Speaker Recognition Using Weighted Dynamic MFCC Based on GMM

Zufeng Weng，Lin Li*，Donghui Guo
School of Information Science and Technology, Xiamen University, Xiamen, Fujian, China
Corresponding author: lilin@xmu.edu.cn

***Abstract*—In this paper, a new algorithm of feature parameter extraction is proposed for application in speaker recognition system, which combines the traditional MFCC and the dynamic MFCC as a new series of coefficients. According to the statistics analysis of the different contribution by the dynamic MFCC and traditional MFCC, these coefficients are weighted as front-end parameters of the GMM, which would decrease the dimension of the mixed weighted GMM and reduce the computation complexity. The experiments based on the TIMIT and VOA speech database were implemented in MATLAB environment, and the results showed the speaker recognition system with the Weighted Dynamic MFCC could obtain better performance with high recognition rate and low computational complexity.**

Keywords - Speaker Recognition; Dynamic MFCC; GMM

I. INTRODUCTION

Text-independent speaker recognition doesn't care about the semantic content of the speech signal, but identifies the speaker with individual characteristics extracted from the speech signal [1]. Therefore, many researchers focus on the extraction of parameters which can effectively represent the characteristics of the speaker's voiceprint feature. Moreover, different feature parameters depict different physical and acoustic characteristics, and influence the performance of speaker recognition system.

Nowadays there are three main types of parameter features widely used, including LPC (Linear Predictive Code) [2], LPCC (Linear Predictive Cepstrum Coefficient) [3] and MFCC (Mel-Frequency Cepstrum Coefficient) [4]. However, both of LPC and LPCC approach the speech signal linearly in the whole frequency scale, which couldn't characterize the acoustic specialization. The feature parameters used in this paper are based on MFCC, which can simulate characteristics of the channel spectrum and the human ear's hearing [5]. Recently, many researches [6-8] make further improvement on the feature extraction of speaker recognition system. Hassan [6] employed a model that using pitch and MFCC as parameters for the classification task, in order to improve the performance by 8% for all speakers. Wang [7] proposed a new speaker recognition algorithm based on the dynamic MFCC parameters, which could obtain higher recognition rate than the one using traditional MFCC. Wang [8] introduced a differential MFCC algorithm for Real-Time Speaker Recognition System and then employed VQ (Vector Quantification) [9] to classify the voice features. In this paper, the concept of the Weighted Dynamic MFCC is proposed, and a speaker recognition system is realized by utilizing this new series of coefficients as the parameters of the classic GMM [10]. The Weighted Dynamic MFCC could characterize the speaker's voiceprint features and the dynamic characteristics of the speech. As a result, by using these feature parameters, the computational complexity of the system is significantly reduced while maintaining high recognition rate.

In order to present our new algorithm in detail, the optimization of the algorithm is introduced in section II, and section III is to present a recognition system based on GMM implemented with exploiting the Weighted Dynamic MFCC. Then, the experiments and analyses are showed in section IV. Finally, we give a conclusion in section V.

II. OPTIMIZATION OF THE PROPOSED WEIGHTED DYNAMIC MFCC

An MFCC process converts linear spectrum into nonlinear Mel-spectrum. The corresponding relationship between the linear-scale frequency f and the Mel-scale frequency f_{mel} is shown below:

$$f_{mel} = 2595 \log_{10}(1 + f / 700) \quad (1)$$

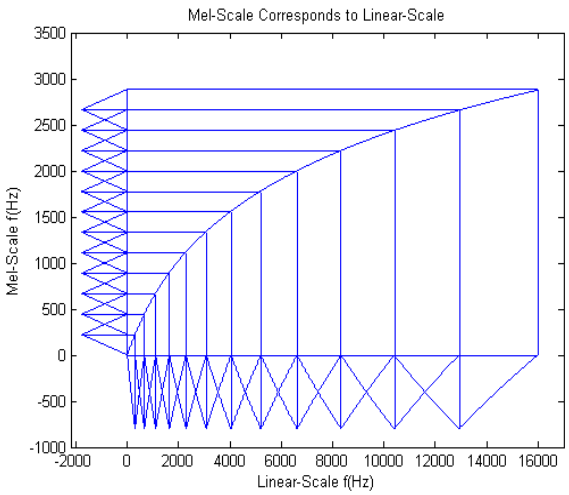

Figure 1. Mel-scale corresponds to Linear-scale

As we can see from figure 1, the mel-filter banks divide the vertical coordinates into several equal parts, while its linear-scale isn't distributed equidistantly in the frequency

