ICIC Express Letters
Part B: Applications
Volume **7**, Number **2**, February **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **249–255**

PARAMETER OPTIMIZATION FOR GAUSSIAN MIXTURE MODEL AND ITS APPLICATION IN SPEAKER IDENTIFICATION

Yuecheng Peng[1] and Xinyuan Huang[2]

[1]School of Arts and Design
Beijing Forestry University
No. 35, Tsinghua East Road, Haidian District, Beijing 100083, P. R. China
pengyuecheng@bjfu.edu.cn

[2]School of Animation and Digital Arts
Communication University of China
No. 1, Dingfuzhuang East Street, Chaoyang District, Beijing 100024, P. R. China

Received July 2015; accepted October 2015

Abstract. *In this paper we study the joint optimization of Gaussian mixture model parameters in speaker identification. First, we introduce the baseline speaker identification system. Second, we study the feature optimization and we simplify the transform matrix into a feature selection vector in speaker identification. Third, The joint estimation of the parameters is proposed based on the shuffled frog leaping algorithm. The traditional expectation-maximization algorithm is embedded in the proposed algorithm. The experimental results show that the proposed optimization framework is effective and achieved a constant improvement in speaker identification.*
Keywords: Speaker identification, Gaussian mixture model, Optimization algorithm

1. **Introduction.** Speaker identification (SI) is an important biometrics field [1, 2, 3]. With the development of speech signal processing and machine learning, it has made significant progresses over the past decade.

Ding and Yen [4] proposed to use Gaussian mixture model (GMM) for speaker identification, and achieved improved results over the traditional methods. Kinnunen and Li [5] studied the text-independent speaker identification problem. In their work, the speaker identification system is generalized to various text content by using super-vectors. Wu and Tsai [6] propose to use a decomposition based algorithm to improve the overall performance. However, the model is largely dependent on the empirical settings. Kinnunen et al. [7] studied the speech conversion influence on the signal features and improved the performance in telephone voice recognition.

Gaussian mixture model is one of the most promising algorithms for speaker identification. However, Gaussian mixture model based classifier is dependent on the empirical parameter setting. Therefore, the effectiveness of the parameter estimation algorithm is key to the success of GMM based speaker identification.

In this paper we propose a novel optimization algorithm for GMM based speaker identification. The overall optimization steps are demonstrated in the flowchart in Figure 1. The number of dimensions is an important factor for the feature space optimization, and the "curse of dimensionality" prevents this number from being too big on a specific recognition problem. The feature selection step in our optimization method is converted from the traditional hard decision into soft decision. A specific feature dimension is either selected or not selected in the traditional feature selection method. In our algorithm, we give each dimension a weight between zero and one indicating the importance of the

