ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

STUDY ON FRAMEWORK FOR CHINESE PRONUNCIATION VARIATION MODELING

LI Jing, XU Mingxing and WU Wenhu

Center of Speech Technology
State Key Laboratory of Intelligent Technology and Systems
Dept. of Computer Science and Technology, Tsinghua University, Beijing, 100084
{lijing, xumx}@sp.cs.tsinghua.edu.cn, wuwh@tsinghua.edu.cn

ABSTRACT

The pronunciation variations, which badly influenced the performance of ASR system, are serious in continuous speech, especially in spontaneous speech. Many research works are focused on pronunciation variation modeling in recent years. A framework for Chinese pronunciation variation modeling is described in this paper. The main idea is that the pronunciation variations are hidden in the recognition errors and it could be found out and modeled. The linguistic knowledge can be used for supervising the finding variations, and it can be used to model the variations directly. Under this framework, the pronunciation variation modeling can be made more reliably and more expediently than conventional methods. *Our primary experimental results show it is effective for Chinese continuous speech recognition.*

1. INTRODUCTION

Compared with the isolated word recognition, more and more research works deal with the continuous speech recognition for different tasks in recent years [1][2]. At the same time more issues must be faced for researchers in continuous speech recognition task, and the pronunciation variations is one of the importants. Many experiments show that the performance of an ASR system would be increased with modeling these variations, especially for spontaneous speech [3].

It is well known that the pronunciation variations are produced due to many aspects, such as context, co-articulation, accent, speaking rate, speaking style[4][5], and so on. In other words, for the same word (or other units) the pronunciation may be very different in different contexts, or pronounced by different speakers. The variations to be modeled must be found out before modeling.

Generally, there are two types of methods, i.e. the knowledge-based[6][7] and the data-driven based[8][9], could be used to model the pronunciation variations. For the knowledge-based approaches, the linguistic knowledge given by experts is needed. Unfortunately, it is impossible to collect enough knowledge for describing the pronunciation variations. And there would be mismatch between the experts knowledge and the database, for the knowledge are usually obtained from limited literatures that often great different from the real speech database. These drawbacks may lead worth pronunciation modeling. The data-driven methods, on the other hand, try to extract the variations information directly from training data and use the information to refine the acoustic model with a bottom-up or top-down strategy. It is difficult to integrate the linguistic knowledge in this processing. Because the knowledge on pronunciation variation are obtained from the data directly, the possible drawback of data-driven methods is that for the new corpus the whole process of transcribing the speech data and obtaining pronunciation variation has to be started again.

In this paper, a framework on Chinese pronunciation variation modeling is proposed. Mainly, this method is data-driven based, but the linguistic knowledge can also be integrated in this framework. The main idea is that the pronunciation variations are hidden in the recognition errors and could be found out and modeled. First at all, a context-dependent Initial/Final acoustic model is constructed by using decision tree based state-tying method, for the contexts greatly influenced the pronunciations. In this step, some Chinese linguistic knowledge are used to determine which states can be tied together. After this, the recognition errors are located, in which the pronunciation variations are hidden. And then the knowledge on pronunciation variations are collected from the recognition errors under the supervising of linguistic knowledge. Finally, the model can be reconstructed and the decoder can be refined to consider these variations. Some modeling methods can also be used in this step.

The proposed framework will be discussed in Section 2, and the details of the framework will be stated in the following Sections (Sction3 to Section 6). Some initial experimental results and conclusions will be given in Section 7.

2. PRONUNCIATION VARIATION MODELING FRAMEWORK

In this section, a framework integrated both linguistic knowledge and data-driven strategy for Chinese pronunciation variation modeling will be described, which consists of five levels showed as follows.

The first level is to prepare a primary acoustic model trained from a database with few pronunciations variations. At this level linguistic knowledge would be considered to build a context-dependent acoustic model (TriIF model in our experiments) by a tree-based state tying method [10].

