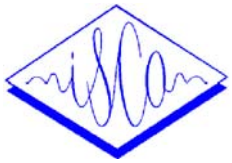

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

STUDY ON TONE CLASSIFICATION OF CHINESE CONTINUOUS SPEECH IN SPEECH RECOGNITION SYSTEM

LIU Jian , HE Xiaodong , MO Fuyuan , YU Tiecheng

Speech Processing Laboratory , Institute of Acoustics , Chinese Academy of Sciences
P.O.Box 2712, Beijing 100080, P.R.CHINA

Email: {lj , hexd , mfy , tcyu}@speech1.ioa.ac.cn
http://www.speech.ioa.ac.cn

ABSTRACT

In this paper, we first introduce the use of Gaussian mixture models (GMM) for Chinese tone classification in continuous speech. Then, we explain how to integrate it with the HMM-based speech recognition system. Finally, we provide the tone classification accuracy of this probabilistic method which is tested with Chinese continuous speech database of national "863" project.

I. INTRODUCTION

Chinese language is a tonal language. Each Chinese character is a monosyllabic lexicon. There are totally 408 syllable sounds if the tones are ignored. But if different tones are distinguished, there will be almost 1300 sounds. Same syllable with different tones will have distinct meanings. Therefore, tone classification plays an important role in Chinese speech recognition.

There are four tones in Chinese isolated syllables. Several studies on tone recognition for isolated syllable have been conducted in the past few years. HMM had been applied for recognizing four tones by Chen *et al.* [1]. Vector quantization and HMM had been combined and used in four-tone recognition by Yang *et al.* [2]. Neural network was used in four-tone recognition by Lim *et al.* [3]. Multi-Layer Perceptron (MLP) had been applied on four-tone recognition by Chang et al. [4]. Fuzzy Sets theory had been applied on four-tone recognition by Xu *et al.* [5].

But in continuous Chinese speech, the variety of tones can't be classified just to four types. At least, a "zero-tone" must be introduced to represent one other kind of tone, while speaker speaks gently and quickly, in one sentence. In general, the shape of pitch contour is the basic rule of classification. The same tone pronunciation has the similar shape. To distill the pitch exactly from the continuous speech has been described as the "hardy perennial" among speech-processing problems. To categorize the pitch contour to right tone class is also complicated, especially in continuous speech. This may be the main reason that most continuous speech recognition systems don't explicitly utilize the tone information in their language processing part.

The method presented here is based on the Gaussian Mixture Model (GMM) which has been successfully applied to both speaker identification [6] and language identification [7]. To our best knowledge, it is the first attempt to use GMM in the field of Chinese tone classification.

Section II will describe the detail about our Gaussian Mixture Tone Models (GMTM). Section III presents the application of GMTM technique for tone classification of Chinese continuous speech. The experiment result will be shown in section IV. The conclusion is finally given in section V.

II. GAUSSIAN MIXTURE TONE MODEL

A. Normalize Pitch Sequence

Let T0 denotes the sequence of raw pitch period for one syllable. Since the length of the sequence may be different due to the arbitrary utterance length, T0 must be normalized to fix length before being used for building GMTM of each tone. If the original pitch period sequence length is K and to be normalized to length D, then the normalized pitch period sequence T0' has

$$T0(i) = \sum_{j=floor((i+1)\cdot K/D+1),k=1}^{ceil(i\cdot K/D),k=k+1} T0(j) * W(k)$$

wherein *Ceil(x)* rounds the elements of *x* to the nearest integers towards infinity, *Floor(x)* rounds the elements of *x* to the nearest integers towards minus infinity. *W(k)* is a triangle window, whose length varies with different *i*, but the summation of its coefficients is always assured to be 1. D is appointed to 16 since 16 points are enough for describing the pitch contour of each tone.

B. Build GMTMs

Under the GMM assumption, each feature vector *X* is assumed to be drawn randomly according to a probability density that is a weighted sum of multi-variate Gaussian densities:

$$p(X \mid \lambda) = \sum_{i=1}^{M} w_i \cdot b_i(X)$$

wherein λ is the set of model parameters, $\lambda = \{ w_i \mu_i \Sigma_i \}$. *M* is the order of the GMM. w_i are the mixture weights and they satisfy the constraint that $\sum_{i=1}^{M} W_i = 1$. $b_i(X)$ are the multi-variate Gaussian densities defined by the means μ_i and variances Σ_i.

For each tone, two GMTMs are created: one for the normalized pitch vectors T0, *{X}* and one for the pitch difference vectors ΔT0, *{Y}*.

There are several techniques available for estimating the parameters of a GMM [8]. Here we use multiple iterations of the estimate-maximize (EM) algorithm for parameter estimation [9]. The initial estimates for mean vectors μ_i of GMTMs are random selected means followed by a single iteration *k*-means clustering [11] on whole train data. Initial variances Σ_i are identity matrix, and equal initial mixture weights are used. The iteration stops when the relative improvement of the model output probability is less than 0.01.