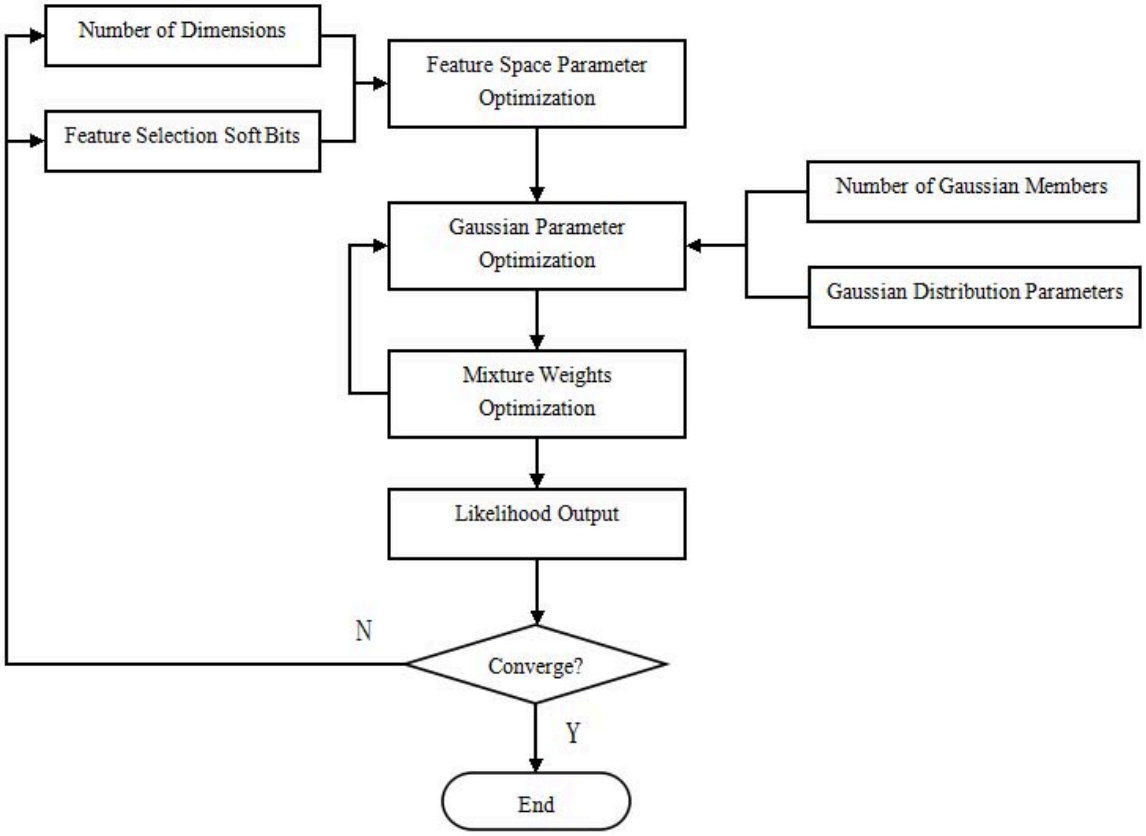

FIGURE 1. A depiction of the overall optimization flowchart

specific feature. The value one is the most important and the value zero is the least important. This is important for some practical situations when part of the features cannot be extracted successfully [8].

After the feature space optimization, the Gaussian parameters can be estimated upon it. First, we need to assume a fixed number of Gaussian members. Second, we can apply certain parameter estimation algorithm to find the optimal result. The mixture weights are generally achieved together in this process. However, the number of dimensions and the number of Gaussian members are usually set empirically in existing methods. This will cause an ill-posed foundation for the Gaussian model parameter optimization. In this paper, we use a joint optimization of all of the above parameters to find a global optimization for both feature space and recognition model.

The rest of the paper is organized as follows. Section 2 gives the general description of our speaker identification system; Section 3 provides a simplified solution of feature space optimization in our application; Section 4 describes the core optimization algorithm used in our system; Section 5 gives the detailed steps of the proposed joint optimization; Section 6 provides the detailed experimental results, and finally, conclusions are given in Section 7.

2. **The General Speaker Identification System.** In this paper, we use the Mel-frequency cepstrum coefficients (MFCC) as the speaker identification features. The MFCC features can be extracted from speech spectrum and used as a unique character of the target speaker. As shown in Figure 2, the 12-th order MFCC features are constructed in the time domain and the Mel-frequency domain. The grey scale indicates the value of a coefficient.

The speaker identification can be achieved by maximizing the likelihood of series of Gaussian mixture model outputs. As shown in Equation (1), the likelihood is computed