978-1-4244-6734-1/10/$26.00 ©2010 IEEE

domain. The flow chart of calculating the traditional MFCC is shown in Figure 2.

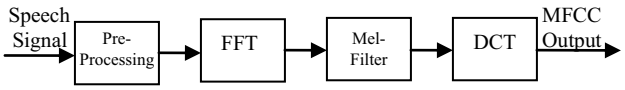

Figure 2. The process of calculating MFCC

The traditional MFCC only represent the speaker's voiceprint feature, but doesn't consider the dynamic characteristics of the speech. In order to improve the performance, the popular method is to combine the traditional MFCC and its first-order differential coefficients [7][8][11], which can effectively reflect the dynamic characteristics of the speech. However, this approach increases both the dimension of the parameters and the computational complexity of the speaker recognition system.

In this paper, an optimization algorithm is proposed using the Weighted Dynamic MFCC. Assuming the new Weighted Dynamic MFCC as $newMFCC$, the equation is shown as following:

$$newMFCC = MFCC + a \bullet \Delta MFCC + b \bullet \Delta^2 MFCC \tag{2}$$

where $\Delta MFCC$ is the first-order differential coefficient, $\Delta^2 MFCC$ is the second-order differential coefficient, with a and b as their weights, respectively. Considering their different contribution to the speech feature parameters, the constraint condition should be set as: $b<a<1$. In the Eq. (2) of this new coefficient $newMFCC$, the element $MFCC$ depicts the voice channel characteristics, $\Delta MFCC$ reflects the dynamic characteristics, and $\Delta^2 MFCC$ is imported as its balance factor. Let $M = MFCC$, $\Delta M = \Delta MFCC$, $\Delta^2 M = \Delta^2 MFCC$, we have:

$$\Delta M_{i,n} = \sum_{k=-2}^{2} k M_{i-k,n} \tag{3}$$

$$\Delta^2 M_{i,n} = \sum_{k=-2}^{2} k \Delta M_{i-k,n} \tag{4}$$

where i=3,4,5…T-2 is the frame of the feature parameters and n=1,2,3…N is the dimension of the feature parameters.

Compared to the popular method [7][8][11], this new algorithm reduces the dimension of the coefficient matrix in great amount. For example, the size of the coefficient matrix in Wang [8] is $T \times (2 \times N)$, and the $r-th$ row data is shown in Eq. (5) as $Mat1_r$. However the dimension of the coefficient matrix in the proposed algorithm could be decreased by $T \times N$, the $r-th$ row data ($Mat2_r$) of which is expressed in Eq. (6).

$$Mat1_r = \begin{bmatrix} M_{r,1} \\ \vdots \\ M_{r,N} \\ \Delta M_{r,1} \\ \vdots \\ \Delta M_{r,N} \end{bmatrix}^T \tag{5}$$

$$Mat2_r = \begin{bmatrix} M_{r,1} + a \bullet \Delta M_{r,1} + b \bullet \Delta^2 M_{r,1} \\ M_{r,2} + a \bullet \Delta M_{r,2} + b \bullet \Delta^2 M_{r,2} \\ \vdots \\ M_{r,N} + a \bullet \Delta M_{r,N} + b \bullet \Delta^2 M_{r,N} \end{bmatrix}^T \tag{6}$$

As shown in Figure 3, the Weighted Dynamic MFCC and the traditional MFCC have the same dimension and similar amplitude curve. The feature parameters with the concept of the Weighted Dynamic MFCC would have better performance on reflecting both the voiceprint and the dynamic characteristics of the speech.

