Send Orders for Reprints to reprints@benthamscience.ae

A New Algorithm for Noisy Speech Classification Based on GMM

Zhongbao Chen, Zhigang Fang, Jie Xu*, Pengying Du and Xiaoping Luo

Zhejiang University City College, Hangzhou, 310015, Zhejiang Province, China

Abstract: Speech can be broadly categorized into voiceless, voiced, and mute signal, in which voiced speech can be further classified into vowel and voiced consonant. With the ever increasing demand of the speech synthesis applications, it is urgent to develop an effective classification method to differentiate vowel and voiced consonant signal since they are two distinct components that affect the naturalness of the synthetic speech signal. State-of-the-arts algorithms for speech signal classification are effective in classifying voiceless, voiced and mute speech signal, however, not effective in further classifying the voiced signal. In view of the issue, a new algorithm for speech classification based on Gaussian Mixture Model (GMM) is proposed, which can directly classify a speech into voiceless, voiced consonant, vowel and mute signal. Simulation results demonstrate that the proposed algorithm is effective even under the noisy environments.

Keywords: Noisy speech signal, energy distribution, Gaussian mixture model.

1. INTRODUCTION

It is an important part of the speech signal processing to divide the speech signal into voiceless, voiced consonant, vowel, mute signal. For example, according to the different characteristics of these kinds of signals, we can use different processing schemes to improve the efficiency and quality of coding algorithm in speech coding. However, due to the dynamic range of many feature parameters of these signals are usually overlapped (for example, short-time energy parameter and zero crossing rate parameter), we can't separate it linear by extracting a feature parameter, what is worse, these parameters are more difficult to distinguish in the disturbed situations. The traditional method is extracting some feature parameters, and then to judge to carry on linear processing and the predetermined threshold. The threshold is usually determined by artificial experience, the method is simple, easy to implement, but can't guarantee the judgment result reliable and accurate. These algorithms have not voiced further classification, the representative is the V/U/S mode classification method[1] based on multi feature parameters which was proposed by Atal and Rabiner, its classification techniques is a Bayesian decision process. Then some improvement and new classification methods were proposed [2, 3], and there were many scholars proposed some classification methods based on different feature parameters and neural network structure [4, 5]. But the traditional artificial neural network methods (such as BP network) have the defects of slow training speed, easy to fall into local minima and poor generalization performance in network training and network design, and those nonlinear methods still need to rely on the user's engineering experience in the selection of network structure and weights setting, which lack of a unified mathematics theoretical foundation. With the emergence of a new pattern recognition method (support vector machine), some scholars put it into the speech classification work [6], they combined two classification support vector machine to realize the voiceless, voiced and mute classification of speech signals. However, the algorithm is very complex, and there was no further classification of voiced speech signals.

Gauss mixture model is a research hotspot of pattern recognition algorithms in recent years, which is widely used in the field of speech signal processing, and achieved good result [7]. The core idea of GMM is to describe the distribution of feature vectors in the probability space with the combination of multiple Gauss distribution probability density function, GMM is indicated by the weighted sum of multiple Gauss distribution probability density function, the number of the probability density function is called the mixed numbers of Gauss model. In this paper, we use the differences of noise and speech signal's energy distribution in the spectrum, avoiding the parameters subject to noise interference, such as the short-time energy and zero crossing rate etc. By making Fourier transform on speech signal, calculating the ratio of the energy distribution on different frequency bands as the feature vector, the algorithm establishes Gauss mixture model (GMM) to classify speech signal, and further divide voiced signal into vowel and voiced consonant. Experiments show that, this method can realize the classification accurately in the environment of low signal-to-noise.

2. ALGORITHM DESCRIPTION

Fig. (**1**) shows the theory of speech classification system based on GMM.

1874-1290/14

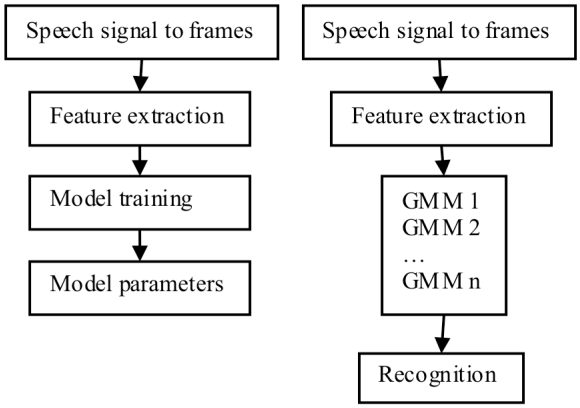

Fig. (1). Theory of Speech Classification System Based on GMM.

Table 1. Speech Spectrum Energy Distribution.

