MULTI-PRONOUNCIATION DICTIONARY CONSTRUCTION FOR MANDARIN-ENGLISH BILINGUAL PHRASE SPEECH RECOGNITION SYSTEM

C. Wang[1], W. Shi[2], Y. X. Zou[1]*

[1]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, 518055, China
[2]Shenzhen Key Laboratory of Intelligent Media and Speech, PKU-HKUST Shenzhen-HongKong Institution, China
*Corresponding author: zouyx@pkusz.edu.cn

ABSTRACT

Generally, in multi-lingual communities, non-native speakers may produce speech sound which is either part of their own native language or established via merging characteristics of native pronunciation with non-native pronunciation. Recently, a Two-pass phone clustering based on Confusion Matrix (TCM) approach has been proposed to address the one-to-one phone mappings between Chinese syllables and English phones using standard Chinese and English data. In this paper, we extend TCM to the one-to-many phone mappings issue since there is the merging phenomenon of native and non-native pronunciation in bilingual speeches. Employing a knowledge-based phone set to TCM as supplements for phone clustering, a novel method termed as the TCM with Initialization and Updating of the Phone Set method (TCM-IUPS). As a result, the pronunciation dictionary is built via using the information learned by our proposed TCM-IUPS as well as canonical pronunciation. Experiments show that, compared with TCM, the Phrase Error Rate (PhrER) of TCM-IUPS is reduced by 5.27% in bilingual testing corpora and 26.09% in mono-English testing corpora compared with TCM, while the same performance is maintained in mono-Mandarin testing corpora.

Index Terms— Accent issue, initialization and updating of the phone set, multi-pronunciation dictionary, bilingual speech recognition

1. INTRODUCTION

Replacing Chinese phrases with English phrases has become a very common linguistic phenomenon in several multilingual countries in Asian for many years. As a result, the demand for Mandarin-English bilingual phrase Speech Recognition System (MESRS) is becoming overwhelming. However, the performance will degrade dramatically due to the accent issue caused by non-native speeches. This paper focuses on the non-native accent issue. According to [1], the accent issue was classified into sound change and phone change. Chinese who take Mandarin as mother tongue are not native to English in which sound change and phone change exist. English who are from different states speak dialect English in which sound change exists. Flege *et al* [2] argued that non-native speakers might produce speech sound which was either part of their own native language or established via merging characteristics of native pronunciation with non-native pronunciation.

Previous researches showed that accent issue has its patterns. There would be some specific phones being varied in a particular area [2]. Therefore, replacing those specific phones was not random, but belonged to a fixed phone set [3] which was found by phone clustering approaches. Recent approaches for phone clustering can be classified into two categories [4]. First, knowledge-based approaches construct multi-lingual phone set by mapping different languages phone sets into the same phone set according to expert knowledge. International Phonetic Alphabet (IPA) [5] [6], Speech Assessment Methods Phonetic Alphabet (SAMPA) [7] and Worldbet [8] are well-known phone sets defined by experts. This method can share the parameters in the acoustic models among different languages. However, the method does not take spectral characteristics into consideration. Second, data-driven approaches merge similar phone units of different languages into the same phone unit according to the spectral characteristics. Phone units with similar spectral properties are combined into one phone unit according to the likelihood or distance between two phone units. Sufficient data is needed to train reliable acoustic models. However, it is difficult to collect sufficient well-labeled non-native and dialect training data in practice.

