English Alphabet Recognition Based on Chinese Acoustic Modeling

Linquan Liu, Thomas Fang Zheng, and Wenhu Wu

Center for Speech Technology, Tsinghua National Laboratory for
Information Science and Technology, Tsinghua University, Beijing, 100084, China
liulq@cst.cs.tsinghua.edu.cn, fzheng@tsinghua.edu.cn, wuwh@tsinghua.edu.cn

Abstract. How to effectively recognize English letters spoken by Chinese people is our major concern in the paper. Some efforts are made to build Chinese extended Initial/Final (XIF) based HMMs for English alphabet recognition which can be integrated with large vocabulary continuous Chinese speech recognition (Chinese LVCSR) system based on a same XIF set. The alphabet-specific XIF HMMs are built using context-dependent modeling, decision tree based state clustering method, state-based phonetic mixture tying and pronunciation modeling techniques. Experiments have been done over a 32-speaker test set. Compared with English phoneme-based acoustic modeling, our proposed method can achieve a relative letter error rate reduction of 5.3% with a letter correctness of 97.2% for Chinese-accented English alphabet recognition. What's more, the XIF-based HMMs for English alphabet can be integrated with Chinese LVCSR seamlessly to recognize Chinese as well as English letters simultaneously.

Keywords: English alphabet recognition, Chinese speech recognition, acoustic modeling, pronunciation modeling, state-based phonetic mixture tying, state clustering.

1 Introduction

Handling non-native speech in automatic speech recognition (ASR) systems is an area of increasing interest. Most of the systems are built on native speech only and as a result the performance for non-native speakers is often not satisfactory. One effective way to deal with this problem is to adapt the acoustic models based on native speech to the non-native speaker [1]. Another important method is to cover the non-native pronunciations in the lexicon [2]. Additionally, in [3], it was shown that training on 52 minutes of non-native data (German-accented English) was much better (with a word error rate of 43.5%) than training on 34 hours of native English data from exactly the same domain (with a word error rate of 49.3%). That is to say, using a small amount of non-native speech data could also achieve good performance for non-native speech recognition. In our task, the speech data is a mix of English letters and standard Chinese words. For Chinese, non-native English speakers, their utterances of English are influenced by their mother tongue more or less. In fact, it is

Q. Huo et al. (Eds.): ISCSLP 2006, pp. 463-472, 2006

likely that Chinese speakers will pronounce an English phone as a similar phone in Chinese. Nowadays, the alphabet recognition has been applied successfully in some practical systems [4], nevertheless, how to recognize Chinese-accented English alphabet effectively and furthermore integrate it into a Chinese Initial/Final (IF) recognizer is still not deserved much attention. Besides, Some Chinese-market-oriented products, such as PlayStation® and GameBox®, do hope that the recognizer can deal with English letters as well as Chinese at the same time during the interaction with player, by doing so, voice-controlled commands can have Chinese words and English letters mixed so that much friendly interface can be provided to players. However, due to the economy consideration and resource limitation of embedded devices, it is impractical to deal with this issue by means of automatic language identification [5] and multilingual speech recognition [6]. Taking these factors into consideration, we propose to work on Chinese-accented English alphabet recognition based on Chinese IF acoustic modeling.

In this paper, we attempt to achieve comparable performance for alphabet recognition under the assumption that our methodology is much suitable for Chinese-accented speech. We attempt to achieve the goal via: 1) Chinese extended Initial/Finals (XIFs) based context-dependent modeling where XIFs are derived from the Chinese Initial/Finals [7]; 2) State clustering, which is performed to tie the acoustically similar states so as to better the baseline XIF HMMs; and 3) State-based phonetic mixture tying, whose goal is to further reduce the redundant Gaussian mixtures and to build robust XIF HMMs, phonetic mixture tying at state level is adopted which can deal well with issue of the underestimation inherent in the mixture tying method; 4) Pronunciation modeling, which is to deal specifically with Chinese-accented alphabet, pronunciation modeling is employed as another measure to improve the accuracy.

