International Symposium on Chinese Spoken Language Processing, pp. 207-210, Oct. 13-15, 2000, Beijing

TONE RECOGNITION OF CHINESE CONTINUOUS SPEECH

Guoliang ZHANG, Fang ZHENG, Wenhu WU
Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing 100084
liang@sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper our approach to the lexical tone recognition of Chinese continuous speech is presented. The Mixed Gaussian Continuous Probability Model (MGCPM) [1] is used for the tone modeling, and the quadric curve is adopted to simulate the Fundamental frequency (F0) contour, whose three coefficients are calculated and taken as the features of the tone models. The tone variety in continuous Chinese speech recognition is an issue that must be faced in the tone modeling. There are two kinds of tone varieties, the change from canonical one to non-canonical one without changing the pitch trend and that from one to another different one. In order to reduce the negative influence caused by the tone varieties, an iterative method is proposed to distinguish the syllables which have tone varieties and remove them from the whole training data, and then the Tone Variety Matrix (TVM) is introduced for improving the performance of tone models. Experiments have been done based on the continuous Chinese speech database named "863" database. The top1 and top2 accuracy for baseline MGCPM is 67% and 90%, while that for MGCPM incorporated with TVM is 70% and 92%.

1. INTRODUCTION

Chinese is a typical tonal language, in which each Chinese character corresponds to a monosyllable and basically has a (C)V phoneme structure with a lexical tone. The lexical tone is mainly associated with the vowel part V that is referred to the Final part. There are four lexical tones (Tone 1, Tone 2, Tone 3, and Tone 4) and a neutral tone in Chinese. As we all know, accurate tone recognition is of a great help to Chinese speech recognition systems, because there will be a large number of homophone words when the tone information is discarded, the language models will promote the performance with the accurate tone recognition result. Another important application of the tone recognition lies in the Speaking Skill Evaluation (SSE) systems [2] where guides can be given to foreigners to learn the difficult-to-learn lexical tones.

Usually, the tone classification is based on the F0 contours. Except for the neutral one, the Chinese tones mostly have stable F0 contours in trend. Several methods have been applied successfully to the tone recognition for isolated syllable in the past few years. For example, HMM had been applied for four-tone recognition [3][4]. However, contours in continuous speech vary due to various reasons, such as voicing of initial consonants, tonal coarticulation, sentential intonation, and etc [5]. The F0 contours vary so greatly that the conventional methods for isolated syllable do not work well. Therefore, robust tone recognition method in continuous speech is still not available up to now.

In continuous speech, speaking rate is often fast, which will cause problems to the methods based on deterministic model. In this paper, the statistical method is used. The method presented here is based on the MGCPM, which has been successfully applied to many fields. The features of MGCPM are three coefficients of the quadric curve that has the ability of describing the variable trend of the F0 contours in a whole syllable.

The detection of tone varieties is also a difficulty for Chinese tone recognition. In actual pronunciations, the varieties of tones are too complicated to describe by linguistics rules and the syllables of tone variety account for a fair proportion. In this paper, two methods are presented to reduce the errors introduced by tone varieties, one is an iterative method that is used to remove the samples with tone variety from the whole training data, another is the Tone Variety Matrix (TVM) that describes the information of varieties of tones. The two methods will be described in Section 3 deathly.

In Section 2 the tone model and the feature extraction are described. In Section 3 the strategy of detecting tones varieties is presented. Some experimental results are given in Section 4 while conclusions are finally given in Section 5.

2. TONE MODEL

2.1 MGCPM

Researches and experiments on the distance measure between models shown that transition probability matrix plays a far less important role in HMM than the observation probability matrix does [6][7][8], so a kind of Segmental Probability Model (SPM) has been proposed based on the desertion of the HMM transition probability matrix. The MGCPM is an example of the SPM, in which the intra-state feature space is described via mixed Gaussian densities (MGDs)[9]. During the recognizing procedure, the modified Viterbi [10] algorithm is adopted for state decoding.

In our Chinese Dictation Machine Engine (CDME), the 6-state 16-MGD based MGCPM is adopted, which has been proved efficient. MGCPM achieves a satisfying top-10 isolated syllable recognition accuracy of 99.05% for 30-male training set and

95.65% for 8-male testing set across the 863 Database. The model is as good as the traditional HMM but much faster and smaller. In this paper, a 16-MGD based on MGCPM is used for tone modeling.

