Optimization of Observation Membership Function By Particle Swarm Method for Enhancing Performances of Speaker Identification

*Jin Young Kim, *So Hee Min, *Seung You Na, **Seung Ho Choi,
*School of Electronics and Computer Engineering
Chonnam National University
Yongbong-Dong 300, Puk-gu, Gwangju 500-757, South Korea

**Dept. of Computer Engineering
Dongshin University
Daeho-Dong 252 Naju-Si, Jeollanam-Do, 520-714 South Korea

Abstract: - The performance of speaker identification is severely degraded in noisy environments. Kim and *et al* suggested the concept of observation membership for enhancing performances of speaker identification with noisy speech [1]. The method is to weight observation probabilities with observation membership values decided by SNR. In the paper [1], the authors suggested heuristic parameter values for observation membership. In this paper we apply particle swarm optimization (PSO) for obtaining the optimal parameters. With the speaker identification experiments using the VidTimit database we verify that the optimization approach can achieve better performances than the original membership function.

Key-Words: - Speaker identification, observation membership, particle swarm optimization

1 Introduction

The need of automatic speaker recognition (ASR) based on speech signal has been increased in Internet and mobile applications. But until now ASR is not successfully adopted in real service domain, because ASR has severe performance degradation in noise environment.

For coping with noise problem, many algorithms have been studied [2-6]. And they can be classified in two approaches. One is to extract parameters robust to noise as like CMS (cepstrum mean subtraction) [2] and RASTA [3]. The other is model adaptation adjusting speaker's model to noisy environments [4-5]. Recently, a new concept of observation confidence was introduced. This method weights observation probabilities depending on noise quantity. In the paper [1] Kim and *et al* proposed a method of GMM training and recognition based on observation confidence.

According to the paper [1], an observation confidence function was adopted and the function was determined by the author's heuristic knowledge. Thus the authors proposed a new possible approach, but they did not generalize the approach in the point of optimization.

In this paper we introduce particle swarm optimization (PSO) for maximizing speaker identification rate using the observation membership function. For optimization we assume that the training speech is clean, for, we think, generally clean speech is used in training aspect.

To verify our approach we perform a speaker identification experiments based on Gaussian mixture model (GMM). We evaluate out method with VidTimit database generally used in audio-visual speech recognition.

2 Speaker Identification Based on Observation Membership

Figure 1 shows the process of speaker identification based on observation membership function. In the figure parts drawn by dotted line represents the module, which estimate observation membership and consider it in probability calculation as a weight. A general

training aspect of speaker identification process is described as follows.

1) Extract feature parameters of input speech. In this paper we use Mel-Cepstrum, which is known as the best in speech processing.
2) Perform CMS (cepstrum mean subtraction) on Mel-cepstrum. CMS can eliminate the effects of channel distortion and partially overcome noise problem.
3) Train GMM models with input Mel-Cepstrum parameters filtered by CMS.

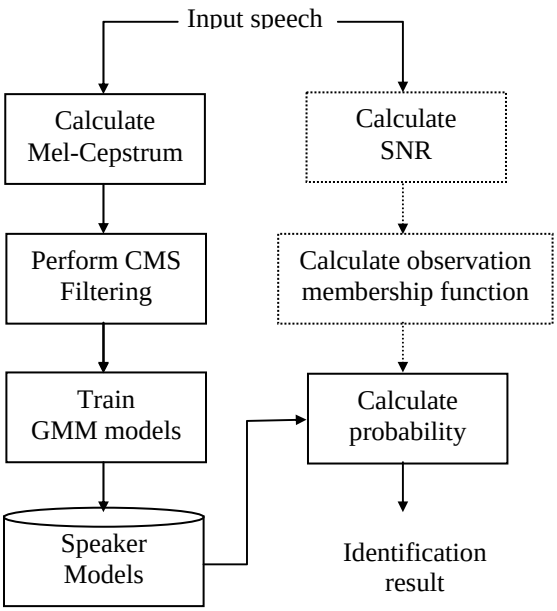

Figure 1. Speaker identification based on observation membership function.

In the testing aspect the speaker, having the highest probability of his/her GMM model, is selected as identified speaker.