The remainder of this paper is organized as follows. In Section 2, some background information for the task is described. In the following section, state clustering and state-based phonetic mixture tying are introduced briefly, and the pronunciation modeling method for alphabet recognition is also presented. In Section 4, experiments based on Chinese XIF acoustic modeling are designed and done to verify the effectiveness for Chinese-accented alphabet recognition. The comparison between Chinese XIF-based HMMs and English phoneme-based HMMs is performed. The effectiveness for integration of English alphabet with Chinese LVCSR is also evaluated. Finally, conclusion is drawn in Section 5.

2 Background

2.1 Database

The database used in the study was of read-style Chinese, consisting of speech data spoken by 132 speakers (with gender balanced and age ranging from 11 to 49) with 236 utterances per speaker. Each speaker uttered 100 long sentences (12~18 words), 100 short phrases (4~8 words), 26 English letters, and 10 digits. To verify the effectiveness of XIF-based alphabet recognition, only the utterances containing English letters were used. That is to say, a sub-database containing only English letters spoken in isolation by 132 speakers was used. All speakers spoke Chinese as their native language, with different levels of Chinese-accented pronunciation for English letters. Speech was recorded with a Logitech USB Headset (LPAC-50000) in a quiet studio and sampled at 48 kHz. Each speaker read the alphabet set once and each utterance consisted of only a single letter. The training set included 50 males' and 50 females' speech, while the test set included 16 females' and 16 males' speech.

2.2 Chinese XIFs

The Initial/Final structure is a particular characteristic of Chinese syllables. Mostly each Chinese syllable consists of an Initial and a Final; however, some of them have a Final only. To make them consistent, 6 zero-Initials are designed so that the syllable without Initial is preceded by a zero-Initial. As a result, the basic speech recognition unit set in our Chinese LVCSR system is an extended Initial/Final set with 27 Initials and 38 Finals, among which 6 so-called zero-Initials are added so that each Chinese syllable is composed of an extended-Initial and a Final. The paper, [8], showed that the adoption of zero-Initials can help improve the performance effectively and build the tri-XIF HMMs consistently.

3 XIF-Based Acoustic Modeling

3.1 Decision Tree Based and Data-Driven State Clustering

In most of the state-of-the-art ASR systems, context-dependent acoustic modeling is commonly used, however, it is common that a quite large set of HMMs are built but a relatively small amount of training data is available for each HMM. In order to reduce the total number of parameters without significantly altering the models' ability to represent the different contextual effects, it is often to tie all of the central states across all models derived from a same mono-XIF. There are two popular dynamic clustering methods, the data-driven clustering method and the decision tree based clustering method, which actually are bottom-up and top-down, respectively [9].

However one limitation of the data-driven clustering procedure is that it does not deal with those tri-XIFs without examples seen in the training data. Although when building the intra-word tri-XIF systems, this problem can often be avoided by careful design of the training database but when building large vocabulary cross-word tri-XIF systems unseen tri-XIFs are unavoidable. As a matter of fact, the decision tree based clustering method can provide a similar quality of clustering but offer a solution to cover the unseen tri-XIFs [10, 11], which is commonly adopted in LVCSR systems. What’s more, decision tree based state clustering can be integrated closely with LVCSR modeling where minor modification is made for alphabet recognition.

3.2 State-Based Phonetic Mixture Tying

Due to the computing limitation of embedded devices, mixture tying is often used to further reduce the complexity in ASR systems. In the mixture tying (MT) [12] method, a single set of Gaussian mixtures is shared by all HMMs while each state has a different composition of mixture weights. As a result, the overlapping mixture distributions can be modeled properly with less Gaussians. As a variant of mixture tying, phonetic mixture tying (PMT) can define a set of Gaussian components independently in which each phone and the triphone variants of the phone share a certain Gaussian set [13]. In both MT and PMT, underestimation is likely to happen since each state has a large number of mixture weight parameters to estimate, and a large number of mixture weights are extremely small in magnitude. In the paper, to address this issue, a state-based phonetic mixture tying (SDPMT) [14] is adopted to build a robust recognizer for alphabet recognition. The key idea is illustrated in Figure 1. The SDPMT is performed based on the decision tree for a tri-XIF. The Gaussian mixtures from tied-states are pooled to become a Gaussian set and shared by these tied-states. In the figure, the second states of the triphones centered by the Initial, *an*, are presented by a decision tree and the leaf nodes stand for the tied-states. In SDPMT, the tied-states share a number of Gaussian mixtures, just encircled by the oval in the figure. The number of mixtures in a Gaussian set is determined by a threshold, which is usually a percentage of the total numbers of mixtures in all tied-states of a decision tree involved.