Speech Classification		Characteristics of the Spectral Energy Distribution
voi-ced	vowel	low frequency(0.1－0.4KHz)high energy intermediate frequency(0.64－4KHz)high energy
	voiced conso-nant	low frequency(0.1－0.4KHz)high energy intermediate frequency(0.64－2.8KHz)high energy
voiceless		high frequency(3.5KHz or above)high energy
noise		distribute in each frequency band

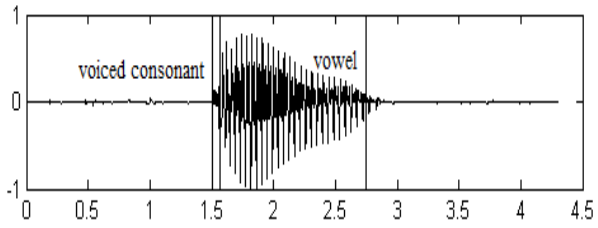

Fig. (2). The Energy Comparison for Voiced Consonant and Vowel.

2.1. Extraction Of Speech Feature Parameters

According to the different forms of excitation, the speech can be divided into two categories: voiced and voiceless, and voiced is again divided into vowel and voiced consonant. Table **1** shows the spectrum energy distribution of speech.

As it can be seen from Table **1**, there have a certain distribution law in various frequency bands of different types of speech signal, so we can classify the speech signal by the law. The feature vector of GMM is the percentage of each frequency band's energy. The speech signal is divided into frames, and each doing FFT transform, according to Table **1**, divided into 4 frequency bands(0-1KHz, 1-3Hz, 3-5KHz, 5-8KHz), and calculating the percentage of energy in the frequency band. The spectral energy distribution of vowel and voiced consonant is approximate, but the energy difference is large, as shown in Fig. (**2**). To distinguish between the two, we adding the normalized energy, composed of 5 dimensional feature vector.

2.2. Establishment of GMM

The distribution of speech signal feature vector samples in the probability space can be said by GMM, a d dimensional vector GMM which has a M mixed number can be shown by:

$$p\left(x\middle|\lambda\right)=\sum_{i=1}^{M}\omega_i p_i\left(x\middle|\mu_i\Sigma_i\right) \quad (1)$$

$$p_i\left(x\middle|\mu_i\ \Sigma_i\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{\frac{-1}{2}\left(x-\mu_i\right)^{T}\Sigma_i^{-1}\left(x-\mu_i\right)\right\} \quad (2)$$

Where:

x is d dimensional feature vector;

ω_i is the mixing weights, and meet $\sum_{i=1}^{M}\omega_i = 1$;

$p_i(x \mid \mu_i \Sigma_i)$ is d dimensional Gauss function, represents the i Gauss component of the GMM; μ_i is the mean vector of Gauss component; $\sum_i$ is the covariance matrix of Gauss component.

The Gauss mixture model (GMM) is described by the mean vector, the covariance matrix and mixed weight of each mixture component, use λ to describe the model:

$$\lambda = (\omega_i \mu_i \Sigma_i) \quad i = 1, 2 \ldots M. \tag{3}$$

2.3. GMM Model Parameters

The method to train GMM model parameters is using Maximum Likelihood estimation algorithm. The purpose of training is to find a set of model parameters λ to get the maximum value of $P(\lambda|X)$.

That is:

$$\hat{\lambda} = \arg\max_{\lambda} L(\lambda|X) \tag{4}$$

Maximum Likelihood estimation usually uses the iterative calculation of Expectation Maximization (EM) algorithm. The algorithm has two main steps: expect step (E) and maximize step (M).E step use current set of parameters to calculate the exception value of complete data likelihood function. M step get the new parameters by maximizing expected function and E step and M step will iterate until convergence.

We define the Q function to illustrate the EM algorithm:

$$Q(\lambda, \lambda') = \sum_{i=1}^{M} \log P(X, i|\lambda)\left[P(X, i|\lambda')\right] \tag{5}$$

i is the serial number of Gaussian component, λ' is the existing model parameters and λ is the new parameters to be calculated.

After defines the Q function, the EM algorithm can be simply described as follows:

(1) E step: calculate the probability $p(i|x_i, \lambda')$ of training data on $i_t = i$ state:

$$p(i_t = i|x_t, \lambda) = \frac{\omega_i' p_i(x_t \ \mu_i' \ \Sigma_i')}{P(x_i|\lambda)} = \frac{\omega_i' p_i(x_t \ \mu_i' \ \Sigma_i')}{\sum_{i=1}^{M}\omega_i' p_i(x_t \ \mu_i' \ \Sigma_i')} \tag{6}$$

(2)M step: Maximize Q function, get $Q(\lambda \lambda')$ partial derivative of 0 relative to $\omega_i, \mu_i, \Sigma_i$ $i = 1, 2, \ldots M$ value:

$$\omega_i = \frac{1}{T}\sum_{t=1}^{T} p(i|x_t \ \lambda') \tag{7}$$

$$\mu_i = \frac{\sum_{t=1}^{T} p(i|x_t \ \lambda')x_t}{\sum_{t=1}^{T} p(i|x_t \ \lambda')} \tag{8}$$

$$\Sigma_i = \frac{\sum_{t=1}^{T} p(i|x_t \ \lambda')(x_t - \mu_i)(x_t - \mu_i)^T}{\sum_{t=1}^{T} p(i|x_t \ \lambda')} \tag{9}$$

Considering the characteristic vector sample of speech is always limited, in the process of training, some component of the covariance matrix of the GMM may be quite small and these model parameters have a large influence on likelihood function which may seriously affect the performance of the system. So in the iterative calculation, we usually set a threshold value of the covariance, that is if covariance value is less than set threshold value, we will use the threshold value instead.

