Particle Swarm 기반 최적화 멤버쉽 함수에 의한 잡음 환경에서의 화자인식 성능향상

Performance Enhancement of Speaker Identification in Noisy Environments by Optimization Membership Function Based on Particle Swarm

민 소 희* · 김 진 영** · 송 민 규* · 나 승 유**
So Hee Min · Jin Young Kim · Min Gyu Song · Seung You Na

ABSTRACT

The performance of speaker identifier is severely degraded in noisy environments. A study suggested the concept of observation membership for enhancing performances of speaker identifier with noisy speech [1]. The method scaled observation probabilities of input speech by observation identification values decided by SNR. In the paper [1], the authors suggested heuristic parameter values for membership function. In this paper we attempt to apply particle swarm optimization (PSO) for obtaining the optimal parameters for speaker identification in noisy environments. With the speaker identification experiments using the ETRI database we prove that the optimization approach can yield better performance than using only the original membership function.

Keywords: Speaker identification, membership function, particle swarm optimization

1. 서 론

음성신호를 사용한 자동 화자인식은 인터넷 및 모바일 응용영역에서 그 수요가 증대되고 있다. 그러나 지금까지 자동 화자인식은 잡음 환경하에서 급격하게 성능이 저하되기 때문에 실생활에 성공적으로 서비스되지 못하고 있다.

이러한 문제점을 극복하기 위한 많은 알고리듬이 연구되어지고 있는데 [2-6] 두 가지 접근 방법으로 분류할 수 있다. 첫째는 CMS(cepstrum mean subtraction)[2]와 RASTA (relative spectra)[3]와 같은 잡음에 강인한 파라미터를 추출하는 방법이고, 둘째는 화자의 모델을 잡음에 맞도록 적응시키는 모델 적응방법이다[4-5].

최근 새로운 시도가 있었는데, 논문 [1]은 관측 신뢰도(observation confidence)라는 개념을 도입하여, 부정확한 관측을 가지고 있는 문제를 해결하기 위해, 변형된 GMM 학습과 인식방법을 제안

* 전남대학교 일반대학원 전자공학과
** 전남대학교 공과대학 전자컴퓨터공학부

한 것이다. 잡음량에 따라 관측 확률을 가중시키는 방법으로 EM 알고리듬과 최적화된 목적함수를 이용하여 VidTimit DB를 가지고 문맥독립 화자인식의 실험결과 기존의 방식보다는 3% 더 효과적임을 증명함으로써 새로운 가능성을 제시하였다. 그러나 이에 사용된 파라미터들은 화자의 경험적 노력에 의해 결정되었기 때문에 최적화의 관점에서는 일반적이지 못하다.

본 논문에서는 관측 신뢰도(멤버쉽 함수)를 최적화하기 위해 Particle Swarm Optimization(PSO) 방법[7]을 도입하여, 인식률을 극대화시키기 위한 최적화 멤버쉽 함수를 제안하였다. 논문 [1]에서는 잡음으로 오염된 일반적인 학습시료를 사용한 경우에 대한 방법론을 제시하였는데, 본 논문은 최적화를 쉽게 하기 위하여, 학습 시료는 잡음이 없는 깨끗한 음성이라고 가정하였다. 학습 시에는 깨끗한 음성을 얻을 수 있는 경우가 많기 때문에 본 논문에서 제시한 방법은 충분히 유용성이 있다. 본 논문에서 제시된 방법은 GMM 모델 기반 화자인식 실험을 통해 검증되는데, 음성 DB로는 ETRI에서 만든 한국어 화자인식용 휴대폰 음성 DB를 사용하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 화자인식의 일반적인 구조와 관측 신뢰도 또는 멤버쉽 함수에 대하여 설명한다. 3 장에서는 PSO 방법에 대하여 설명하고, PSO를 본 논문의 문제에 적용하기 위한 알고리듬에 대하여 설명한다. 본 논문에서 제시한 방법에 대한 타당성을 4 장에서 화자인식 실험을 통하여 검증한다.

2. 화자인식 구조와 관측 신뢰도

<그림 1>은 논문 [1]에서 제시한 관측 신뢰도의 개념과 화자식별 과정을 보여주고 있다. 먼저 일반적인 화자인식 과정은 다음과 같다.