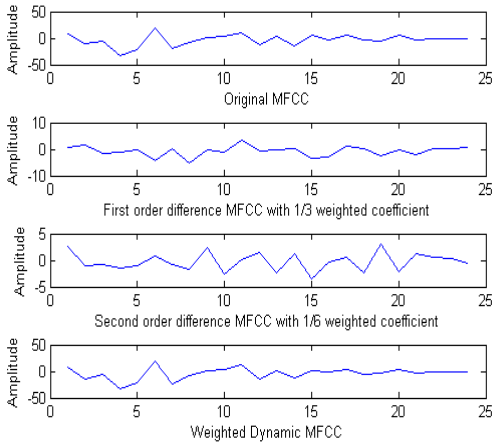

Figure 3. The Weighted Dynamic MFCC of single-frame speech signal

III. IMPLEMENTATION BASED ON GMM

In the process of training GMM speaker model, we use LBG [12] and KMeans [13] of VQ (Vector Quantification) [9] algorithm to get the stable cluster center as the training parameters of the GMM. With these cluster centers, the $newMFCC$ got from the previous section could be converged into a stable N-dimensional mean vector μ, where N is the dimension of the Weighted Dynamic MFCC.

The principle of GMM is to abstract a random process from the speech, then to establish a probability model for each speaker. It is relatively independent between the various probability models. Assuming the variable M in the M-order GMM probability density function is the number of Gaussian probability density functions. And set X as the Weighted Dynamic MFCC of the speech:

$$P(X/\lambda)=\sum_{i=1}^{M}\omega_i b_i(X) \tag{7}$$

where $b_i(X)$ is the sub-distribution and is given by the following equation:

$$b_i(X)=\frac{1}{(2\pi)^{N/2}\bullet\left|\sum_i\right|^{1/2}}\bullet\exp(-\frac{1}{2}(X-\mu_i)^T\bullet\sum_i^{-1}(X-\mu_i)) \tag{8}$$

where μ_i is the mean vector，$\sum_i$ is the full covariance matrix, $\sum_i^{-1}$ is the inverse matrix of $\sum_i$, $|\sum_i|$ is the determinant of $\sum_i$. The mixed weights ω_i should satisfy the following condition：

$$\sum_{i=1}^{M}\omega_i=1 \tag{9}$$

The full GMM is constituted of the mean vectors, the covariance matrix and the mixed weights, which could be expressed as: $\lambda=\{\omega_i,\mu_i,\sum_i\},i=1,2,...M$. For a given time series $X=\{X_t\},t=1,2,...,T$, where T is the total number of frames of the Weighted Dynamic MFCC. The logarithm likelihood obtained from the GMM can be defined as the following equation:

$$L(X/\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log P(X_t/\lambda) \tag{10}$$

In Eq. (8), both of $\sum_i^{-1}$ and $|\sum_i|$ require a large amount of computation. Besides, $\sum_i$ would be singular if the rank is not equal to the dimension, which may cause obstacle to calculate $\sum_i^{-1}$ and $|\sum_i|$, so we use diagonal covariance matrix instead. Set the diagonal covariance matrix as $Sigma_i$, which can be defined as:

$$Sigma_i=\begin{bmatrix}\sigma_1 & 0 & 0\ldots 0\\ 0 & \sigma_2 & 0\ldots 0\\ & \cdots & \\ 0 & 0 & \ldots\ \sigma_N\end{bmatrix} \tag{11}$$

where $\sigma_1,\sigma_2\ldots\sigma_N$ are the main diagonal elements of $\sum_i,i=1,2,...,M$

Assuming we have S speakers in a closed-set which is different from each other. For a given speech feature vector $\{X_t\},t=1,2,...,T$, the purpose of speaker recognition is to find the speaker k in the closed-set $k\in\{1,2,...,S\}$, whose corresponding model λ_k will obtain the largest posterior probability $P(\lambda_k/X)$. The general structure based on GMM for speaker recognition system can be shown as follows, where GMM 1, GMM 2, …, GMM S represent the GMM models of Speaker 1, Speaker 2, …, Speaker S, respectively.