In Figure 1, the modules represented by dotted line show algorithms proposed in the paper [1]. First, SNR is calculated from input speech. Second, the value of observation membership function (OMF) is calculated by SNR value. And then the probability of input speech is calculated by (Eq. 1).

$$\overline{P}_k(X) = \prod_{n=1}^{N} p_x^{\mu_n}(x_n \mid M_k), \qquad \text{(Eq.1)}$$

where $\overline{P}_k(X)$ is the probability of k-th speaker for given observation sequence x_n, $p_x(x_n \mid M_k)$ is the probability of n-th parameter and k-th speaker model, and μ_n is the value of observation membership function.

Observation membership as observation confidence represents how an observation is exact. According to the reference [1], OMF is represented with SNR. In this paper we adopt Sigmoid function as a mapping function from SNR to observation membership. That is,

$$\mu(SNR) = \frac{1}{1+e^{-a(SNR-b)}}, \qquad \text{(Eq.2)}$$

where a is a scaling parameter and b is a shift parameter. In the paper [1], 0.25 and –12.5 are used respectively for a and b.

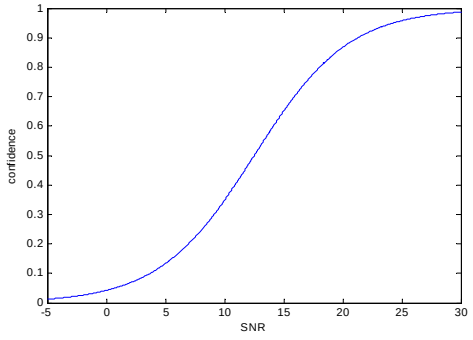

Figure 2. Example of observation membership function (a= 0.25 and b=-12.5)

According to the paper [1], the parameter values of a and b was determined empirically. And the authors showed that the performance of speaker identification could be enhanced with that observation membership function.

But, there is no any evidence that the parameter values used in the paper [1] were optimal in the point of identification rate. So the scaling and shift parameters of OMF should be obtained by some optimization theory. In this paper we adopt a method of particle swarm optimization for determining the parameter values. In section 3, we discuss about PSO and optimization of the membership function.

3 Particle Swarm Method and Optimization of Membership Function

3.1 Particle Swarm Optimization

Particle swarm optimization (PSO) was proposed by R. Eberhart and J. Kennydy in 1995 [7]. PSO was devised by imitating movements of bird flock. PSO is similar to genetic algorithm (GA) in the point that initial solutions are set to random values. Differently from GA, PSO performs renewal of each latent solution by combining previous latent solutions and random velocity vectors. These latent solutions are called particle swarm.

As a general problem, let's assume that a function $f(\cdot)$ should be optimized with the parameter P. Then PSO method is described as follows.

1) Randomize each latent solution { P_{i0} }

2) For each *j*-th iteration, perform the followings

2-1) Calculate $f(P_{ij})$ for each P_{ij}

2-2) Test the convergence. If converged, break the loop.

2-3) For each *i*, calculate optimal solution on iterations { $0,..., j-1$ }. Let's them be $pbest_{ij}$.

2-4) Calculate the global optimal solution on $pbest_{ij}$. Lets it be $gbest_j$

2-5) Calculate the velocity of each particle as follows.

$$v_{ij} = v_{ij-1} + c_1 r_1 (pbest_{ij} - P_{ij-1}) + c_2 r_2 (gbest_j - P_{ij}) \quad \text{(Eq.3)}$$

where c_1 and c_2 are constants and r_1 and r_2 are random variables.

2-6) Renew each particle.

$$P_{ij} = P_{ij-1} + v_{ij} \quad \text{(Eq.4)}$$

3) Determine the optimal solution as $gbest_j$.

3.2 Optimization of Observation Membership Function Based on PSO

For optimizing the observation membership function described by Sigmoid function we need to define a object function $f(\cdot)$. In this paper we define $f(\cdot)$ as speaker identification rate. That is, the object function could be defined by

$$f(a,b) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k} \delta(\arg\max_m P_m(X_{kl}), k)}{K\sum_{k=1}^{K} L_k}, \quad \text{(Eq.5)}$$

where $\delta(i, j)$ is delta function, X_{kl} is *l*-th speech feature vectors of *k*-th speaker, and P_m is observation probability of given feature sequence for *m*-th speaker. And $\arg\max_m P_m$ is the index of the speaker having max probability. It is defined by (Eq.1)

The object function defined (Eq.5) is a nonlinear function. So it is impossible to obtain the parameters *a* and *b* of sigmoid membership function in a closed form. Thus we apply PSO approach for optimizing the object function. We expect the parameters obtained by PSO enhance performance of speaker identification.