2.2 Feature Extraction

Typical F0 contour shapes for 4 basic lexical tones are illustrated in Fig.1, which are often observable in continuous speech, and even in isolated syllable cases.

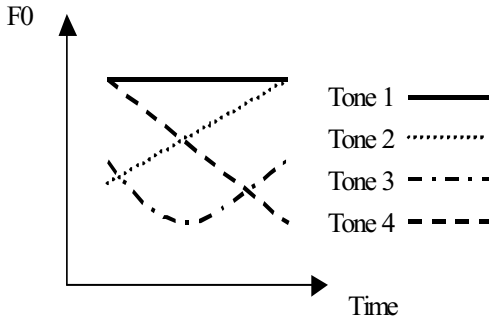

Fig.1. Typical F0 contour shapes for 4 basic lexical tones

Due to physiological articulatory constraints of the human vocal cord vibration, these patterns are rarely kept unchanged when uttered in continuous speech, furthermore some utterances change so greatly that they will be recognize as another tone. Therefore, the F0 contours shapes of syllable will be normalized at first. The syllable sequence length must be normalized to a fixed length as follows:

$$T_n(i) = (L/l) * i \qquad i = 1,2,\cdots,l$$

where l is the length of syllable's F0 contours and L is the normalized length. The range of F0 varieties also must be normalized into this fixed range. If the range of F0 exceeds the threshold, then the normalized $F_n 0$ sequence has

$$F_n 0(i) = (F0(i) - \min(F0)) * (\max(F0) - \min(F0))/H$$
$$i = 1,2,\cdots,l$$

where H is the normalized varieties range of F0.

The quadric curve f is used to simulate the F0 contour, since the quadric curve has the ability of describing the variable trend of the F0 contours in a whole syllable. The f is shown as:

$$f(x) = ax^2 + bx + c$$

Given the $T_n(i)$ and $F_n(i)$, minimize the following

$$E = \sum_{i=1}^{l} (f(T_n(i)) - F_n 0(i))^2$$

The three coefficients of the quadric curve f are calculated and taken as the features of the tone models.

F0 contour variations in continuous speech lead to some inaccuracies in the extraction of F0 contour, so the features that are extracted from the inaccurate F0 contour will also result in errors. Therefore, some rules are introduced to distinguish the "error" features. For example, the quadric curve with high variance is marked as an “error”.

2.3 Tone Models

The mixed Gaussian densities (MGD) is described as follows,

$$p(x \mid \lambda) = \sum_{m=1}^{M} g_m p(x \mid \mu_m, \Sigma_m)$$

where $\lambda = \{g_m, \mu_m, \Sigma_m\}$ is the model parameters set, M is the number of density mixtures, $p(x \mid \mu_m, \Sigma_m)$ is a Gaussian pdf and g_m is the gain or weighting of corresponding Gaussian pdf. The covariance matrix Σ of a Gaussian pdf is often a diagonal one as $\Sigma = (\sigma_d^2)_{D \times D}$.

The parameter estimation of MGD is very important. There are several techniques available for estimating the parameters of a MGCPM [11], in this paper, the famous estimate-maximize (EM) algorithm [12] is chosen for training, while the initial estimation for mean vectors μ_i and variances Σ_i are selected by the LBG algorithm [13].

The Chinese neutral tone is complicated and has a very short duration in actual pronunciations, so we leave it in future study. Each lexical tone has two sets of model parameters, $\{\lambda_i, \lambda_i^{\Delta}, i = 1,2,3,4\}$, associated with the F0 contour and delta F0 contour respectively.

The log likelihood is defined as

$P(x \mid \lambda_i, \lambda_i^{\Delta}) = \log p(x \mid \lambda_i) + \log p(x^{\Delta} \mid \lambda_i^{\Delta})$From the above description, the tone classification result can be easily combined with the result of acoustic models in the language model, since the tone classification result is given in probability.