While training a GMTM, it has been observed that variance elements can become quite small value. This is particularly true for a large mixture order. These small variances produce a singularity in the model's likelihood function and can degrade classification performance. To avoid these singularities, a variance limiting constraint is applied. For an arbitrary element of mixture component variance, the minimum value is set to 0.01. This constrained version of EM algorithm has been shown to provide more robust parameter estimates than the unconstrained version [8][10].

C. GMTM Classification

For tone classification in Chinese continuous speech, five tones are represented by GMTMs $\{ \lambda_k^{T0}, \lambda_k^{\Delta T0}, k=1,2,3,4,5 \}$. Obviously, if for isolated Chinese syllable tone recognition, only 4 GMTMs need to be built not 5. The log likelihood for one tone is defined as

$$Q(\{X,Y\} \mid \lambda_k^{T0}, \lambda_k^{\Delta T0}) = \log p(X \mid \lambda_k^{T0}) + \log p(Y \mid \lambda_k^{\Delta T0})$$

So the objective is to find the tone model â, which has the maximum *a posteriori* probability for a given sequence, where

$$\hat{a} = \arg\max_a Q(\{X,Y\} \mid \lambda_a^{T0}, \lambda_a^{\Delta T0}) \quad 1 \le a \le 5$$

From above description, we can see both the training procedure and the recognizing process of Gaussian Mixture Tone Model are very simple. The GMTM we introduced here has the following characteristic: It needs no artificial thresholds to distinguish the pitch trends of different tones. So there is no limitation that it can be only applied for Chinese speech. It can be used as a common tone recognition method for any tonal language, only if this tonal language has fix number of tone modes.

III. TONE CLASSIFICATION FOR CHINESE CONTINUOUS SPEECH

Most continuous speech recognition systems for spoken Chinese do not provide tone classification information at their acoustic model level output. To archive the final recognition result of the whole sentence, they only depend on the language model. Although this kind of systems performs quite well, we think it needs to be improved further.

As one of three essential factors in Chinese phonology, tone information is an essential source for correctly understanding the meaning of the sentence. By introducing the tone classification information into the speech recognition system, not only the complication inducted to the language processing part can be reduced, but also some kinds of illegal errors produced by the acoustic model part can be corrected. So we believe that tone recognition will play an important role in the final Chinese speech recognition system, especially in its

speech understanding part. This is also the motivation of our being interested in tone classification for Chinese continuous speech.

In this section, we suggest a new way, which can be easily combined with most popular recognition systems nowadays, i.e. HMM-based ASR system, to provide further tone information besides the common recognition result of the phone-like unit (PLU).

Chinese syllable is made up of initial part and final part. Almost all Chinese speech recognition systems are initial-final structural based system. Tone information only exists in the final part (voiced part). Therefore, it is not necessary to extract pitch period for whole sentence, only those voiced parts of the sentence need to be extracted, and then classified to the different tones.

Generally, when the HMM-based recognition system finishes its acoustic model matching, it can also provide a segmentation path of the phone-like units for one sentence. As for Chinese speech recognition system, this path usually gives the initial-final segmentations. Upon this segmentation path, the pitch extraction process can be applied only on those voiced units to get their pitch contours. Here, for example, pitch extraction is based on the de-emphasized LPC residual signal, which can be easily obtained from the feature vectors if LPC-CEP feature is used by the recognition system. Once the pitch sequence of the final part is obtained, GMTM is introduced into the tone classification procedure as described in section II. Thus, the initial part recognition result, final part recognition result and its corresponding tone category can be available at the same time. All the information can be utilized in the language processing part of the continuous speech recognition system. Since the tone classification result is given in probability, it can be easily combined with the PLU recognition result in the language model. This will benefit for the speech understanding process in ASR system.

IV. EXPERIMENTS

Experiments have been done based on the Chinese continuous speech database of national "863" project (data of male speaker "M00" were used). The speech recognition system is a baseline system for Chinese continuous speech recognition in speaker-independent mode. The top-one PLU accuracy of this system is above 60%, with almost 10% insertion errors and 3% deletion errors. Since any incompleteness of PLU segmentation will make wrong pitch sequence or even fail to extract the pitch of this PLU, the amount of insertion errors and deletion errors directly reflects the effect on the time-domain segmentation by recognition system, and further impacts the final result of tone classification. Pitch extraction method used here is called Area Difference Function method, and it will bring almost 8% pitch extraction error into the final result [12].

Tone Duration Probability

According to Chinese phonetics, the duration of different tones may have little difference. To improve the accuracy of tone classification, tone duration probability is introduced as follow:

$$\hat{p}(X \mid \lambda) = p(X \mid \lambda)\left[N(d_t, \mu, \sigma^2)\right]^{\gamma_D}$$

where N is a normal density with mean μ and variance σ^2, and γ_D is a weighting factor on tone duration. d_t is the segmentation duration where the pitch vector X extracted from.