Researchers had noticed that knowledge-based approaches and data-driven ones verified and complemented each other for building multilingual phone set [9]. So it is straightforward to combine these two kinds of methods to derive a more feasible mapping relation from English phones to Chinese initials and finals (IFs) for MESRS. In this paper, a knowledge-based approach called Initialization and Updating of the Phone Set (IUPS) is proposed. It includes initializing the phone mapping set according to expert knowledge and updating each mapping individually based on phone similarities that learned from data-driven approach as well as TCM. At last, a multi-pronunciation

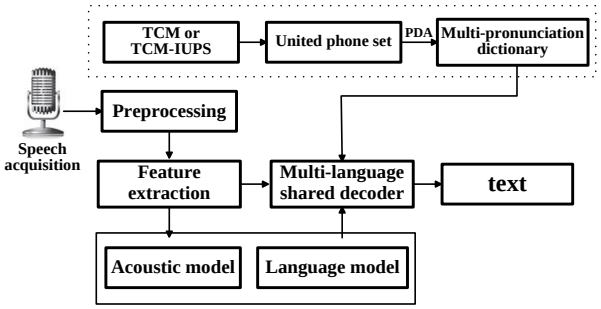

Fig.1. Block diagram for MESRS

dictionary is built based on the one-to-many phone mappings obtained by the TCM-IUPS procedure. Experiments show that, by utilizing this new multi-pronunciation dictionary rather than canonical dictionary, the performance of MESRS can be effectively improved.

2. METHODOLOGY

The block diagram for MESRS is shown in Figure 1. It consists of six parts: preprocessing, feature extraction, acoustic model, language model, multi-language shared decoder and multi-pronunciation dictionary. This paper focuses on constructing the multi-pronunciation dictionary.

As described above, the performance of MESRS can be affected by accent issue dramatically. In order to obtain a Mandarin-English phone set which reflects the non-native and dialect pronunciation variations, knowledge-based and data-driven approaches are combined together to cluster 65 Mandarin IFs (including 6 zero initials) and 44 English phones by TCM-IUPS. The clustering result is the basic unit of acoustic model [12].

2.1. TCM

TCM is a two-pass phone clustering approach similar to automatic phone mapping using confusion matrix that usually used in fast acoustic modeling for a new target language [4] [10]. In this paper, Mandarin and English take turns as the source language and the target language in each pass [11]. The algorithm is described in details as follows:

1) Target reference: Target language speech utterances are force-aligned by 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 acoustic model is applied to the same utterances to obtain the phonetic transcriptions, which 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) Calculate co-occurrence: 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, the phones of the target and source language are put 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 [13]. Figure 2 shows an example of the co-occurrence between initial 'o' and phone 'oh' 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 $\boldsymbol{A}_{ST}(M, N)$ be the confusion matrix and A_{ij} be the i^{th} row and j^{th} column element of this matrix. Given the target language phone t_j and the source language phone s_i, the confusion probability can be computed as:

$$A_{ij} = count(t_j \mid s_i) / \sum_{n=1}^{N} count(t_n \mid s_i) \quad (1)$$

where $A_{ij} \in \boldsymbol{A}_{ST}(M,N)$, $i = 1,2,\cdots,M$, $j = 1,2,\cdots,N$.

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 ($\boldsymbol{A}_{man,eng}$, $\boldsymbol{A}_{eng,man}$). The final confusion matrix after two-pass process is calculated as:

$$\boldsymbol{A}_{TCM} = \frac{1}{2}(\boldsymbol{A}_{man,eng} + \boldsymbol{A}^{T}_{eng,man}) \quad (2)$$

In our application, we assumed that the Mandarin and English model are of equal importance and they have the same weight of 0.5 respectively.

6) The final phone set: After the final confusion matrix $\boldsymbol{A}_{TCM}$ is obtained, the clustering information can be derived from this matrix. If the i^{th} row and j^{th} column element of $\boldsymbol{A}_{TCM}$ 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 $\boldsymbol{A}_{TCM}$.

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.