1) 입력음성의 특징파라미터를 추출한다. 본 논문에서는 지금까지 가장 우수한 성능을 보인다고 알려진 멜켑스트럼(*Mel-Cepstrum*)을 사용하였다.
2) 멜켑스트럼 파라미터에 대하여 CMS를 수행한다. CMS는 채널에 의해 발생하는 채널왜곡을 제거하고 잡음에 의한 파라미터의 오염을 일부 제거하는 성질을 가지고 있다.
3) 입력된 CMS 결과 파라미터를 이용하여 GMM 학습을 수행한다. 학습결과는 화자별로 저장한다. 화자인식 시에는 입력음성의 열에 대하여 GMM 모델에 대한 발생확률을 계산하여 가장 높은 확률을 갖는 화자로 판단하게 된다.

<그림 1>에서 점선으로 표시된 부분은 관측 신뢰도를 계산하여 가중치로써 확률계산에 반영하기 위한 전처리 과정이다. 먼저 입력음성으로부터 SNR을 측정한 후 SNR에 따른 관측 신뢰도를 계산하여 입력 음성의 확률 계산 시 반영한다.

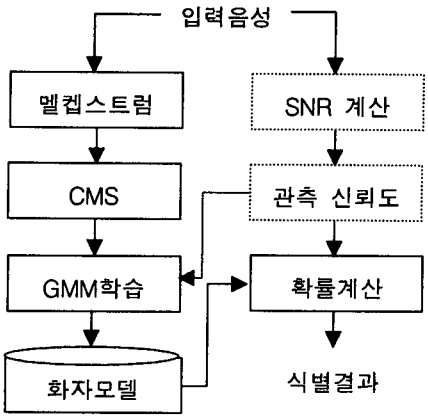

그림 1. 화자식별 알고리듬(논문 [1])

즉, 식 (1)과 같이 표현된다.

$$\overline{P_k}(X) = \prod_{n=1}^{N} p_x^{\mu_n}(x_n|M_k) \tag{1}$$

위 식에서 $\overline{P_k}(X)$는 주어진 관측 열에 대한 k번째 화자의 확률, $p_x(x_n|M_k)$는 k화자 모델에 대한 n번째 파라미터 x_n의 관측 확률이고, μ_n는 n번째 관측 파라미터의 관측 신뢰도이다. 관측 신뢰도란 주어진 관측 파라미터의 신뢰도, 즉 얼마나 정확한 측정인가를 나타내는 값이다. 논문 [1]에 의하면 관측 신뢰도는 잡음의 함수로 나타내는 것이 타당하다. 본 논문에서는 논문 [1]과 같이 시그모이드(sigmoid) 함수를 매핑(mapping) 함수로 이용하였다. 즉, 관측 신뢰도(멤버쉽 함수)는 식 (2)와 같이 정의 된다

$$\mu(SNR) = \frac{1}{1+e^{a(snr-b)}} \tag{2}$$

위 식에서 a는 스케일 파라미터이고, b는 이동 파라미터이다.

$a=-0.25$이고 $b=12.5$인 경우의 멤버쉽 함수를 <그림 2>에 나타내었다. 논문 [1]에서는 문맥독립 화자식별 실험에서 멤버쉽 함수를 사용한 경우 인식 성능이 개선됨을 확인하였다. 이 때 사용된 시그모이드 함수의 파라미터 a와 b는 경험적으로 결정된 값이다. 그러므로 최적화 이론에 의한 파라미터 a와 b를 구할 수 있어야만 한다. 본 논문에서는 최적화 파라미터 값을 결정하기 위해 PSO 방법을 도입하여 문제를 해결하고자 한다.