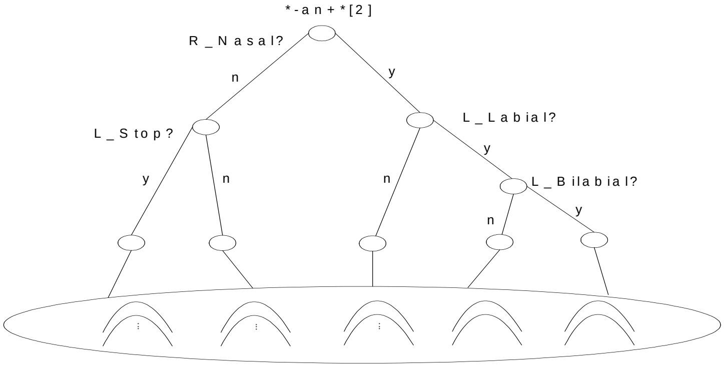

Fig. 1. State-based phonetic mixture tying for Initial *an*.

In SDPMT, due to sharing the overlapping Gaussian mixture distributions, the same amount of training samples is used to estimate the reduced number of mixtures, which leads to enhancement of model robustness. In addition, less time consumption is required during decoding which is of great benefit for embedded devices.

3.3 Pronunciation Modeling for Alphabet Recognition

It is well known that pronunciation modeling plays a significant role in non-native speech recognition. It is expected that it will also be helpful in the Chinese-accented alphabet recognition. For example, letter *c* is pronounced as /*si:*/ in English, but often as either [*s i*] or [*s uei*] by some (or even many) Chinese people. In the paper, pronunciation modeling is used to derive or to predict Chinese-accented pronunciations for English letters from English pronunciations. A decision tree which predicts the mappings between the base form English phoneme and the surface form XIF obtained by the following procedures similar to [2] is adopted as a pronunciation modeling method:

1. The base form transcription of English alphabet is obtained from a canonical lexicon with a single standard pronunciation for each English letter.

2. The surface form XIF transcription is obtained from the output of continuous Chinese XIF-based recognizer, where the Viterbi search algorithm is applied with an unconstrained network. As a result, these non-native English letters are represented by Chinese XIFs in surface form transcription. In the surface form transcription, some pronunciation variants with less consistency are removed which is based on the assumption that if a phoneme is correctly recognized, it always appears more often than a phoneme that is erroneously recognized.

3. The Chinese XIF-based transcriptions are used to train a decision tree, which maps the base form English pronunciation to the Chinese variants. The decision tree is then used to predict Chinese-accented variants from the English ones, which are added to the lexicon of the English alphabet ASR system.

4 Experiments and Results

4.1 Baseline

The context-dependent HMMs were intra-word tri-XIFs built upon the training set. Each XIF was modeled by a 3-state, left-to-right, non-skip, and non-loop HMM except that the skip and the loop among the states were allowable for the silence model. The number of Gaussian mixture components for each state was 8. Each speech frame was represented by a 39-dimensional feature vector with 12 MFCCs, log energy, and their first and second order time derivatives. Cepstral Mean Normalization (CMN) [15] was performed. The covariance matrices for each model were diagonal. The training and the evaluation procedures were both conducted via HTK v3.2 [16].

Aiming at evaluating the effectiveness of the methods of interest, we only used a subset of XIFs to construct the acoustic model for alphabet recognition. The adopted subset of XIFs was the necessary ones for representing the alphabet in the base form lexicon, which consisted of 30 out of the 65 mono-XIF units as listed in Table 1. Initially the context-dependent tri-XHF HMMs were built based on the context-independent mono-XIF ones, which resulted in. 57 tri-XIF HMMs and totally 171 states. We refer to these tri-XIF HMMs as the baseline. Results for the baseline were evaluated by the 32-speaker test set and a letter correctness of 95.9% was achieved.

Table 1. XIFs adopted in the base form lexicon for English alphabet recognition.