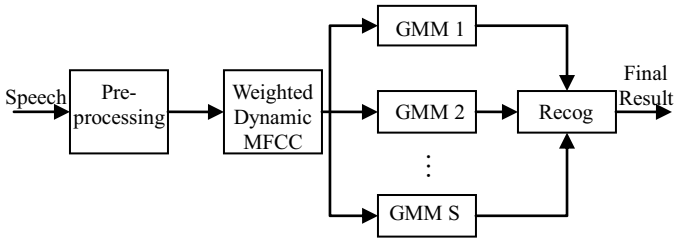

Figure 4. Speaker recognition system structure based on GMM

IV. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

A text-independent speaker recognition system based on GMM is realized using the Weighted Dynamic MFCC in MATLAB environment. The speech data used in this paper consist of TIMIT Speaker Speech Database and VOA news report data including 201 different speakers, which is recorded in a quiet environment in American English with 16000Hz sample frequency. In order to test the flexibility of the Weighted Dynamic MFCC, three different speech databases (named SPEECH1, SPEECH2 and SPEECH3, respectively) are made of: (1).201 speakers (168 in TIMIT, 33 in VOA); (2).168 speakers (all in TIMIT); (3).89 speakers (33 in VOA, 56 in TIMIT with women speech).

The following three tables show performance (recognition rate and time-consuming) under the condition of different Gaussian Mixed Weights and different feature parameters for each speech database (SPEECH1, SPEECH2 or SPEECH3).

TABLE I. ANALYSIS OF PERFORMANCE FOR SPEECH1(201 SPEAKERS)

Gaussian Mixed Weights	Parameters	Dimensions	Time(s)	Recognition Rate (%)
2	MFCC	24	0.81	73.2
	MFCC+ Δ MFCC	48	1.27	74.7
	Weighted Dynamic MFCC	24	0.83	76.5
4	MFCC	24	1.16	83.1
	MFCC+ Δ MFCC	48	2.02	85.6
	Weighted Dynamic MFCC	24	1.22	86.2
8	MFCC	24	1.87	91.6
	MFCC+ Δ MFCC	48	3.47	93.1
	Weighted Dynamic MFCC	24	1.98	94.6

TABLE II. ANALYSIS OF PERFORMANCE FOR SPEECH2(168 SPEAKERS)

Gaussian Mixed Weights	Parameters	Dimensions	Time(s)	Recognition Rate (%)
2	MFCC	24	0.84	72.1
	MFCC+ ∆ MFCC	48	1.26	73.6
	Weighted Dynamic MFCC	24	0.86	74.5
4	MFCC	24	1.13	81.6
	MFCC+ ∆ MFCC	48	1.98	82.1
	Weighted Dynamic MFCC	24	1.15	82.8
8	MFCC	24	1.73	91.1
	MFCC+ ∆ MFCC	48	3.41	91.7
	Weighted Dynamic MFCC	24	1.77	92.8

TABLE III. ANALYSIS OF PERFORMANCE FOR SPEECH3(89 SPEAKERS)

Gaussian Mixed Weights	Parameters	Dimensions	Time(s)	Recognition Rate (%)
2	MFCC	24	0.58	76.4
	MFCC+ ∆ MFCC	48	0.89	78.7
	Weighted Dynamic MFCC	24	0.62	79.9
4	MFCC	24	0.75	88.7
	MFCC+ ∆ MFCC	48	1.44	88.7
	Weighted Dynamic MFCC	24	0.82	91.5
8	MFCC	24	0.97	92.2
	MFCC+ ∆ MFCC	48	1.62	94.4
	Weighted Dynamic MFCC	24	1.04	95.5

The experimental results in the tables above show that for each speech database, the recognition rate is improved by increasing the number of Gaussian Mixed Weights. In addition, for each GMM Mixed Weight (2, 4 or 8), the system using the Weighted Dynamic MFCC obtains higher recognition rate than the one employing the traditional MFCC by 3.3%, while 1.5% higher than the system with MFCC+△ MFCC. Analyzing the computational complexity, the proposed system needs only 4.8% more time-consuming than the one exploiting traditional MFCC, but saves nearly 39% compared with the sytem using MFCC+△MFCC.