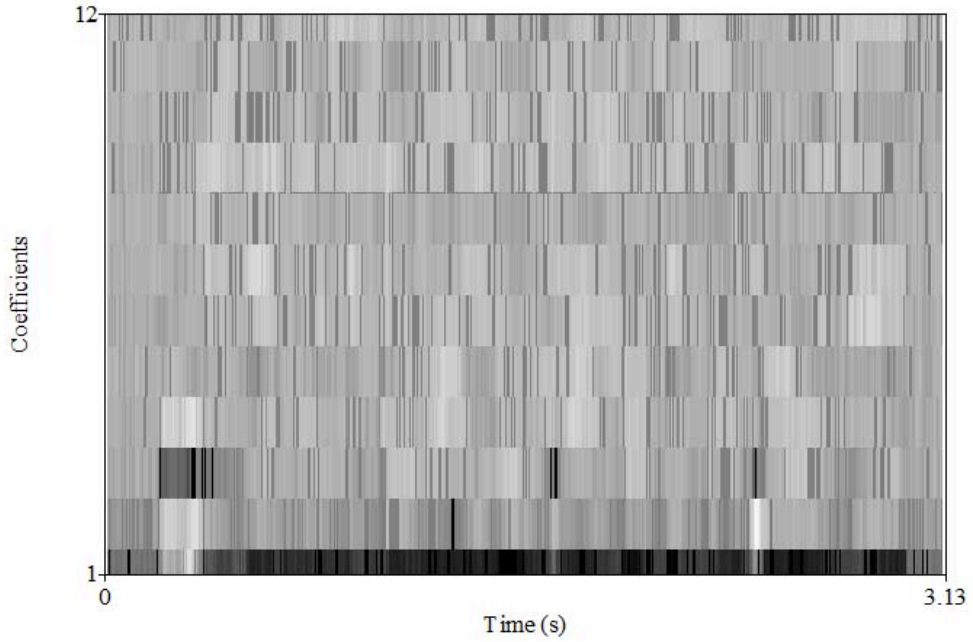

FIGURE 2. MFCC features for speaker identification system

over the entire speech utterance.

$$H(\mathbf{x}_t) = \prod_{t=0}^{T} \sum_{i=1}^{M} \omega_i * p_i(\mathbf{x}_t) \tag{1}$$

where H is the likelihood of a given speaker, $\mathbf{x}_t$ is the feature vector in time domain, t is the discrete time index corresponding to speech frames, T is the total length of speech frames, M is the number of Gaussian members, ω_i is the mixture weight, and p_i is the i-th Gaussian model output.

The similarity of two speakers is then measured by this likelihood ratio, and by maximizing it we can find the target speaker identity, as shown in Equation (2).

$$SpeakerID = \arg\max_{j} \{H_j\} \tag{2}$$

where H_j is the likelihood of a specific speech utterance computed by the method in Equation (1).

3. The Feature Space Optimization.

Feature space optimization is an important step for GMM parameter estimation. Although the estimation algorithm converges to a certain degree, the estimated parameter cannot bring a good recognition performance without the proper setting of the features.

In our application, the speaker features involve 12-order MFCC features. The total number of the original feature dimension is 13 including the zero-order MFCC feature which relates to the intensity. The selected feature dimensions can be an number between 1 to 13, though the training stage may require some minimum dimension numbers dependent on the character of different algorithms.

An $n \times m$ conversion matrix can transform the n-dimensional feature vector into an m-dimensional feature vector, with reduced dimensionality, as shown in Equation (3).

$$\mathbf{x}_{m\times 1} = \mathbf{C}_{n\times m} \times \mathbf{f}_{n\times 1} \tag{3}$$

where $\mathbf{f}$ is the original 13-dimensional feature vector and $\mathbf{x}$ is the reduced feature vector.

In this paper, we only select subset of the MFCC features for the speaker identification problem, and the conversion matrix $\mathbf{C}_{n\times m}$ is reduced down to an n-dimensional vector with m non-zero values and $n-m$ zeros.

4. **The Shuffled Frog Leaping Algorithm.** In this paper, we adopt a powerful meta-heuristic optimization algorithm namely shuffled frog leaping algorithm (SFLA) for the parameter searching. In reference [9], an improved version of the SFLA is introduced and we apply this variant to our speaker identification problem.

Velocity-Verlet algorithm is introduced to the original SFLA. It is a powerful solution to the motion equation in the molecular dynamics, as shown in Equation (4) and Equation (5).

$$r(\tau+\Delta\tau) = r(\tau) + v(\tau)\Delta\tau + 0.5a(t)\Delta\tau^2 \tag{4}$$

$$v(\tau+\Delta\tau) = v(\tau) + 0.5[a(\tau) + a(\tau+\Delta\tau)]\Delta\tau \tag{5}$$

r is the position, v is the velocity, τ is the time and a is the acceleration.