Initial(18)	Final(12)
b, p, m, f, d, t, g, k, q, zh, ch, z, s, _a, _o, _e, _u, _v	*a, ai, ei, en, er, ou, i, i2, iao, iou, ua, uei*

4.2 Results for XIF-Based Acoustic Modeling

In this section, experiments on state clustering, state-based phonetic mixture tying and pronunciation modeling were performed sequentially based on the baseline, and the overall results are listed in Figure 2. First we compared the decision tree based state clustering method with the data-driven state clustering method in Table 2. Results of the decision tree based state clustering method are listed in the left part while those of data-driven method in the right part.

Table 2. Results for the decision tree based vs. the data-driven clustering methods. (Two columns, *Threshold*, between the decision tree based and the data-driven methods are not comparable.)

Decision tree based state clustering			Data-driven state clustering		
Threshold	States	Letter Correctness	Threshold	States	Letter Correctness
100.0	175	96.5%	0.18	153	96.7%
250.0	161	96.8%	0.22	148	97.0%
300.0	156	97.1%	0.26	140	97.2%
350.0	155	97.1%	0.30	136	97.1%

It can be seen from Table 2 that the threshold can be used to adjust the number of the states effectively. Both the total number of states decreases with the increase of *Threshold*. For the decision tree based state clustering method, the XIF HMMs can reach an optimal point at 350.0, with 11.4% states reduction; accordingly for the data-driven method, 0.26 is the optimal threshold with 140 states in total. Furthermore, the tree-based and the data-driven methods, make a close match, 97.1% vs. 97.2%. Generally speaking, the data-driven state clustering is more appropriate for the small vocabulary tasks with sufficient training data. Thus, to some extent, it is more preferable for the data-driven method to be tailored to the current task. However the decision tree based clustering is more extensible which is able to cover unseen tri-XIFs in the training data, and what is more, it is expected that decision tree based state clustering can be integrated with Chinese LVCSR in which the decision tree based state clustering method is also adopted, we chose decision tree as our state clustering method in the following experiments. As a result, the letter correctness is improved from 95.9% in baseline to 97.1%, which is denoted by the columns, *CD* and *DTBST*, respectively in Figure 2.

With application of SDPMT preceded by decision tree based state clustering, the number of Gaussian mixtures among tri-XIF HMMs was reduced from 1,240 to 1,100 without causing performance degradation. The result, 97.1% in letter correctness, corresponds to column *SBPMT*, in Figure 2. SBPMT achieved an equal accuracy in comparison with decision tree based state clustering. It can be seen that 1) redundant Gaussian mixtures can be shared by tied-states at state level; 2) equally robust HMMs can be obtained with fewer Gaussian mixtures; 3) SDPMT results in no performance degradation for English alphabet recognition.

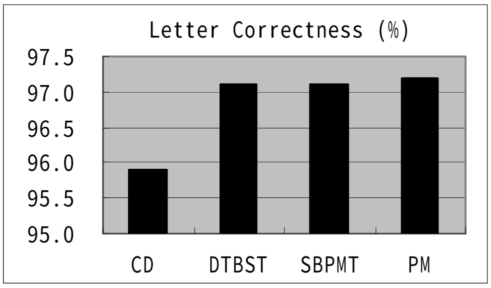

Fig. 2. Overall results for XIF-based acoustic modeling for alphabet recognition. CD stands for context-dependent modeling (baseline), DTBST stands for decision tree based state clustering, SBPMT stands for state-based phonetic mixture tying, and PM stands for pronunciation modeling.

In our task, most of the letters were designed to have one pronunciation entry in the lexicon and only several letters were designed to have 2 pronunciation entries, which is based on the experience that too many pronunciations can lead to performance degradation [17]. In total, 36 pronunciation entries were contained in the alphabet lexicon. By combining XIF-based acoustic modeling with pronunciation modeling, a letter correctness of 97.2% was achieved at best, as listed in column *PM,* in Figure 2, that is to say, the pronunciation modeling can lead to a letter correctness increase of 5.3% relatively for alphabet recognition. In a sense, the pronunciation modeling is modestly beneficial for alphabet recognition.