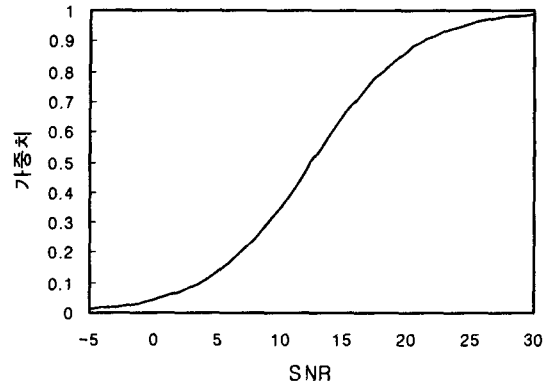

그림 2. 멤버쉽 함수의 예 (a=-0.25, b=12.5)

3. Particle Swarm Optimization를 사용한 최적화 멤버쉽 함수

3.1 Particle Swarm Optimization (PSO)

PSO는 Eberhart와 Kennedy에 의하여 1995년 제안된 방법으로서, 새 떼 또는 물고기 떼들의 먹이를 찾는 움직임을 모방하여 개발된 방법이다[7]. PSO는 초기 불규칙한 해들의 모임으로 시작한다는 면에서 유전자 알고리듬과 유사하지만, 각 잠재적인 해들이 다시 random한 속도와 이전 잠재적인 해들의 결합으로 구성된다는 측면에서 다르다. 이 잠재적인 해들의 모임을 Particle Sswarm (PS)이라고 한다. 일반적인 문제로서 파라미터 P에 의하여 최적화 되어야 할 함수 $f()$가 있다고 하자. 그러면 PSO 방법은 다음과 같다.

1) Random하게 잠재적인 해들 $\{P_{i0}\}$를 결정한다.

2) 각 iteration j에 대하여 다음을 반복한다.

2-1) 각 P_{ij}에 대하여 $f(P_{ij})$를 구한다.

2-2) 최적 f값의 변화를 계산하고, 수렴한 경우 루프를 빠져나간다.

2-3) 각 i에 대하여 $\{0...j-1\}$에 대하여 가장 최적인 해를 저장한다.
이를 $pbest_{ij}$라고 하자.

2-4) 모든 $pbest_{ij}$를 대상으로 가장 최적인 해를 저장한다. 이를 $gbest_j$라고 하자.

2-5) 각 particle의 속도를 다음과 같이 계산한다.

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

위 식에서 c_1과 c_2는 상수이며 r_1과 r_2는 임의의 수이다.

2-6) 각 particle의 값을 갱신한다.

$$P_{ij} = P_{ij-1} + v_{ij} \tag{4}$$

3) $gbest_j$를 최적의 해로 결정한다.

PSO 방법은 수학적으로 해를 구하기 어려운 비선형 문제에 일반적으로 널리 응용되고 있다. 즉 최적화 함수 f가 비선형인 경우, PSO 방법은 해가 국부적인 최적값(local optimum)을 피하면서 전체 최적인 해(global optimum)를 구하는데 사용된다.

3.2 Particle Swarm 기반 최적화 멤버쉽 함수

시그모이드 함수로 표현된 멤버쉽 함수를 최적화하기 위해서는 최적화 함수 또는 목적함수라고 불리는 함수 f가 정의되어야 한다.

본 논문에서는 화자식별의 인식률을 목적함수로 정의하였다. 즉, 최적화 함수는 다음과 같은 식으로 정의 될 수 있다.

$$f(a,b) = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k}\delta(\arg_m \max(P_m(X_{kl}),k)}{K\sum_{k=1}^{K}L_k} \tag{5}$$

위 식에서 $\delta(i,j)$는 $i=j$일 때 1이고 그렇지 않으면 0인 함수이다. X_{kl}는 k번째 화자의 l번째 음성시료이고, P_m은 주어진 시료에 대한 m번째 화자의 관측 확률로서 식 (1)로 정의된다. 그리고 $\arg_m \max P_m$는 가장 큰 확률을 갖는 화자의 인덱스(index)를 의미한다. 식 (5)로 표현된 최적화 함수는 비선형 함수로서 시그모이드 멤버쉽 함수의 파라미터 a와 b에 대하여 만족할 수학적 해를 얻기 힘들다. 따라서 본 논문에서는 3.1 절에서 설명한 PSO 방법을 이용하여 화자인식의 성능을 향상시키기 위한 최적화 파라미터 a와 b를 결정한다.