3. REDUCING THE ERRORS INTRODUCED BY TONE VARIETIES

The results of the above experiments are not satisfying. After analyzing the results, we find that the samples of tone varieties account for a fair proportion in the whole database, which leads to the confusion in the tone recognition. As we know, the detection of tone varieties is also a key point in the tone recognition of Chinese continuous speech. Moreover, varieties of tones are too complicated in actual pronunciations to describe by linguistics rules. The conventional methods almost pay attention to the reorganization of the F0 contour and neglect the description of the tone varieties. Indeed, the accurate tone varieties description is precondition of the accurate tone recognition in Chinese continuous speech. In this paper, the following methods are presented to reduce the errors.

First, an iterative method is used to remove the samples with tone variety from the whole training data so that the tone models are not affected by tone varieties. The method can be described as follows.

Step 1: The training data are initialized as the whole training data.

Step 2: The tone models are trained by the training data.

Step 3: Recognize every sample in the training data. The sample whose classification does not consist with its labeled tone will be regarded as having tone variety and it will be removed from the training data.

Step 4: Decide whether the result satisfy convergence requirement. If it does not, repeat step 2 and step 3 until convergence.

Second, the Tone Variety Matrix (TVM) is used. The TVM describes the information of varieties of tones, whose elements are the conditional probability of sample's labeled tone given the classification of the current sample and the last sample by our tone models.

$$f_{k,ji} = p\left(label_t = k \mid recog_{t-1} = j, recog_t = i\right)$$

$$i,\ j,\ k\ = 1,2,3,4$$

In fact the TVM is a right-dependent confusion matrix, which describes the influence of last sample on current one in continuous speech. The product of MGCPM and TVM gives the final result as follows

$$Q\left(x_t \mid \lambda_k, \lambda_k^{\Delta}\right) = \sum_i \sum_j \{P\left(x_t \mid \lambda_i, \lambda_i^{\Delta}\right) + P\left(x_{t-1} \mid \lambda_j, \lambda_j^{\Delta}\right) + \log f_{k,ji}\} \qquad i,j,k = 1,2,3,4$$

4. EXPERIMENTS

Tone recognition experiments have been carried out across the continuous Chinese speech database named 863 database, uttered naturally by 80 people aged form 16 to 25 from all over the country. The training data include 250,595 syllables of 30 male's data, and the testing data include 68,904 syllables of 8 male's data. The corpus offers syllable and lexical tone labels together with sentence text transcriptions, and the boundaries of all syllables in each utterance are pre-labeled manually. In the 863 database, speech signal is sampled at 16kHz sampling rate with 8kHz cut-off through the SoundBlaster under the PC environment. F0 is calculated for every frame with 16ms step. The F0 extraction method will bring some F0 extraction error, so some rules are introduced to distinguish the "error" utterance, which are present in Section 2.2, almost 10% utterance are eliminated from the whole database.

The baseline method is introduced in Section 2, in which the tone models are based on the MGCPM and the three coefficients of the quadric curve are adopted as the feature. The top1 and top2 accuracy for baseline MGCPM is 67% and 90%, while that for MGCPM incorporated with TVM is 70% and 92%.

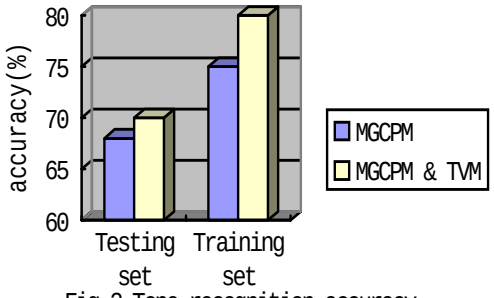

Fig.2 Tone recognition accuracy

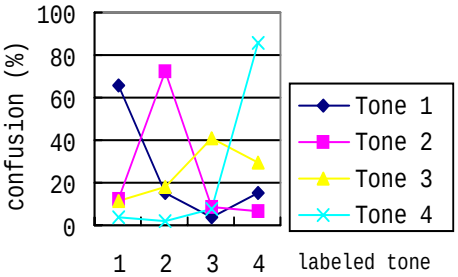

Fig.3.Confusion of Tone recognitio

Recognition results for the proposed method are shown in Fig.2, and the confusions of Tone recognition are shown in Fig.3. From the result, we may conclude:

1. The proposed method using MGCPM has some advantages: good performance and high efficiency. It can be easily combined with the Chinese Dictation Machine (CDM), since the tone classification result is given in probability.