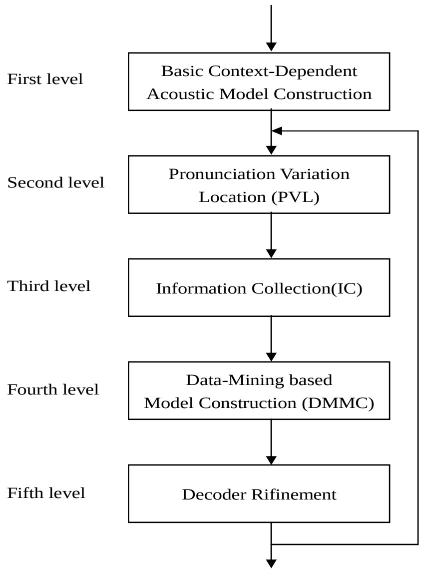

Figure1. Pronunciation Variation Modeling Framework

The second level is a Pronunciation Variations Location (PVL) for a spontaneous speech database. Similar to data-driven methods, a dynamic programming algorithm is used to generate force alignments for the training data. And then a confidence measurement is performed to determine which segments of speech stream are aberrant. After this, another recognition is invoked to find transcriptions for those variational slices.

The third level is called Information Collection (IC). Since recognition errors regarded as the results of pronunciation variations are located at PVL level, we could try to find the external and internal behaviors related to the variations. For example, the duration time of the syllable or the Initial/Final, the tone of the Final, the context, the Initial/Final confusion matrix, the score of the syllable or the Initial/Final, the speaker identifier, and so on.

The fourth level is named Data-Mining based Model Construction (DMMC). With data mining on the information collected at IC level, a special model is constructed with linguistic knowledge provided by the experts as well. In addition to this, some linguistic knowledge would be re-selected by verifying through the training data.

The final level of the framework is to integrate the model built at DMMC level into an acoustic decoder modified according to linguistic knowledge or combined with language models for spontaneous speech recognition.

In this framework, from the second level to the final level can be repeated to refine the model and decoder if needed.

3. CONTEXT-DEPENDENT ACOUSTIC MODELING

3.1 Basic Speech Recognition Units

Unlike the western languages, Chinese is naturally a syllabic language and each syllable has an Initial-Final structure. Previous experiments show that the acoustic modeling based on either the syllable or the Initial/Final (IF) can achieve a good performance [10]. Many Chinese speech recognition systems have been established, using syllables or IFs as the SRUs [13][14]. In this paper, the Extended Initial/Final (XIF) units are used as the basic units[10].

3.2 TriIF Modeling

In systems where the context information is completely or partly not taken into account to reduce the computational complexity, the accuracy is often not satisfying because the co-articulation of the speech is not well modeled. In order to gain higher recognition accuracy, the context-dependent modeling becomes an important part in the acoustic modeling. For Chinese speech recognition systems, two major issues should be considered. One is the selection of SRUs, while the other is the reduction of the scale of the context-dependent (CD) model.

Considering the contexts (e.g. the near left and right units), the XIF units become the TriIF units, like the triphone units. After designing linguistic knowledge based question set, the tree-based state-tying method is used to build the TriIF model. The recognition results show that using this kind of units to build the context-dependent model can get a better performance for Chinese continuous speech recognition [10]. In other words, the variations related to the contexts are modeled very well.

4. PRONUNCIATION VARIATIONS LOCATION (PVL)

4.1 Standard tanscription

A more accurate transcription is needed for a speech corpus, but it is difficult to transcribe a large speech corpus manually. Commonly, a fraction of the corpus is manually transcribed at first, and then some automatic segmentation algorithm is used to transcribe the speech. In our experiments, the original segmentation information is generated by the Merging-Based Syllable Detection Automation (MBSDA) [11] and the manual adjustment. For the TriIF model provides high performance in our experiments, the transcription is refined by using a dynamic programming algorithm to generate force alignments for the training data. And the new segment information is regarded as the standard transcription.

4.2 Pronunciation Variation Location

As mentioned above, the recognition errors are simply regarded as the results of pronunciation variations in our framework. So the task of PVL becomes to locate the errors.

The PVL processing consists of the following steps. Firstly, the standard transcription is obtained as mentioned in section 4.1. Secondly, all the speech data are recognized to get the results including the time alignment information. Finally, variations are located through comparing the recognition results with the standard transcription.

In our experiments, the PVL includes two levels, the syllable level and the Initial/Final level. It is well known that the Chinese is a syllabic language, and the syllable plays an important role in

Chinese speech recognition. So the pronunciation variation at syllable level should be located. On the other hand, the initial-final structure is a characteristic of Chinese language, and the XIFs have been chosen as the basic SRUs. Thus the pronunciation variation location at Initial/Final level is also performed.