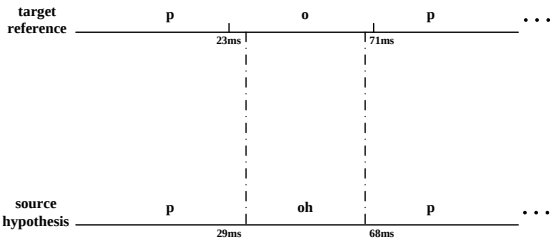

Fig.2. Example of the co-occurrence between IFs 'o' and phone 'oh' when Mandarin is the target language

2.2. The proposed TCM-IUPS

The TCM method described above can learn one-to-one phone mapping set from training data. The learned phone set may have limitations due to insufficient amount of training data. To find a more feasible Mandarin-English phone set that can contain one-to-many mapping information, in this paper, a novel phone clustering algorithm called TCM-IUPS is proposed in which TCM and IUPS are combined.

In TCM-IUPS, first, 4 Mandarin IFs are selected as Initial Mapping Set (IMS) for each English phone. Then, the phone set is updated by the result of TCM-IUPS. Finally, k ($1 \le k \le 4$) IFs are clustered for each English phone. The algorithm is described as follows:

1) Initialization of the phone set: 4 IFs are selected for each English phone based on expert knowledge, which are derived from Hoste *et al* [9], IPA, SAMPA and Worldbe. For example, 'z' is one of English phones (EPs), its IMS is {z, s, z r, s r}. For English phone j_E, its IMS $\boldsymbol{I}$ is described as follows:

$$\{i_{initial1}, i_{initial2}, i_{initial3}, i_{nitial4}\} \tag{3}$$

where $j_E \in EPs$, $\boldsymbol{I} \subseteq IFs$, and $\boldsymbol{I}$ have been sorted in accordance with the similarity descending order.

2) Calculation of confusion probability: The 1-5 steps of TCM are executed to obtain the Confusion Matrix $\boldsymbol{A}_{TCM}$, in which the rows stand for IFs, and the columns stand for EPs.

3) Updating of the phone set: A Clustering Set (CS) is defined for each English phone, which represents the clustering results for each English phone. The largest 4 elements in Confusion Matrix $\boldsymbol{A}_{TCM}$ are chosen as $\{a_{max1}, a_{max2}, a_{max3}, a_{max4}\}$, which are sorted in accordance with the similarity descending order.

For the largest value a_{max1}, suppose its row and column index are i_M ($1 \le i_M \le 65$) and j_E ($1 \le j_E \le 44$), respectively. Then, i_M is added to the CS of j_E. The CS of j_E is matched with the corresponding IMS $\boldsymbol{I}$ of j_E for the union set $\boldsymbol{i}_m$. Suppose the number of elements in $\boldsymbol{i}_m$ is n. The mapping set of j_E in the phone set is updated as follows:

$$\begin{cases} i_M, i_{initial1} & n = 0 \\ \boldsymbol{i}_m & n = 1,2,3,4,5 \end{cases} \tag{4}$$

where the element of $\boldsymbol{i}_m$ is sorted in accordance with the added time sequence of the CS of j_E. Then j_E and i_M (when $n=0$) or the first value in $\boldsymbol{i}_m$ (when $n=1, 2, 3, 4, 5$) are removed from $\boldsymbol{A}_{TCM}$.

For the other three elements a_{max2}, a_{max3} and a_{max4}, the corresponding IFs are added to the corresponding clustering sets, respectively. If the number of elements in the CS of any English phone (called j_{any}) is equal to 4, then the CS of j_{any} is matched with the corresponding IMS of j_{any} immediately. The mapping set of j_{any} in the phone set is updated the same as (4), except when n is zero. When n is zero, the clustering results of j_{any} is only the first value in the IMS of j_{any}. Then, j_{any} and the first value in the IMS of j_{any} (when n=0) or the first value in the union set of j_{any} (when n=1, 2, 3, 4, 5) are removed from $\boldsymbol{A}_{TCM}$.

The entries with the largest value among the rest elements are found and the corresponding phones will be clustered. This clustering procedure continues until the desired number of phone classes is reached. After that, the multi-pronunciation dictionary is built based on the results of the phone clustering.