2. Tone recognition based on MGCPM and TVM is better than that based on MGCPM only. This result shows that our methods for reducing the tone variety errors are effective.

3. The utterances of tone varieties account for a fair proportion in Chinese continuous speech, so the accurate detection of tone varieties is the key point in tone recognition. Tone 3 is the most complicated one in continuous Chinese speech. The recognition accuracy for tone 3 is lower than 50%, whereas tone 1 has a relatively stable shape. Therefore, further study is needed on the description of the tone varieties.

4. Even the highest tone recognition rate for testing set is still rather low for actual use. Besides the possible incorrect methods of tone modeling, the reason also lies in the negative influence exerted by tone varieties, which affect the accuracy of tone recognition greatly.

5. CONCLUSION

We apply a new method based on MGCPM to recognize the lexical tones of continuous Chinese speech. Moreover, an

iterative method is proposed to remove the syllables that have tone varieties from the training data, and then the Tone Variety Matrix (TVM) is introduced for improving the performance of tone models. Although the proportion of top1 tone recognition accuracy is just 70%, we believe that further study is needed on this topic. How to describe the varieties of tones is the key to the tone recognition in continuous speech recognition. A database with detailed tone labels in actual pronunciations is probably helpful.

6. REFERENCES

[1] Zheng F., Mou X.-L., Wu W.-H., and Fang D.-T. (1998), "On the embedded multiple-model scoring scheme for speech recognition". *International Symposium on Chinese Spoken Language Processing (LSCSLP'98)*, ASR-A3, PP.49-53, Dec.7-9, 1998, Singapore

[2] Song ZH.-J., Zheng F., Xu M.-X., Wu W.-H., (1999), "An Effective Scoring Method for Speaking Skill Evaluation System". *EuroSpeech'99*. Vol.1, pp.187-190, Budapest, Hungary, Sept. 1999

[3] Chen X.-X., Cai CH.-N., Guo P., Sun Y., (1987), "A Hidden Markov Model Applied to Chinese Four-Tone Recognition", *ICASSP 1987*, PP.797-800

[4] Yang W.-J., Lee J.-CH., Chang Y.-CH., Wang H.-CH., (1988), "Hidden Markov Model for Mandarin Lexical Tone Recognition". *IEEE Trans. on ASSP,* Vol.36, No.7, July 1988, pp.988-992

[5] Hirose K., Zhang J.-S., (1999), "Tone Recognition of Chinese Continuous Speech Using Tone Critical Segments". *EuroSpeech'99,* Vol. 2, pp.879-882, Budapest, Hungary, Sept. 1999

[6] Juang B.-H., and Rabiner L.-R., (1985), "A Probabilistic Distance Measure for Hidden Markov Models", *AT&T Technical Journal*, Vol.64, No.2, pp.391-408, Feb., 1985

[7] Rabiner L.-R., and Juang B.-H., (1986), "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Vol.3, No.1, pp.4-16, Jan., 1986

[8] Lee K.-F (1989), "Automatic Speech Recognition – The Development of the SPHINX System", *Kluwer Academic Publishers*, Boston, 1989

[9] Bahl L. R., Brown P. F., de Souza P. V., Mercer K. L., (1990), "Speech Recognition with Continuous-parameter Hidden Markov Models". *Readings in Speech Recognition.* Alex Waibel & Kai-Fu Lee (eds.), 1990, pp.332-339

[10] Zheng F., Wu W.-H., and Fang D.-T.(1998), "Center-distance continuous probability models and the distance measure". *J. Of Computer Sci. & Tech.,* 13(5): 426-437, Sept. 1998

[11] McLachlan G., (1988), "Mixture Models", New York: Marcel Dekker, 1988.

[12] Dempster A. P., Laird N. M. and Rubin D. B., (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Proc. R. Stat. Soc. B*. 39(1):1-38, 1977

[13] Linde Y, Buzo A and Gray R M. (1980), "An Algorithm for Vector Quantization Design". *IEEE Trans. on COM*-28(1), Jan., 1980