4. 시뮬레이션 및 결과 고찰

4.1 화자인식 DB 및 문맥종속 화자인식 실험개요

본 논문에서는 제안된 방법의 성능을 확인하기 위하여 ETRI에서 만든 한국어 화자인식용 휴대폰 음성 DB를 사용하여 문맥종속 화자식별 실험을 하였다. 음성데이터의 샘플링 주파수는 8 KHz이며, 8 비트 μ-law PCM 방식으로 코딩되어 제공되었다. 그리고 DB의 전체 화자의 수는 남녀 모두

49 명이고, 화자 당 음성파일은 모두 20 개로 이중 10 개씩을 학습용과 실험용으로 나누어 사용하였다. <그림 3>은 문맥종속 인식실험에 사용한 음성데이터의 파형으로 발성 시간이 약 3 초 정도로 화자모델 학습에 사용된 음성데이터는 파일 10 개를 합친 평균 약 30 초 정도의 분량이다.

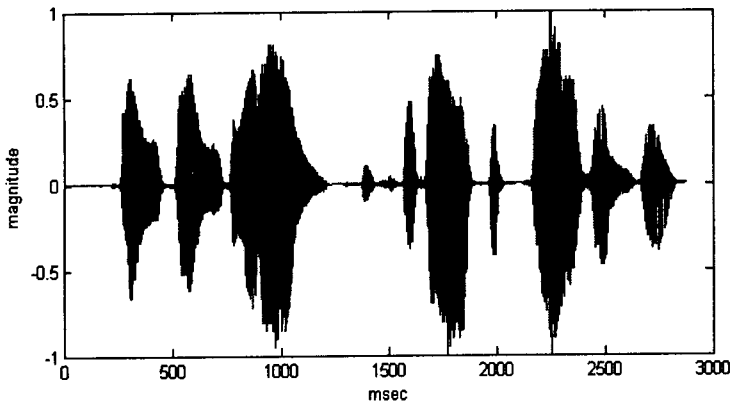

그림 3. 실험에 사용한 음성신호의 파형
(음성: 동생들은 수시로 학교에 간다.)

실험에서 입력 음성데이터의 한 프레임은 40 ms로 하였고, 20 ms씩 중첩되어 처리되도록 하고, 음성의 특징벡터는 12 차 멜켑스트럼 계수와 로그 에너지를 포함하였으며, 채널 왜곡을 보상하기 위하 CMS 방법을 적용하였다. GMM 화자모델에 포함된 Gaussian 개수는 10 개이고, EM 알고리듬에 의해 GMM 모델 λ_i의 파라미터를 반복적으로 훈련하여 구하였다. 이 과정에서 공분산 값은 full covariance를 사용하였고, 알고리듬의 초기과정에서는 fuzzy C-means clustering 방법을 사용하였다. 이에 대한 내용은 다음 <표 1>과 같다.

표 1. 문장종속 화자식별 실험의 개요

항목	규격	항목	규격
음성 DB	ETRI 휴대폰 화자인식용 음성 DB	음성특징벡터	12차 멜켑스트럼과 로그 에너지
샘플링/음성코딩	8000 Hz / 8 bits μ-law PCM	프레임 길이/중첩	40 ms/20 ms
화자수	49	채널 보상	Cepstral Mean Subtraction
화자당 학습 음성파일의 개수	10	GMM 모델	EM 알고리듬, full covariance
화자당 테스트 음성파일의 개수	10	Gaussian mixture 개수	10

4.2 실험결과 및 고찰

본 논문에서는 제시한 PSO 기반 멤버쉽 함수를 검증하기 위해 두 가지 실험을 수행하였다. 첫 번째 실험은 고정된 크기의 잡음신호에 대하여 각기 별도로 최적화를 수행한 것이고, 두 번째는 다양한 크기의 잡음에 대하여 통합적으로 최적화를 수행한 것이다. 본 실험에서는 모든 입력음성 신호의 최대크기를 +1로 정규화하고 실험을 수행하였으며, 잡음음성은 가우시안 잡음을 더하여 얻었다. 식 (6)은 깨끗한 음성으로부터 잡음 음성을 얻는 과정을 수식으로 표현한 것이다.

$$s_\eta(n) = s(n) + \alpha\eta(n) \tag{6}$$

위 식에서 $s(n)$은 잡음이 섞이지 않은 깨끗한 음성이며, $\eta(n)$은 평균전력이 1인 가우시간 불규칙 잡음이다. 또한 α는 가산되는 잡음의 양을 결정하는 파라미터이며, $s_\eta(n)$는 잡음에 오염된 신호를 의미한다.