The Velocity-Verlet algorithm is applied to the original SFLA algorithm to update the worst individual's position, velocity and acceleration, as shown through Equation (6) to Equation (9) [9].

$$a(k) = \lambda e^{|r_g - r(k)|}(r_g - r(k)) \tag{6}$$

$$r(k+1) = r(k) + v(k) + 0.5a(k) \tag{7}$$

$$a(k+1) = \lambda e^{|r_g - r(k+1)|}[r_g - r(k+1)] \tag{8}$$

$$v(k+1) = v(k) + 0.5[a(k) + a(k+1)] \tag{9}$$

where $r(k)$, $v(k)$ and $a(k+1)$ are the position, velocity and acceleration of the worst individual. r_g is the position of the global optimal. $r(k+1)$, $v(k+1)$ and $a(k+1)$ are the position, velocity and acceleration of the updated individual.

The advantage of this improved variant of SFLA is that the local depth searching can be enhanced by the randomization character. The balance between the population diversity and the searching efficiency brings reliable algorithm convergence.

5. **The Joint Optimization of Gaussian Mixture Model.** In this section we apply the SFLA based method to the parameter optimization in Gaussian mixture model. The details of the optimization algorithm are shown in Algorithm 1.

Algorithm 1 Parameter Optimization based on SFLA

Require: L-fold cross-validation fold number L, training dataset Φ

Ensure: EM parameters ϵ and Γ; Feature space parameters m and $\boldsymbol{c}$; GMM mixture parameter M.

1: Initialize the individual frogs in SFLA: $r_0 = [m, M, \epsilon, \Gamma, c_1, c_2, \ldots, c_n]$.
2: **for all** $q = 1, \ldots, Q$ **do**
3: Search the global optimal r_q^{opt}.
4: Update the worst individual r_{q+1} according to Equation (6) to Equation (9).
5: Estimate the Gaussian parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and ω_i in EM algorithm.
6: Update the fitness function f^q.
7: Terminate when the stopping criteria is met: $f^{q+1} < \theta$, where θ is the threshold for SFLA.
8: **end for**

The notations of the parameters in our optimization framework are listed as follows: the number of the original feature dimensions is denoted as n; the number of the selected feature dimensions is denoted as m; the importance evaluation of each dimension is denoted as c_k, where k is the index of feature dimension; the number of Gaussian members is denoted as M; the mean vector of a Gaussian member is denoted as $\boldsymbol{\mu}_i$, where i is the

index of the Gaussian member and $i = 0, 1, \ldots, M$; the covariance matrix of a Gaussian member is denoted as $\mathbf{\Sigma}_i$, where i is the index of the Gaussian member and $i = 0, 1, \ldots, M$; the weight of a Gaussian member is denoted as ω_i, which satisfies $\sum_{i=0}^{M} \omega_i = 1$.

When the SFLA is applied to our optimization problem, the fitness function is designed based on the final speaker identification error rate e, as shown in Equation (10).

$$f = -\ln\left(\frac{1}{L}\sum_{l=0}^{L} e_l\right) \tag{10}$$

where L is the number of folds in the L-fold cross validation, and l is the index of each validation.

The individual frog position r is initialized with the following parameters:

$$r = [m, M, \epsilon, \Gamma, c_1, c_2, \ldots, c_n] \tag{11}$$

The selected features are determined by the ranking of the evaluation vector $[c_1, c_2, \ldots, c_n]$ of the feature dimensions. The Gaussian parameters, including the weights, are optimized using the traditional expectation-maximization (EM) algorithm. ϵ is the threshold in EM algorithm, and it is the stop criteria that is optimized in SFLA. Γ is the maximum allowed iteration in EM algorithm, and it is also optimized in SFLA. In this way the SFLA is combined with traditional EM algorithm, and it is responsible for searching for the global optimization of the parameters that is not estimated in the traditional EM iterations.

6. **Experimental Result.** In order to verify optimization framework for the GMM based speaker identification, we carry out a number of tests in this section.

First, the baseline speaker identification system based on MFCC features and empirically set GMM classifier is presented in Figure 3. We can see that the false acceptance rate changes along with the false rejection rate. When the false acceptance rate reaches the minimum value, more samples are likely to be rejected by mistake. When the false rejection rate reaches the minimum value, more samples are likely to be accepted by mistake.