4 Simulation and Discussion

4.1 Outline of Speaker Identification Experiments

To evaluate our proposed algorithm, we used a speech database constructed by ETRI. The DB was developed for text-dependent speaker recognition in mobile environment. Speech signal was sampled by 8kHz, 8bit μ-law PCM codec. The DB includes 49 speaker's speech and 20 utterances per a speaker. We divide the DB into training set and testing set. Each set include 10 utterances. Figure 3 shows an example of sample utterance. The length of each utterance is about 3sec. So total 30sec speech is used for training speaker model.

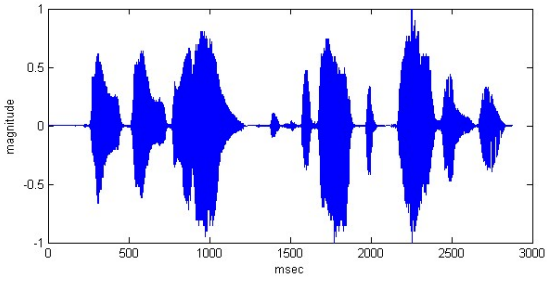

Figure 3. Waveform of sample utterance

In the experiments the frame length is 40msec and each frame is overlapped by 20msec. 12-th order Mel-cepstrum is used for analyzing speech. Also, log energy is included in speech features. For compensating noise effect, CMS is adopted. To modeling probability density function for each speaker GMM having 10 mixtures are adopted. For GMM training we apply full covariance modeling and fuzzy C-means clustering as initializing GMM models. Table 1. shows the outline of our text-dependent speaker recognition experiments.

Table 1. Summary of text-dependent speaker identification experiments

Speech DB	ETRI DB for speaker recognition in mobile environments
Sampling rate/ speech coding	8000 Hz / 8 bits μ -law PCM
Speaker number	49
Number of training files	10
Number of testing files	10
Speech features	12-th order mel-cepstrum, log energy
Frame length/ Frame increment	40ms/20ms
Channel compensation	Cepstral Mean Subtraction
GMM modeling	EM algorithm, full covariance
Gaussian mixture number	10

4.2 Experimental Results and Discussion

For verifying our PSO-based optimization method of the observation membership function, we performed two kinds of experiments. One is to optimize separately for each fixed SNR. The other is to obtain just one optimized parameters on speech signals having varying SNRs. In this paper we normalized each speech signal so that all utterances have +1 as their maximum value. Gaussian noise was added to clean speech as additive noise. (Eq.6) shows this simple process.

$$s_\eta(n) = s(n) + \alpha\eta(n), \quad \text{(Eq.6)}$$

where $s(n)$ is a clean signal, $\eta(n)$ is a random noise with the power 1 and α is a mixing parameter, which determines noise quantity (SNR). $s_\eta(n)$ is a corrupted signal by noise. For obtaining the optimal parameters of *a* and *b*, we used the DB used for GMM training. Of course, the testing is performed with the utterances not used in training.

Table 2 shows the optimal parameter values for several SNR values.

Table 2. Optimal parameter values depending on SNR

α	0.05	0.025	0.0125	0.00625
Avg SNR	7.9	11.8	16.3	20.8
Optimal *a*	-0.47	-0.51	-0.53	-0.53
Optimal *b*	9.89	10.1	12.3	12.1

From Table 2 it is observed that the optimal scaling parameter increases a little with SNR. All the values are near –0.5. However, in the case of the shift parameter, it varies from about 10 to 12 while SNR increase. The change amount is around 20%.

Figure 4 shows the identification rate with and without optimization. No-weighting means that weighting based on the observation membership function does not applied. Without-optimization means that the parameter values of the paper [1] were used in the experiments. The figure shows that SNR-based probability weighting improves the identification performance. Comparing the results of the paper [1] and our optimization, we can confirm that using the optimal parameters has better results than the original values.