PSO 기반 최적화에서 스케일 파라미터 a와 이동 파라미터 b를 구하기 위하여 각 화자의 GMM 모델을 학습하기 위하여 사용된 DB를 사용하였다. 그리고 검증은 학습에 참여하지 않은 DB를 대상으로 이루어졌다. 첫 번째 실험에 대한 결과를 <표 2>에 보였다.

표 2. 잡음량 별 최적화에 따른 화자식별 실험결과

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

<표 2>에서와 같이 SNR이 증가함에 따라 최적 스케일 파라미터 a는 약 -0.5 dB 근처에 있음을 알 수 있다. 한편, SNR이 약 13 dB 정도 증가함에 따라 이동 파라미터 b는 약 2 dB 정도 증가하였음을 알 수 있다.

<그림 4>는 잡음량별 최적화에 따른 인식률을 보여주고 있다. no-weighting은 멤버쉽 함수를 적용하지 않은 경우이며, without-opt의 경우는 논문 [1]의 파라미터가 실험에 사용된 경우이다. with-opt는 본 논문에서 제안한 PSO 방법으로 구한 파라미터를 사용한 경우로 논문 [1]의 방법보다 더 좋은 인식률을 나타내고 있음을 확인할 수 있다.

<그림 5>는 잡음량 독립 최적화를 사용하였을 경우와 잡음량 종속 최적화를 사용했을 경우의 인식률의 차이를 보여주고 있다. 잡음량 독립 최적화 파라미터 값은 <표 2>의 평균을 사용하는데, 스케일 파라미터 a는 -0.51 dB, 이동 파라미터 b는 11.1 dB의 값을 갖는다. 실험결과 잡음량 독립 최적화를 사용하더라도 크게 인식률의 저하가 없음을 확인할 수 있다.