Second, the SFLA based optimization framework in Algorithm 1 is used for improved results. The error rates of the optimized system is shown in Figure 4. We can see that the performance of the speaker identification is more reliable after the parameter

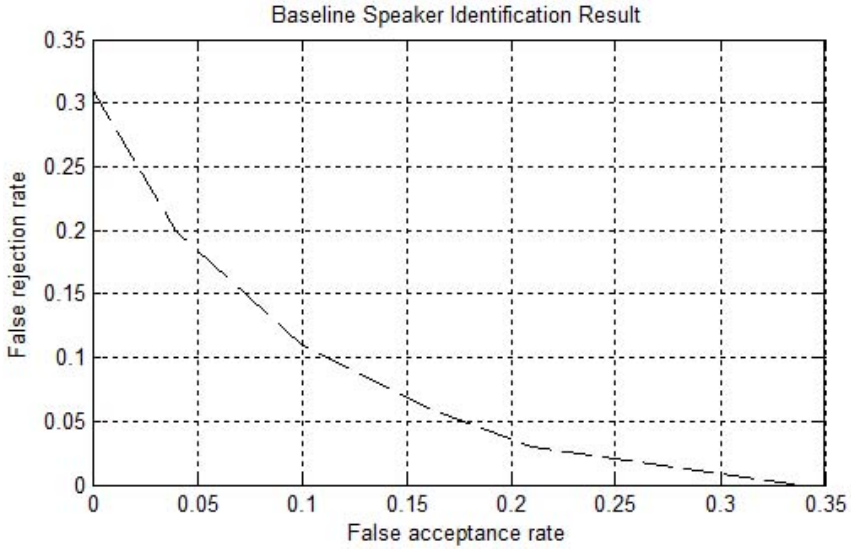

FIGURE 3. Baseline speaker identification result with false rejection rate and false acceptance rate

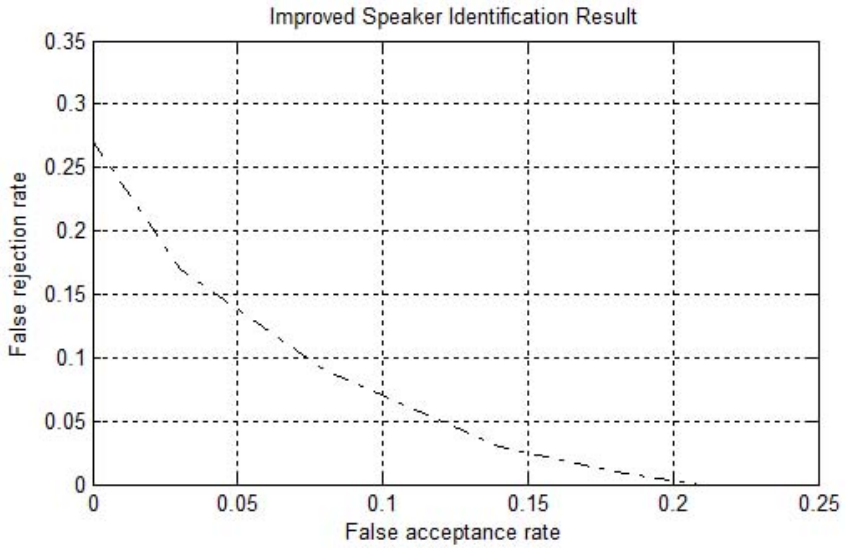

FIGURE 4. Improved speaker identification result with false rejection rate and false acceptance rate