Figure 5 shows the experimental results of α -independent optimization. The optimal parameter values of α -independent optimization are –0.51 and 11.1 respectively for *a* and *b*. According to Figure 5, there is no significant difference between α -independent optimization and α -dependent optimization. As a result, we don't need to use SNR-varying parameter values of the observation membership function.

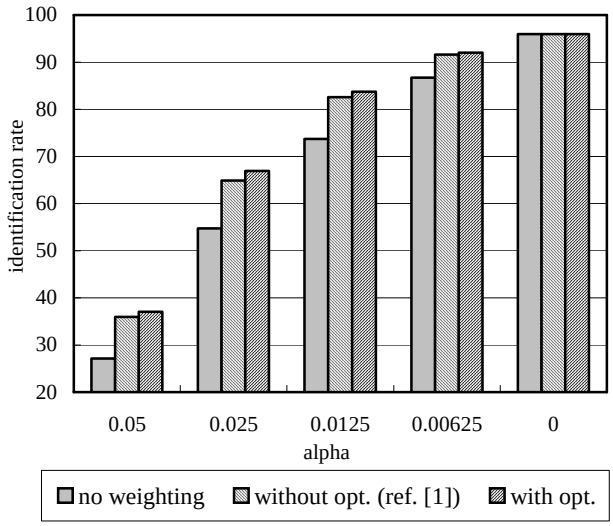

Figure 4. Identification results with/without optimization.

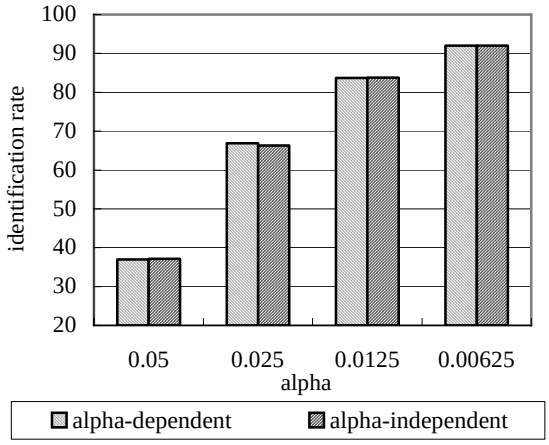

Figure 5. Experimental results with α -independent optimization (a=-0.51, b=11.1)

5 Concluding Remarks

We verified the recently proposed method of SNR-based probability weighting. We applied PSO-based approach for the optimizing observation membership function. We obtained the scaling and shift parameter values of sigmoid-type OMF by maximizing the identification rate. Through text-dependent speaker identification experiments we confirmed that PSO-based parameter optimization could enhance the performance of SNR-based probability weighting.

In the future we will develop more general optimization approach, which can be applied to the case that training utterances are corrupted by noise.

Acknowledgment

This work was supported by Chonnam National University as a sabbatical grant.

References:

[1] Jinyoung Kim, et. al, “Modified GMM Training for Inexact Observation and Its Application to Speaker Identification,” sumbitted to *Speech Science.*

[2] Rosenberg A. et al, “Cepstral channel normalization techniques for HMM-based speaker verification,” *Proc. ICSLP-94*, pp. 1835-1838, 1994.

[3] Zhen Bin, Wu Xihong, Liu Zhimin, CHI Huisheng “An Enhanced RASTA processing for speaker identification,” *Proc of 2000 ICSLP*, pp. 251-254, 2000.

[4] Mengusoglu E, “Confidence Measure based Model Adaptation for Speaker Verification,” *Proc. of the 2nd IASTED International Conference on Communications, Internet and Information Technology*, 2003.

[5] Chin-Hung Sit, Man-Wai Mak, and Sun-Yuan Kung, "Maximum Likelihood and Maximum A Posteriori Adaptation for Distributed Speaker Recognition Systems," *Proc of 1st Int. Conf. on Biometric Authentication*, 2004.

[6] Mammone R. J., Zhang X. and Ramachandran R. P., "Robust Speaker Recognition, A Feature-based Approach," *IEEE Signal Processing Magazine*, Vol. 13, No.5, 58-71, 1996.

[7] R. Eberhart and J. Kennedy, "A New Optimizer Using Partilce Swarm Theory," *Proc. of Sixth International Symposium on Micro Machine and Human Science*, pp.39-43, 1995.