Figure I Classification Result of Each Tone

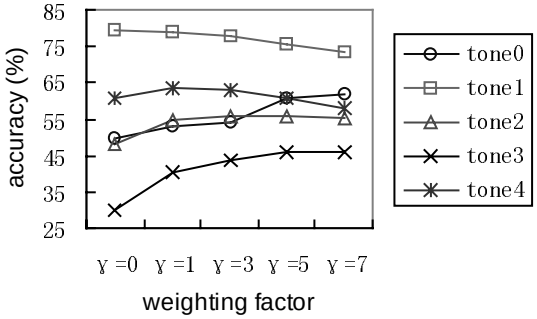

Figure II Overall Classification Result

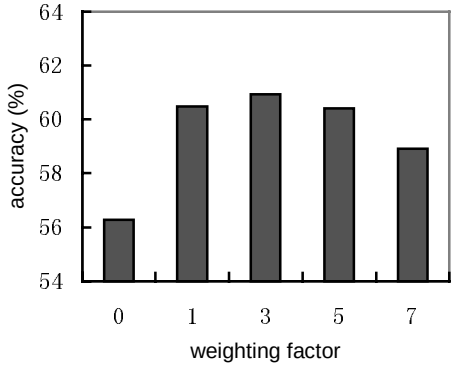

Figure I shows that tone duration probability has the different influence on different tones. Tone 0, i.e. “zero-tone” in continuous speech, and tone 3 have the distinct improvement while other three tones have no regular rule in their result. Figure II shows that introducing of the tone duration makes the overall tone classification accuracy improve about 4% on the best situation while Υ_D equals to 3. The result also shows that tone 3 has the lowest accuracy while tone 1 owns the highest. This result just accord with the tone sandhi rule in Chinese phonetics. Tone 3 has the most complicated rules of inflection in Chinese language [13], especially in continuous speech, whereas tone 1 has a relatively little number of modified tones. So the classification accuracy curves of tone 1 and tone 3 have a distinct distance with other three tones in figure I.

V. CONCLUSION

This paper provides a new attempt on how to obtain the tone classification information for a tonal language and a way of combining it with the popular speech recognition system. Although the top-one tone recognition result is just above 60% in our baseline system now, it shows the advantage of feasibility to be integrated with ASR system. And not only can it be applied in Chinese language, but also it can be used as a common tone recognition method for any tonal language, since most tonal languages in the world have the fix number of tone modes for themselves. Further study is worthy of being done on this topic. We believe that there will be various prospective usage of tone information that can be applied into the continuous speech recognition system for tonal language in the near future.

REFERENCES

[1] Xi-Xian Chen, Chang-Nian Cai, Peng Guo and Ying Sun, "A Hidden Markov Model Applied to Chinese Four-Tone Recognition", ICASSP 1987, pp. 797-800.

[2] Wu-Ji Yang, Jyh-Chyang Lee, Yuen-Chin Chang and Hsiao-Chuan Wang, "Hidden Markov Model for Mandarin Lexical Tone Recognition", IEEE Trans. on ASSP, Vol. 36, No.7, July 1988, pp. 988-992.

[3] Zhiwei Lin, DongXin and Taiyi Huang, "A Neural Net approach for Chinese four tone recognition", International Conference on Computer Processing of Chinese and Oriental Languages, April 1990, Changsha, China.

[4] Pao-Chung Chang, San-Wei Sun and Sin-Horng Chen, "Mandarin Tone Recognition by Multi-Layer Perceptron", ICASSP-90, pp. 517-520.

[5] Shilin Xu and Samuel C. Lee, "A fast real time Chinese tone recognition system using Fuzzy Sets", An International Journal of the Chinese & Oriental Languages Information Processing Society, Vol. 2, No. 1, April 1992.

[6] A.R. Douglas, R.C. Richard, “Robust text-independent speaker identification using Gaussian mixture speaker models”, IEEE *Trans.* on ASSP, Vol. 3, No.1, pp.72-83, Jan. 1995.

[7] M.A. Zissman, “Automatic language identification using Gaussian mixture and hidden Markov models”, *Proc*. ICASSP ’93, Vol. 2, pp.399-402, Apr. 1993.

[8] G. McLachlan, “Mixture Models”, New York: Marcel Dekker, 1988.

[9] A. Dempster, *et al.*, “Maximum likelihood from incomplete data via the EM algorithm”, *J.Royal Stat. Soc.*, Vol. 39, pp.1-38, 1977.

[10] R. Hathaway, “A constrained formulation of maximum-likelihood estimation for normal mixture distributions”, *Ann. Stat.*, Vol. 13, No. 2, pp.795-800, 1985.

[11] Yoseph Linde, Andres Buzo, Robert M. Gray, “An algorithm for vector quantizer design”, IEEE Trans on Communications, Vol. COM-28, No.1, January, 1980.

[12] Liu Jian, Yu Tiecheng, “A Novel Pitch Extraction Method and Its Application to Automatic Classification of Chinese Tones” In Proceedings of the Sixth Western Pacific Regional Acoustics Conference *(WESTPRAC VI 97)*, Nov. 19, 1997, HONG KONG, Vol. I, pp. 259-264.

[13] Wang, William S.Y.,. Li, K.P., “Tone 3 in Pekinese”, Journal of speech and hearing research, Vol. 10, pp. 629-636, 1967.