using Model_M and Model_NE for forced alignment/phone recognition, and using TrainA/TrainM data for computing the numbers of co-occurrences. For the consistency of the model size, the number of phone classes is selected to be 70, which is the same as the number in Model_NE_TCM70_ReT.

tems indicate that the proposed TCM outperforms LL algorithm consistently. These results show the effectiveness of these proposed methods in improving bilingual recognition performance, given the limited amount of nonnative adaptation data. Although the task is domain specific for music retrieval, we believe the research findings presented in this paper can be applicable to other multilingual recognition tasks as well.

REFERENCES

1. C. Myers-Scotton, *Duelling Languages*: *Grammatical Structure in Code-Switching*, Clarendon Press, Oxford, Vol. 6, 2007.
2. Z. Wang, U. Topkara, T. Schultz, and A. Waibel, "Towards universal speech recognition," in *Proceedings of International Conference on Multimodal Interfaces*, 2002, pp. 14-16.
3. S. Martinčic-Ipšic, J. Žibert, I. Ipsic, F. Mihelič, and N. Pavešic, "Bilingual speech recognition for a weather information retrieval dialog system," *The IEEE Region 8*, Vol. 2807, 2003, pp. 380-387.
4. S. Yu, S. Zhang, and B. Xu, "Chinese-English bilingual phone modeling for cross-language speech recognition," in *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, Vol. 1, 2003, pp. 603-609.
5. H. Ye and S. Young, "Improving the speech recognition performance of beginners in spoken conversational interaction for language learning," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 289-292.
6. K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Vol. 3, 2000, pp. 1683-1686.
7. I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Joint pronunciation modeling of non-native speakers using datadriven methods," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Vol. 3, 2000, pp. 622-625.
8. J. Y. C. Chan, P. C. Ching, and T. Lee, "Development of Cantonese-English code-mixing speech corpus," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 1533-1536.
9. Y. C. Chan, P. C. Ching, T. Lee, and H. Cao "Automatic speech recognition of Cantonese-English code-mixing utterances," in *Proceedings of the 9th International Conference on Spoken Language Processing*, 2006, pp. 113-116.
10. J. B. Mariño, J. Padrell, A. Moreno, and C. Nadeu "Monolingual and bilingual Spanish-Catalan speech recognizers developed from speechdat databases," in *Proceedings of International Workshop on Very Large Telephone Speech Databases*, 2000, pp. 57-61.
11. L. M. Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," in *Proceedings of Multilinguality in Spoken Language Processing*, 2001.
12. Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, 2003, pp. 540-543.
13. The CMU Pronuncing Dictionary v0.6, The Carnegia Mellon University, 2009, 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," *Studies in Second Language Acquisition*, Vol. 14, 1992, pp. 131-158.
15. IPA, The International Phonetic Association, "IPA chart," *Journal of the International Phonetic Association*, Vol. 23, 1993.
16. A. Li, Z. Yin, T. Wang, Q. Fang, and F. Hu, "RASC863-A Chinese speech corpus with four regional accents," in *Proceedings of the International Conference on Speech and Language Technology*, 2004.
17. J. Köhler, "Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks," in *Proceedings of ESCA-NATO Tutorial and Research Workshop*: *Multi-Lingual Interoperability in Speech Technology* 1999, pp. 79-84.
18. Y. Yan, E. Barnard, and R. A. Cole, "Development of an approach to automatic language identification based on phone recognition," *Computer*, *Speech Language*, Vol. 10, 1996, pp. 37-54.
19. P. Beyerlein, *et al*., "Towards language independent acoustic modeling," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 1999, pp. 1029-1032.
20. C. L. Huang and C. H. Wu, "Generation of phonetic units for mixed-language speech recognition based on acoustic and contextual analysis," *IEEE Transactions on Computers*, Vol. 56, 2007, pp. 1225-1233.
21. P. Y. Shih, J. F. Wang, H. P. Lee, H. J. Kai, H. T. Kao, and Y. N. Lin, "Acoustic and phoneme modeling based on confusion matrix for ubiquitous mixed-language speech recognition," in *Proceedings of IEEE International Conference on Sensor Networks*, *Ubiquitous and Trustworthy Computing*, 2008, pp. 500-506.
22. R. Bayeh, *et al*., "Towards multilingual speech recognition using data driven source/target acoustical units association," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Vol. 1, 2004, pp. 521-524.
23. News weekly web page, http://www.newsweekly.com.au, 2009.

Qing-Qing Zhang (張晴晴) 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 Sciences (CAS). Her research interests include speech signal processing and speech recognition.

Jie-Lin Pan (潘接林) received his B.S. from Peking University in 1986 and his Master degree in Electronic Engineering from Tsinghua University in1989. From January 2000 to July 2001 he was working in Intel China Research Center as a senior researcher and speech recognition group manager. In December 2002 he joined ThinkIT laboratory as a Professor, his current research includes: LVCSR, speech analysis, acoustic model and search algorithm.

Yong-Hong Yan (顏永紅) 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.