4.3 Comparison with Phoneme-Based Acoustic Model

In order to evaluate the effectiveness of the proposed method, we also built the acoustic model based on English phonemes to compare with the proposed Chinese XIF-based acoustic model. For simplicity, in the remaining part of this paper we will refer to the two models as the XIF-based modeling and the Phoneme-based modeling, respectively. For the Phoneme-based modeling, the exactly same training set and test set were utilized. The Phoneme-based modeling consisted of 27 English phonemes [15] for context-independent modeling. Additionally, similar measures were also taken to build the Phoneme-based model (triphone HMMs), *i.e.*, context-dependent modeling, decision tree based state clustering, state-based phonetic mixture tying and pronunciation modeling. As a result, a letter correctness of 97.1% was achieved by Phoneme-based HMMs which was slightly lower than XIF-based HMMs with a letter correctness of 97.2%. It is shown that the XIF-based method outperforms the phoneme-based method with a letter correctness increase of 5.3% relatively. Moreover, detailed comparisons between the two methods for every test speaker are listed in Figure 3. By analyzing the results, we found out XIF-based modeling could improve effectively for the test speakers with strong Chinese accent, namely, *Chinglish* speakers. In the test set, the speakers, No. 1, No. 3, No. 11, No. 12, *etc.*, were the ones with much strong Chinese accent. Taking speaker No. 1 for example, who was a native Beijing teenager, the utterance for letter *n* sounded much more like Chinese pronunciation [*_e eng*] than English pronunciation [*en*], and *k* was pronounced as "*ke*" [*k e*] or "*ki*" [*k i*]. With respect to speakers with less Chinese accent, the English phoneme-based modeling showed better performance than the XIF-based modeling, which was consistent with the expectation.

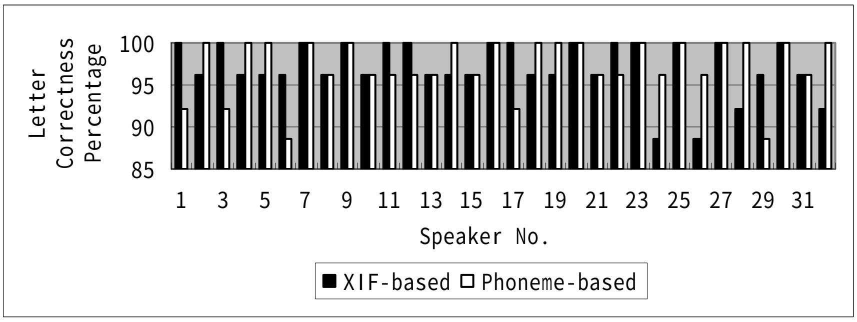

Fig. 3. Each speaker's letter correctness percentage for the XIF-based and the Phoneme-based modeling methods.

4.4 Integration with Chinese LVCSR System

To evaluate the effectiveness of Chinese LVCSR in combination with English alphabet modeling, A Chinese LVCSR model was built together with English alphabet. It is assumed that when a Chinese LVCSR system attempts to recognize the utterances containing English letters, the performance will deteriorate due to the increasing confusability introduced by English letters. Our goal here was to achieve good performance for continuous Chinese speech as well as English letters. In the experiment, exactly the same database as English alphabet modeling was used to train Chinese LVCSR acoustic model. A set of 65 XIFs was exploited as basic unit set. The identical 100 speakers were used as training set in which each speaker had 100 long sentences and 26 English letters. Accordingly the same group of 32 speakers was used for testing. In this experiment, two test sets, *test*1 and *test*2, were selected. In *test*1, each speaker had 100 utterances each of which consisted of one or two commands in Chinese; and a lexicon was composed of 200 commands for game interaction. In *test*2, besides the data in *test*1, each utterance was mixed with one or several English letters; likewise, another lexicon was composed of 200 identical commands and 36 pronunciations for English letters. The results are presented in Table 3.

From Table 3, it can be seen clearly that the performance is deteriorated due to the combination of English letters with Chinese syllables, dropping from 96.9% to 94.8% in syllable error rate, which lies in the fact that English letters are always confused with some Chinese syllables. To a great extent, the result was consistent with the expectation that no great degradation was brought in Chinese LVCSR system by the integration with English alphabet. In other words, it is shown that our proposed methods for alphabet recognition can well collaborated with Chinese LCVSR system on the basis of a same XIF set. Among the errors, some confusion is inherent in XIF-based acoustic modeling, such as *'yi'* vs. *'E'*. However, some confusion, such as *'er'* vs. *'R'*, can be discriminated by refined modeling.

Table 3. Evaluation with and without English letters via Chinese LVCSR system.