2.4. GMM Recognition Algorithm

Our purpose is to find a kind of speech whose corresponding model λ_i makes the characteristic vector group X have the maximum posteriori probability $P(\lambda_i|X)$ in recognition.

According to the Bayes theory, the maximum posteriori probability is expressed as:

$$\hat{S} = \arg\max_{1 \le k \le M} P(\lambda_k|X) = \arg\max_{1 \le k \le M} \frac{P(X|\lambda_k)P(\lambda_i)}{P(X)} \tag{10}$$

Assuming that the prior probability of each voice is equal, that is $P(\lambda_i) = 1/S$ and to each speech, $P(X)$ is the same, the expression can be simplified as:

$$\hat{S} = \arg\max_{1 \le k \le M} P(X|\lambda_k) \tag{11}$$

3. SIMULATION AND RESULTS

Quiet environment speech signal is collected as the experiment data including 10 people and the male to female ratio is 1:1.Everyone speak for one minute and speech content contains all Chinese phonemes. Signal sampling frequency is 8000HZ and quantitative level is 16 bit. Extracting voiceless, voiced consonant, vowel, mute each 1000 characteristic vectors used in the experiment. The experimental steps are as follows:

1. Frame processing speech signal, each frame has 256 sampling points. Extract voiceless, voiced consonant, vowel, mute each 900 frames. Calculate the percentage of each frequency band energy and normalized energy forming a feature vector.

2. Determine the initial value of λ. This paper adopts the method of clustering classifying feature vector according to nearest neighbor, and determining the covariance and mean

Table 2. Recognition Accuracy Under Different SNR.

	SNR60dB	SNR50dB	SNR45dB	SNR40dB
vowel	100%	100%	100%	98%
voiced consonant	100%	98%	89%	83%
voiceless	100%	96%	82%	75%
mute	100%	100%	100%	100%

of each class respectively as the initial matrix and average value. Weights are the percentages of the each classes' feature vectors number in the total feature vectors.

3. Take voiced consonant, vowel, voiceless, mute each 900 feature vectors for training respectively, to determine the corresponding GMM to each voice type, and using EM algorithm to iterate to determine the coefficients of each GMM.

4. Take voiced consonant, vowel, voiceless, mute each 100 feature vectors, adding Gauss noise with different SNR, and sending to each model respectively. The model whose has the largest output posterior probability is the speech type.

Table **2** shows the experiments results.

CONCLUSION

The Gauss mixture model based on speech signal energy distribution has a high accuracy in a high SNR condition, but it is easy to confuse the voiceless, voiced consonants at low SNR, because the two normalized energy is very close, and the noise energy in each frequency band has a uniform distribution, its superposition makes the frequency band energy distribution boundary fuzzy, which causes miscarriage of justice. This problem will be improved in the future research work.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

Our Research Project was fully sponsored by Zhejiang Nature Science Foundation (LY12F03018). Thanks for the sponsorships.

REFERENCES

[1] Atal B., Rabiner L. "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition". *IEEE Transactions on A coustics, Speech and Signal Processing*, vol.24(3), pp.201-212,1976.

[2] Rabiner L. R., Sambur M. R. "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25(4), pp.338-343,1977.

[3] Cox B., Timothy L. M. "Nonparametric rank-order statistics applied to robust voiced-unvoiced-silence classifica2tion". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28(5), pp.550-561,1980

[4] Qi Y. Y., Hunt B. R. "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier". *IEEE Transactions on Speech and Audio Processing*, vol.1(2), pp.250-255,1993

[5] Ahn R., Holmes W. H. "Voiced / unvoiced / silence classification of speech using 2-Stage neural networks with delayed decision input. B Boashash, et a l". *Proc ISSPA96. B risbane, Australia: Queensland University of Technology*, pp.389-390,1996.

[6] Qi F. Y., Bao C. C. "New method based on support vector machine(SVM) to classify speech signals with noise into voiced consonant", *Vowel and Mute*, vol. 4(4), pp.605-611,2006.

[7] Reynolds D. A., Rose R. C. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE Transactions on Speech and Audio Processing January*, vol.3(1), pp.72-83,1995.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014