Large Vocabulary Mandarin Chinese Continuous Speech Recognition System Based on Tonal Triphone

Long Yan, Rencai Zhao,Gang Liu, Jun Guo

School of Information Engineering
Beijing University of Posts and Telecommunications, Beijing
yanlong@pris.edu.cn

Abstract

Large vocabulary mandarin Chinese continuous speech recognition has been a difficult problem in speech recognition area because of several reasons. First, it is a tone language. There are five lexical tones that are important in distinguishing the confusable words in mandarin. So the modeling of tones plays an important role in mandarin speech recognition. Second, the variation of tones in spontaneous mandarin speech would have some effects on the performance. Third, the co-articulation is inevitable in spontaneous mandarin speech recognition.

In this paper, a large vocabulary mandarin Chinese continuous system based on tonal triphone was constructed. The experimental results shows that a good performance in acoustic level has been achieved while poor performance in word level.

1. Introduction

Large vocabulary speaker-independent continuous speech recognition has been an important problem for speech recognition researchers. As far as Mandarin Chinese has been concerned, it has some characteristics that are important for speech recognition task. Unlike the western languages, Chinese is naturally a syllabic language and each syllable has an Initial-Final structure. Five lexical tones play an important role in distinguishing the confusable words. In large vocabulary continuous speech recognition, pitch information does not improve the performance much. However, pitch information serves an important role in isolated word recognition[1].

Nowadays, in speech recognition, a lot of tools that make the research convenient have arisen. Such as Hidden Markov Toolkit(HTK)[2] which is developed by Cambridge University Engineering Department, CMU-Cam_Toolkit[3] which is co-developed by Carnegie Mellon University and Cambridge University, Julius[4] which is developed by Kyoto University and IPA.

The structure of this paper is as follows. In Section 2, the acoustic modeling is described. The language modeling is mentioned in Section 3. In the following section, the recognition engine is described. The experimental condition and results are given in Section 5. And at last we come to the conclusion in Section 6.

2. Acoustic modeling

The acoustic modeling is one of the primary processes in large vocabulary mandarin Chinese continuous speech recognition systems. There are some relations between acoustic modeling and mandarin Chinese pronunciation.

2.1. Selection of basic acoustic units

The selection of basic acoustic units is the very important in acoustic modeling. Typically there are three types of basic acoustic units for mandarin Chinese Context-Dependent acoustic modeling: syllable, Initial/Final or phone. Mandarin Chinese is a tone language, each character is pronounced as a monosyllable with a tone association. There are about 410 toneless syllables and 1250 tonal syllables totally. Choosing syllables as basic acoustic units will increase the computation and storage complexities greatly. Compared with syllable, the phone unit is rather small and there are only a small number of phones, but phones vary very much in pronunciation. There are often phone deletions, phone insertions and phone changes in continuous speech. However, Initial/Final is relatively steady. There are 24 Initials and 25 Finals in mandarin Chinese. In this study, we selected the Initial/Final approach. The table 1 shows the final units with tones.

Table 1.*Final units with tones.*

Final Units With Tones	a(0-4) e(0-4) i(0-4) u(0-4) an(0-4) ao(0-4) en(0-4) ing(0-4) iang(0-4) iao(0-4) o(0-4) ai(0-4) in(0-4) ou(0-4) uo(0-4) vn(0-4) ong(0-4)v(1-4) ei(1-4) ia(1-4) ie(1-4) iu(1-4) ua(1-4) ui(1-4) un(1-4) ve(1-4) eng(1-4) ian(1-4) uai(1-4) uan(1-4) van(1-4) uang(1-4) ang(1-4) er(2-4) iong(1-3)

2.2. Context Dependent Modeling[5]

If we make acoustic units context dependent, we can significantly improve the recognition accuracy, provided there are enough training data to estimate these context-dependent parameters. Context-dependent phonemes have been widely used for large-vocabulary continuous speech recognition. A context usually refers to the immediately left and/or right neighboring phones. A triphone model is a phonetic model that takes into consideration both the left and the right neighboring phones. Triphone models are powerful because they capture the most important coarticulatory effects. Triphones has four types in mandarin Chinese.(SilB-Initials+Final, Initials-Finals+SilE, Initials-Finals+Initials, Finals-Initials+Finals).In this paper, we use tonal triphone model in acoustic modeling.

2.3. HMM definition

Figure.1 shows a simple left-right HMM with five states in total. Three of these are emitting states and have output probability distributions associated with them.

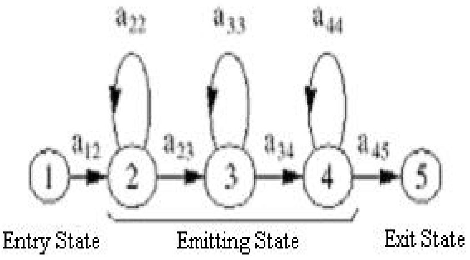

Figure 1. *Simple Left-Right HMM.*

2.4. Model Training

Figure.2 shows the training procedure which is based on HTK. The HTKBook[2] can give the details.