3. BILINGUAL CORPORA FOR TRAINGING AND TESTING

Before introducing the experiments, it is necessary to describe the bilingual corpora in details, as it is important to the results of experiments. In this paper, all speech data are acquired at 16KHz sampling rate with 16-bit resolutions. The speech feature vector consists of 39 components (12 MFCC parameters, 1 frame energy, and their first and second delta-MFCC).

3.1. Training corpora

Our training corpora are divided into two categories: the native Mandarin corpora (labeled as TrainM), the native accented English corpora (labeled as TrainE). The National 863 Hi-Tech Project (DB863) is a standard corpus which is published by governmental research program 863 for reading speech in Mandarin. TrainM consists of 300 hours' native Chinese speech from DB863 which includes 400 male and 400 female accented Mandarin residents. While the 205 hours' TrainE is from Wall Street Journal and includes 133 male and 133 female English utterances. Table 1 shows the summary of two training corpora.

3.2. Testing corpora

This study focuses on dealing with the accent issue with limited amount of non-native training data. The testing statements are allowed to be either monolingual or bilingual. In our testing corpora, there are 20960 phrases which consist of 18674 mono-Mandarin phrases (labeled as TestM), 1261 mono-English phrases (labeled as TestE) and 1025 bilingual phrases (labeled as TestB). Mono-Mandarin phrases consist of names of singers, titles of songs, machine instructions and utterances from DB863. It includes 15 male and 18 female Mandarin residents. Bilingual phrases are intra-sentence language switching phrases including 5 male and 5 female Mandarin residents. Mono-Mandarin phrases and

Training corpora label	Language	Source	Time/h
TrainM	Mandarin	DB863	300h
TrainE	English	WSJ	205h

Tab.1. Summary of two training corpora

Training corpora lable	Language	Number	Example
TestM	Accented Mandarin	18674	今天星期几? 打开导航 心太软；周杰伦
TestE	Accented English	1261	THIS WAS EASY FOR US
TestB	Accented Bilingual	1025	打开 GPS Peter 张

Tab.2. Summary of three testing corpora

bilingual phrases were collected under realistic conditions such as in restaurants, streets, cars and other noisy places, which cover variations in background noise, microphones, volumes, speaker fluency and accents. Mono-English phrases are selected randomly from TIMIT which includes 630 speakers in American English with eight major dialects. Examples of these three types of test utterances are shown in Table 2.

4. EXPERIMENTS AND ANALYSIS

4.1 Experimental setting

MESRS adopts TCM-IUPS and TCM to cluster 65 IFS into 44 English phones for one phone set mapping. We build a multi-pronunciation dictionary based on the phone set mapping. The bilingual acoustic model is trained with TrainM by the hidden markov model toolkit (HTK), which is monophone HMMs with 16-component Gaussian mixture output densities per state. The other experiment condition of TCM-IUPS is the same as the TCM.

4.2. Experiment of the proposed TCM-IUPS and analysis

Figure 3 shows the performance comparison between TCM and TCM-IUPS phone clustering approaches. As is shown, the PhrER of TCM-IUPS is much lower than that of TCM for non-native and bilingual test data sets. The noticeable point is the great performance gap between TestM and TestE (7.02% V.S. 45.44%). This is caused by two major reasons. First, insufficient amount of clustering data results in data sparseness problem, and furthermore leads to unreliable acoustic model. Second, the testing corpora contain American English speeches with eight major dialects, and serious sound change exists in those spoken English. Another performance gap between TestM and TestB (7.02% V.S. 15.90%) is caused by that the test utterances which are from Mandarin residents, and their spoken English has serious accent issue.