Model	Syllable Error Rate	
	*Test*1	*Test*2
Chinese LVCSR	96.9%	94.8%

5 Conclusion

In the paper, aiming at the Chinese-accented alphabet recognition, we propose to build the context-dependent tri-XIF HMMs, upon which the decision tree based state clustering is performed to refine the models. Subsequently, state-based phonetic mixture tying is adopted to further reduce the complexity while no performance degradation is introduced. The pronunciation modeling is performed as well to make it much suitable for Chinese-accented speech. Eventually, a letter correctness of 97.2% was achieved on English alphabet. In contrast to traditional English phoneme-based HMMs, our proposed method has achieved comparable accuracy for English alphabet recognition. Specifically, it is much effective for strong Chinese-accented alphabet recognition. Integrated with Chinese LVCSR system, it achieves the acceptable degradation on the utterances comprising Chinese words and English letters in comparison with those comprising Chinese words only. At present, to some extent, the alphabet has negative effect on overall performance in speech recognition, to which much effort should be made in the future.

Acknowledgments

The project was partially supported by Sony Computer Entertainment Inc, Japan. The authors would also like to give sincere thanks to Mr. Mingxing Xu, Mr. Jing Li, and Mr. Yuchun Pan for their help and valuable advice.

References

1. Tomokiyo, L.-M.: Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in LVCSR, PhD Thesis, Carnegie Mellon University, 2001.
2. Goronzy, S., Rapp, S., Kompe, R.: Generating Non-native Pronunciation Variants for Lexicon Adaptation, Speech Communication, Vol. 42, pp. 109-123, 2004

3. Wang, Z.-R., Schultz, T., Waibel, A.: Comparison of Acoustic Model Adaptation Techniques on Non-native Speech, IEEE ICASSP, 540-543, 2003
4. Loizou P.-C., Spanias, A.-S.: High-Performance Alphabet Recognition, IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 6, pp. 430-444, November, 1996.
5. Zissman, M.-A., Berking, K.-M.: Automatic Language Identification, Speech Communication, Vol. 35, pp. 115-124, 2001
6. Sipil, J.-I., Moberg, M., Viikki, O.: Multi-Lingual Speaker-Independent Voice User Interface for Mobile Devices, IEEE ICASSP, 2006
7. Li, J. Zheng, F., Zhang, J.-Y. Xu, M.-X., Wu W.-H.: The Definition and Extension of the Question Set for Decision Tree Based State Tying in Chinese Speech Recognition, International Conference on Chinese Computing, pp. 106-110, Nov. 27-29, 2001, Singapore
8. Zhang J.-Y., Zheng, F., Li, J. Luo, C.-H., Zhang, G.-L.: Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition, EuroSpeech, pp. 3:1617-1620, 2001, Aalborg, Denmark
9. Hwang, M.-Y., Huang, X.-D.: Shared-Distribution Hidden Markov Models for Speech Recognition, IEEE Transaction on Speech and Audio Processing, Vol.1, No. 4, pp. 414-420 October, 1993
10. Liu C.-J., Yan, Y.-H.: Robust state clustering using phonetic decision trees, Speech Communication, Vol. 42, pp. 391-408, 2004
11. Hwang, M.-Y., Huang, X.-D., Alleva, F.-A.: Predicting Unseen Triphones with Senones, IEEE Transaction on Speech and Audio Processing, Vol.4, No.6, pp.412-419, November, 1996
12 Zavaliagkos, G. McDonough, J., Miller, D., El-Jaroudi, A., Billa, J., Richardson, F. Ma, K.,Siu, M., Gish, H.: The BBN BYBLOS 1997 Large Vocabulary Conversational Speech Recognition System, IEEE ICASSP, pp. 905-908, 1998
13. Lee, A., Kawahara, T., Takeda, K., Shikano, K.: A New Phonetic Tied-mixture Model for Efficient Decoding, IEEE ICASSP, pp. 1269-1272, 2000
14. Liu, Y., Fung, P.: State-Dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition, IEEE Transaction on Speech and Audio Processing, Vol. 12, pp. 351-364, July 2004
15. Huang, X.-D. Acero, A., Hon, S.-W.: Spoken Language Processing, Prentice Hall, 2001
16. Yong, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.2), Cambridge University, UK, 2002
17. Lussier, E.-F.: A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, pp.38-77, 2003