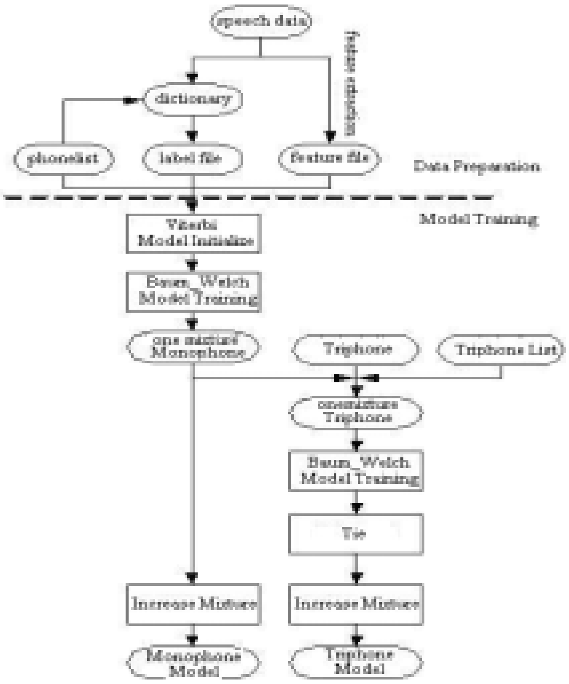

Figure 2. HTK Training.

3. Language modeling

A large vocabulary speech recognition system is generally critically dependent on linguistic knowledge embedded in the input speech. Therefore, for large vocabulary speech recognition, incorporation of knowledge of the language, in the form of a “language” model, is essential. N-gram language models are constructed based on the lexicon. Specially, word 2-gram and 3-gram models are trained using back-off smoothing. Witten-Bell discounting method is used to compute back-off coefficients. The language model is obtained by using CMU-Cam_Toolkit which can be downloaded from the Internet[3].

4. Recognition engine

A recognition engine Julius is developed for evaluation of both acoustic and language models by the IPA (Information-technology Promotion Agency) and School of Information, Kyoto University. It is a sharable software. It can deal with various types of the models. It accepts not only wave files (16bit PCM) and acoustic parameter files (HTK format) but also microphone input(Sun/SGI workstation, Linux PC, via DAT-LINK/netaudio). Speech analysis is implemented only for those parameters adopted by the acoustic model of toolkit.

Julius performs a two-pass (forward-backward) search using word 2-gram and 3-gram on the respective passes. In the first pass, a tree-structured lexicon assigned with language model probabilities is applied with the frame-synchronous beam search algorithm. It assigns pre-computed 1-gram factoring values to intermediate nodes, and applies 2-gram probabilities at the word-end nodes. Cross-word context dependency is handle with a approximation which applies the best model for the best history. In the second pass, 3-gram language model and accurate sentence-dependent acoustic model is applied for rescoring. In this paper, we only use the first pass in decoding.

5. Experiments

A continuous Chinese speech corpus from 863 materials is used. 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. 40 speakers’ data are used as the training set while the remaining part is used for testing. The speech data were sampled at 16KHz and 16bit. Twelfth-order mel-frequency cepstral coefficients (MFCC) are computed every 10ms. Temporal difference of the coefficients ($\Delta\mathrm{MFCC}$) and power ($\Delta\mathrm{LogPow}$) are also incorporated. So the feature vector at each frame consists of 25(12+12+1) variables. Each model consists of three states excluding the initial and final states that have no distributions. The state transitions are all left-to-right, and the path from the initial state and that to the final state are limited to one, and the tools in HTK v3.0 are used for building our acoustic models. The experimental results below are given in the acoustic level and word level. For training set, the correct rate of Initial/Final is 86.49% and correct rate of word is 72.4%. For testing set, the correct rate of Initial/Final is 85.14% and correct rate of word is 59.46%.

Table 2.*Performance of the system.*

	IF result(%)	word result(%)
Training set	86.49	72.24
Testing set	85.14	59.46

6. Conclusion

In this paper, by using the tools (HTK, CMU-Cam_Toolkit and Julius),a method of building a large vocabulary mandarin Chinese continuous speech recognition system has been presented. The experimental results show that the performance of the system in the acoustic level is good, but in word level the performance is not very good. The poor language model may affect the performance in word level. In our future work, we will try to improve the language model and utilize more information of the linguistic knowledge.

7. References

[1] Y.W.Wong and Eric Chang,”The Effect of Pitch and Lexical Tone on Different Mandarin Speech Recognition Tasks,” Eurospeech 2001,vol 4, pp. 2741–2744.

[2] Steve Young, Dan Kershaw, et al. The HTK Book, July 2000.

[3] The CMU-Cambridge Statistical Language Modeling Toolkit v2 document, 1997.

[4] Akinobu Lee, Tatsuya Kawahara, Kiyohiro Shikano” Julius— an Open Source Real-Time Large Vocabulary Recognition Engine,” Eurospeech 2001,vol 3,pp 1691-1694.

[5] Jiyong Zhang, Fang Zheng , Jing Li,” Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition,” Eurospeech 2001,vol 3,pp. 1617–1620.