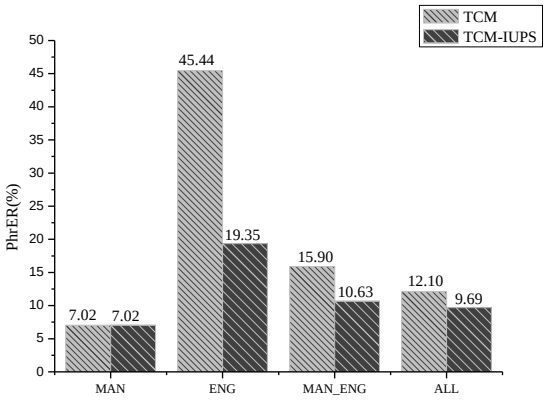

Fig.3. Performance comparison between TCM and TCM-IUPS

The PhrER of TCM-IUPS is reduced by 26.09% and 5.27% relatively for TestE and TestB compared to TCM, and the performance on TestM is comparable to that of TCM.

In conclusion, experiment results show that the multi-pronunciation dictionary produced by TCM-IUPS is much feasible than the single-pronunciation dictionary produced by TCM. By utilizing TCM-IUPS, the accent issue can be overcome effectively.

5. CONCLUSION

This paper presents a novel approach called TCM-IUPS for Mandarin-English phone clustering and multi-pronunciation dictionary construction. The original phone clustering method TCM can only learn one-to-one phone mapping, which suffers the data sparseness problem. By cooperating with the IUPS process, universal phone mapping rules defined by experts and information learned by TCM is both included in the final multi-pronunciation dictionary. Compared with TCM, the PhrER of TCM-IUPS is reduced by 5.27% in Mandarin-English bilingual testing corpora and 26.09% in American English testing corpora with eight major dialects. At the same time, the TCM-IUPS maintains the same performance in mono-Mandarin testing corpora.

6. ACKNOWLEDGEMENTS

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science Research Program (No. CXZZ20140509093608290).

7. REFERENCES

[1] L. Liu. “Small data set based acoustic modeling for dialectal Chinese speech recognition”, Journal of Tsinghua university (Natural Science Edition), vol. 48, no. 4, pp. 604-607, 2008.

[2] O. S. Bohn, J. E. Flege. “The production of new and similar vowels by adult German learners of English”, SSLA, pp. 131-158, 1992.

[3] Y. Yang, B. Ma, L. Wang, *et al*. “Multi-pronunciation dictionary based Uyghur accent modeling for speech recognition”, Journal of Tsinghua university (Natural Science Edition), vol. 51, no. 9, pp. 1303-1306, 2011.

[4] Q. Zhang, J. Pan, Y. Yan, “Development of a Mandarin-English bilingual speech recognition system”, Journal of

Chongqing university of Posts and Telecommunication (Natural Science Edition), vol. 20, no. 4, pp. 391-396, 2008.

[5] International Phonetic Association (IPA), "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet," Cambridge University. Press, Cambridge, U.K., 1999, available: https://www.internationalphoneticassociation.org/

[6] IPA, the International Phonetic Association (revised to 1993)-IPA Chat. J. Int. Phonetic Assoc., 1993, 23, available: https://www.internationalphoneticassociation.org/redirected_home

[7] J. C. Wells, "Computer-coded phonemic notation of individual languages of the European community," Journal of the International Phonetic Association, vol. 19, pp. 31–54, 1989.

[8] J. L. Hieronymus, "ASCII phonetic symbols for the world's languages: Worldbet", Journal of the International Phonetic Association, 1993, available: http://www.ling.ohio-state.edu/~edwards/WorldBet/worldbet.pdf

[9] V. Hoste, W. Daelemans, S. Gillis. "Using rule induction techniques to model pronunciation variation in Dutch". Computer Speech and Language, vol. 18, no. 1, pp. 1-23, 2004.

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

[11] S. Yu, S. Zhang, B. Xu. "Chinese-English bilingual phone modeling for cross-language speech recognition", ICASSP 2004, vol. 1, pp. 917-920.

[12] K. Reinhard, P. Jochen, "Semantic clustering for adaptive language modeling", ICASSP 1997, vol. 2, pp. 779-782.

[13] P. Beyerlein *et al*, "Towards language independent acoustic modeling", ICASSP 2000, vol. 2, pp. 1029-1032.