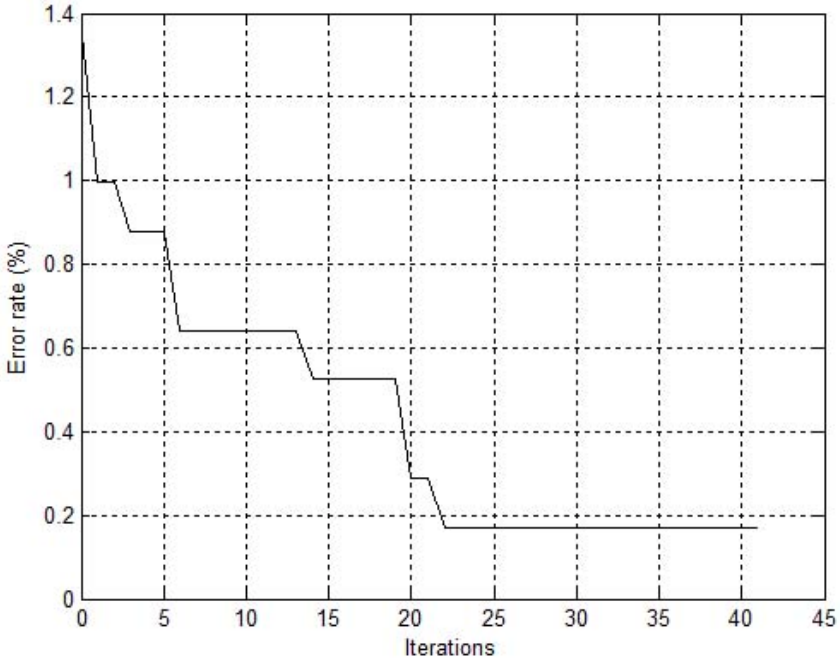

FIGURE 5. The convergence curve of Algorithm 1

optimization. The maximum false acceptance rate drops to 21% with zero false rejection rate, and the maximum false rejection rate drops to 27% with zero false acceptance rate. The SFLA algorithm may improve the feature optimization and the Gaussian mixture number determination that cannot be achieved in the EM algorithm. The convergence curve of the proposed optimization algorithm is depicted in Figure 5. We can see that after 23 iterations, the algorithm converges. Further experiments show that it takes 31 iterations to converge for the basic particle swarm optimization and 36 iterations for genetic algorithm. The error rate after each iteration is not increasing, and it may jump out the local minimal due to its meta-heuristic property. The EM algorithm is embedded in this optimization framework, and the overall convergence is ensured by SFLA.

7. **Conclusions.** In this paper we analyze the combination of SFLA and EM algorithm in a unified optimization framework. We apply the proposed algorithm in the parameter

optimization problem in Gaussian mixture model. The fitness function is designed based on the error rate of the speaker identification system. The improvement over the baseline system shows that the optimization is successful. In future work, we will further explore the possibility of dataset optimization. The possible replacement of cross validation may further enhance the robustness of the speaker identification system.

Acknowledgement. This work is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2015ZCQ-YS-02.

REFERENCES

[1] P. French, An overview of forensic phonetics with particular reference to speaker identification, *International Journal of Speech Language and the Law*, vol.10, no.2, pp.169-181, 2013.

[2] R. Togneri and D. Pullella, An overview of speaker identification: Accuracy and robustness issues, *IEEE Circuits and Systems Magazine*, vol.11, no.2, pp.23-61, 2011.

[3] H. B. Kekre and V. Kulkarni, Speaker identification by using vector quantization, *International Journal of Engineering Science and Technology*, vol.2, no.5, pp.1325-1331, 2010.

[4] J. Ding and C. T. Yen, Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications, *Multimedia Tools and Applications*, vol.4, no.1, pp.1-10, 2014.

[5] T. Kinnunen and H. Li, An overview of text-independent speaker recognition: From features to supervectors, *Speech Communication*, vol.52, no.1, pp.12-40, 2010.

[6] J. D. Wu and Y. J. Tsai, Speaker identification system using empirical mode decomposition and an artificial neural network, *Expert Systems with Applications*, vol.38, no.5, pp.6112-6117, 2011.

[7] T. Kinnunen, Z. Z. Wu and K. A. Lee, Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, pp.4401-4404, 2012.

[8] C. Huang, D. Han, Y. Bao, H. Yu and L. Zhao, Cross-language speech emotion recognition in German and Chinese, *ICIC Express Letters*, vol.6, no.8, pp.2141-2146, 2012.

[9] X. Zhang, F. Hu, L. Zhao and C. Zou, Improved shuffled frog leaping algorithm based on molecular dynamics simulations, *Journal of Data Acquisition and Processing*, vol.27, no.3, pp.327-332, 2012.