5. INFORMATION COLLECTION (IC)

It is well known that the pronunciation variations are produced due to many aspects, such as context, co-articulation, accent, speaking rate, speaking style, and so on. But which contexts influence the pronunciations? And which variations are due to the speaking style? Obviously, existing linguistic knowledge is not sufficient yet. So, we try to find the external and internal behaviors related to the variations from the data under supervising of the linguistic knowledge. The following information that related to the pronunciation variation can be collected in IC level.

- **basic errors:** insertion, deletion and substitution errors (to find out which is the main error)
- **speakers' errors:** the errors according to the speakers (speaking rate, speaking style, accent and so on)
- **syllables'/XIFs' errors:** the errors related to the syllables and XIFs (to find out which models are bad described)
- **duration time information:** the average duration time of the syllables and XIFs
- **confusion matrix:** syllables' and XIFs' confusion matrix
- **syllables/XIFs error scores:** the scores of the wrong recognized syllables/XIFs, can be compared with standard scores
- **tone related errors:** the errors are related to which original tone (which unit with which tone are badly modeled)
- **contexts related errors:** the errors related to which contexts (which syllable or XIF in which contexts are badly modeled)
- **insertion and deletion error location:** where these errors are taken place, at the tail/head of the syllable or cross the syllable
- **relationship of the successive errors:** which errors are often taken place successively
- **others**

After IC processing, some information related to the pronunciation variations are collected, which will be used in the next level.

6. DATA-MINING BASED MODEL CONSTRUCTION AND DECODER REFINEMENT

The task of this level is to find the pronunciation variations from the collected information in the previous level. This level is a Data Mining processing, e.g. a knowledge discovering processing. It is expected to get such information,

- **influence of the contexts:** which contexts are badly modeled
- **influence of the speakers:** related to the speaking rate, speaking style, accent and so on
- **confusions:** related to the substitution errors
- **insertion errors:** related to the insertion errors
- **correlation of errors:** related to the successive errors
- **others**

The extracted information on pronunciation variations are regarded as the new knowledge that will be used to reconstruct the model or refine the decoder for pronunciation variation modeling with the linguistic knowledge provided by the experts as well. These knowledge can be used for the model reconstruction from the following aspects,

- re-estimating model parameters
- defining new basic units
- adding pronunciation variants to the lexicon
- building (weighted) pronunciation networks
- others

And some knowledge can also be used to refine the search algorithm to model the pronunciation variations.

7. RESULTS AND CONCLUSIONS

7.1 Speech Corpus

A continuous Chinese speech corpus from 863 materials is used in this paper. The corpus contains 80 speakers' data and 520 utterances are available for each speaker. All the recorded materials are obtained in a low noise environment through a close-talk noise-canceling microphone. 70 speakers' data are used as the training set while the remaining part is used for testing. They are digitized at 16KHz sampling rate. 13 dimensional MFCC, 1 dimensional log energy, and their 1st and 2nd order derivatives extracted in 24ms Hamming windowed frames every 12ms are used as the features. Cepstral Mean Normalisation (CMN) is applied over the 1s long segment. Each unit is divided into 3 states, and the tools in HTK v3.0 are used to build the acoustic models [12].

7.2 Primary Experimental Results

Some Primary experimental results including speaker related pronunciation and the insertion and deletion errors will be given in this section, and the results obtained by using some common used technologies like the confusion matrix are not discussed here. The in-depth study will be done in the future.

The recognition processing has been done on the training set, and the recognition results show that the TriIF model is relatively good. The accuracy is about 95.30%, and the primary errors are the substitution errors. They are great larger than the insertion and deletion errors. After studying every speaker's results, it is found that there are about 5 speakers in the training set, their speech results in a worse accuracy. The main reason is that all these 5 persons are pronounced with some accent, and this leads the recognition errors higher than other people. In other words, the current TriIF model used in our experiments can not catch the pronunciations. And the confusion matrix is used to get the confusion information in our experiments.