V. CONCLUSION

In this paper, a new algorithm using Weighted Dynamic MFCC is proposed to extract the feature parameters of the speakers for the speaker recognition system based on GMM, which considers both the voiceprint and the dynamic characteristics of the speech. The experimental results show that the recognition system with the Weighted Dynamic MFCC could achieve higher recognition rate than both the systems with traditional MFCC and MFCC+ △ MFCC. Moreover, the proposed system using the Weighted Dynamic MFCC would significantly reduce the computational complexity compared to the one using the MFCC+△MFCC.

REFERENCES

[1] A.Revathi and Y.Venkataramani, “Source and system features for text independent speaker identification using iterative clustering approach”, IEEE International Conference on Signal and Image Processing Applications 2009, pp.1-5, Nov. 2009.

[2] M.Bouzid, “Robust quantization of LPC parameters for speech communication over noisy channel”, Second International Conference on the Applications of Digital Information and Web Technologies 2009, pp.713-718, Aug. 2009.

[3] X.Y.Zhang, Y.L.Guo and X.M.Hou, “A Speech Recognition Method of Isolated Words Based on Modified LPC Cepstrum”, IEEE International Conference on Granular Computing 2007, pp.481-485, Nov. 2007.

[4] D.Hosseinzadeh and S.Krishnan, “Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs”, IEEE 9th Workshop on Multimedia Signal Processing 2007, pp.365-368, Oct. 2007.

[5] Z.J.Wu and Z.G.Cao, “Improved MFCC-Based Feature for Robust Speaker Identification”, TSINGHUA Science and Technology, vol.10, pp. 158-161, Apr. 2005.

[6] H.Ezzaidi and J.Rouat “Pitch and MFCC dependent GMM models for speaker identification systems”, Canadian Conference on Electrical and Computer Engineering 2004, vol.1, pp.43 – 46, May. 2004.

[7] Y.TWang, B.Li, X.Q.Jiang, F.Liu, and L.H.Wang, “Speaker Recognition Based on Dynamic MFCC Parameters”, International Conference on Image Analysis and Signal Processing 2009, pp.406-409, April,2009.

[8] C.Wang, Z.J.Miao, and X.Meng, “Differential MFCC and Vector Quantization used for Real-Time Speaker Recognition System”, Congress on Image and Signal Processing 2008, vol.5, pp.319-323, May 2008.

[9] K. Mori and S. Nakagawa, “Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition”, IEEE International Conference on Acoustics, Speech, and Signal Processing 2001, vol.1, pp. 413-416, 2001.

[10] D.A.Reynolds, “Speaker Identification and Verification Using Gaussian Mixture Speaker Models”, Speech Communication, vol.17, pp.91-108, August 1995.

[11] Y.P.Lai, M.H.Siu and B.Mak, “Joint Optimization of the Frequency Domain and Time-Domain Transformations in Deriving Generalized Static and Dynamic MFCCs”, IEEE Signal Processing Letters, vol.13, pp.707-710, Nov.2006.

[12] Y.Linde, A.Buzo, and R.M.Gray, “An algorithm for vector quantizer design”, IEEE Transaction on Communications, vol.28, pp.84-95, 1980.

[13] O.Ben-Harush, I.Lapidot, and H.Guterman, “Segmental K-Means initialization for SOM-based speaker clustering”, 50th International Symposium on ELMAR 2008, vol.1, pp.305-308, 2008