Although the insertion and deletion errors are not the primary errors, it is also badly influenced the model performance. So, the related information is collected to study the causes. The XIFs' dependent insertion and deletion errors are studied firstly. The results show that for the insertion errors, the errors related to {i, u, _y, _w, _a, a, _v, v, _e, e} are more than 52.67%, and for the deletion errors, the {_y, i2, _e, _w, u, l, v, _v, _a, er} are 58.81%. The results show that the insertion and deletion errors are closely related to the zero-Initials and its corresponding Finals (exclusive _o, for its samples are too fewer).

The insertion and deletion errors are located in PVL processing. After studied the recognition results, some helpful knowledge are generated by comparing with the standard transcriptions. For example: for units in {xie, xing, jiu, jing, jie, ju, you}, it is easy to insert an error insertion with unit _y, for units in {jie, jing, ji, jiu,jia}, unit i will be inserted into the syllable. For {ha, za , da, fa, ta} will attached with unit _a a , and so on. And the deletion will be occurred in the following units strings, such as " di yi", "yi yi", "yu yi", "li yi ai yi" and "xi yu".

7.3 Conclusions

In this paper, a framework for Chinese pronunciation variation modeling is described. The main idea is that the pronunciation variations are hidden in the recognition errors and could be found out and modeled. Under this framework, the pronunciation variation modeling can be made more reliably and more expediently than conventional methods. Our primary experimental results show it is effective for Chinese continuous speech recognition. The future work is to find more usable knowledge on pronunciation variations and to study how to model these variations better.

8. REFERENCES

[1] Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., Rosenfeld, R., "The SPHINX-II Speech Recognition System: An Overview," *Computer Speech and Language*, v. 2, 1993, p. 137-148

[2] Young S., "Large Vocabulary Continuous Speech Recognition: a Review", *IEEE Workshop on ASR*, 1995, Snowbird, December 1995, pp. 3-28

[3] Strik, H. and Cucchiarini, C., "Modeling pronunciation variation for ASR: A survey of the literature," *Speech Communication 29*, 1999, pp. 225-246

[4] Finke M., Waibel A., "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition," *In: Proceedings of Eurospeech Conference*, Rhodes, Greece, September 1997, pp. 2379-2382

[5] Ostendorf M., Byrne B., Bacchiani M., Finke M., Gunawardana A., Ross K., Roweis S., Shriberg E., Talkin D., Waibel A.,. Wheatley B, and Zeppenfeld T. "Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode," *In: International Conference on Spoken Language Processing,* Philadelphia, USA, 1996.

[6] Aubert X., Dugast C., "Improved acoustic-phonetic modeling in Philips's dictation system by handling liaisons and multiple pronunciations," *In: Proceedings of Eurospeech'95,* Madrid, pp.767-770

[7] Nock, H.J., Young, S.J., "Detecting and correcting poor pronunciations for multiword units," *In: Strik, H., Kessens, J.M., Wester, M. (Eds.), Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition,* Rolduc, Kerkrade, 4-6 May 1998. A2RT, University of Nijmegen, pp. 85-90.

[8] Cremelie, N., Martens, J.-P., "In search of better pronunciation models for speech recognition," *Speech Communication* 29 (2-4), 1999, pp.115-136.

[9] Torre, D., Villarrubia, L., Hernandez, L., Elvira, J.M., "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," *In: Proceedings of ICASSP97*, Munich, 1997, pp.1463-466

[10] Zhang J.Y., Zheng F., Li J., Luo C.H., and Zhang G.L., "Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition," *EuroSpeech*, 3:1617-1620, Sept. 3-7, 2001, Aalborg, Denmark

[11] Zhang J. Y., Zheng F., Du S., Song Z. J., and Xu M. X. "Merging-based Syllable Detection Automation in Continuous Speech Recognition". *J. of Software*, 10(11):1212-1215, 1999 (In Chinese)

[12] Yong, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., The HTK Book (for HTK Version 2.2), Cambridge University, 1999

[13] Zheng, F., Song, Z.-X, and Xu, M.-X, "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine", EuroSpeech'99, Vol.2, pp.819-822, Budapest, Hungary, 1999

[14] Li, J., Zheng, F., and Wu, W.-H., "Context-Independent Chinese Initial-Final Acoustic Modeling," International Symposium on Chinese Spoken Language Processing, pp. 23-26, Beijing , 2000