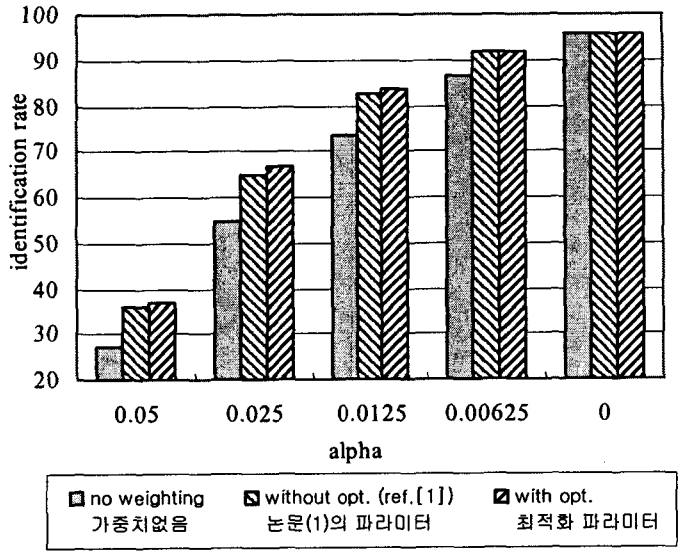

그림 4. 잡음량별 최적화에 따른 화자인식 실험결과

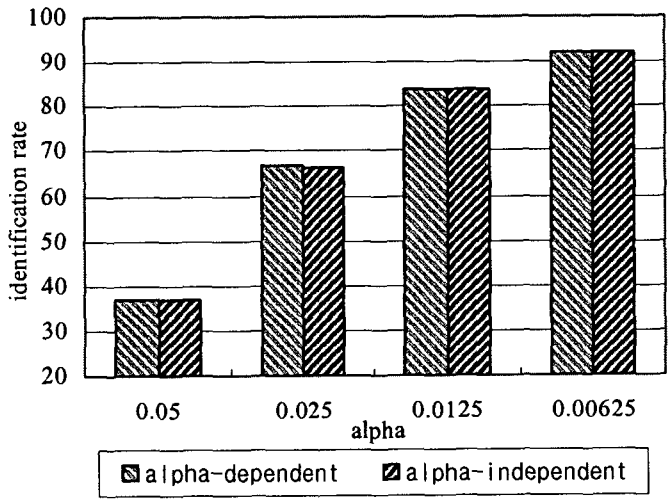

그림 5. 잡음량 독립/종속 최적화에 따른 화자인식

한편, <그림 6>은 논문 [1]에 주어진 멤버쉽 함수와 본 논문에서 PSO을 이용하여 구한 최적화 멤버쉽 함수를 비교하고 있다. 그림에서 실선은 논문 [1]의 멤버쉽 함수이고, 파선은 본 논문의 결과이다. 그림에서 볼 수 있듯이, 최적화 멤버쉽 함수값이 0.5인 경우 SNR이 2 dB 정도 아래로 이동하였다. 즉 2 dB 정도 잡음이 강화된 환경에서도 인식률이 떨어지지 않고 변함이 없음을 의미한다. 더욱 뚜렷한 특징은 천이영역이 [5 dB, 20 dB] 정도로 짧아져 10 dB이상의 잡음 환경 하에서 인식률이 향상되었음을 알 수 있다.

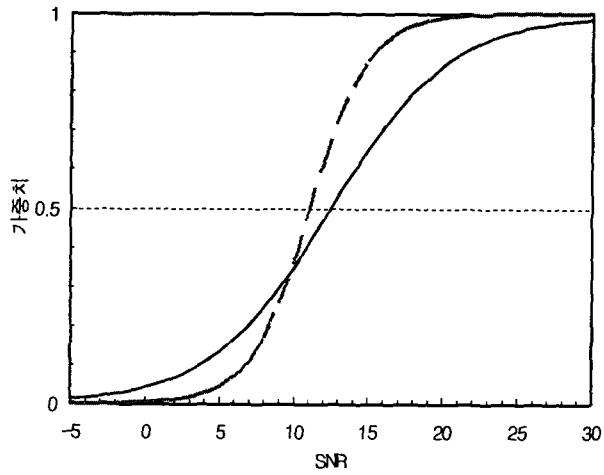

그림 6. 최적화 멤버쉽 함수와 논문[1]의 멤버쉽 함수 비교
(파선 : 최적화 멤버쉽 함수, 실선 : 논문[1])

5. 결 론

화자인식에서 잡음문제를 해결하기 위해 제시된 멤버쉽 함수 기반 확률가중 방법이 최근에 제시된 바 있다. 본 논문에서는 기존에 경험적으로 결정되었던 멤버쉽 함수를 최적화하기 위하여 Particle Swarm Optimization을 도입하였고 얻어진 최적화 파라미터가 문맥종속 화자식별 실험을 통해 가중치를 적용하지 않은 경우와 논문[1]의 파라미터를 사용한 경우 보다 더욱 효과적인 인식률을 보여줌으로써 최적화가 성공적으로 이루어졌음을 확인하였다.

향후, 인식 시뿐만 아니라 학습 시에도 잡음으로 오염된 음성 경우에도 적용할 수 있는 일반화된 방법을 개발할 것이며, 화자인식 뿐 아니라 음성인식 분야에서도 적용할 수 있도록 제안된 방법을 확장해 나갈 것이다.

참 고 문 헌

[1] Kim, Jinyoung et. al. 2007. "Modified GMM training for inexact observation and its application to speaker identification." *Speech Sciences* 14(1), 163-175.

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

[3] Zhen Bin, Wu Xihong, Liu Zhimin, Chi Huisheng. 2000. "An enhanced RASTA processing for speaker identification." *Proc of 2000 ICSLP*, 251-254.

[4] Mengusoglu, E. 2003. "Confidence measure based model adaptation for speaker verification."

Proc. of the 2nd IASTED International Conference on Communications, Internet and Information Technology.

[5] Chin-Hung, Sit, Man-Wai Mak, & Sun-Yuan Kung. 2004. "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems." *Proc of 1st Int. Conf. on Biometric Authentication.*

[6] Mammone, R. J., Zhang, X. & Ramachandran, R. P. 1996. "Robust speaker recognition, a feature-based approach." *IEEE Signal Processing Magazine* 13(5), 58-71.

[7] Eberhart, R. & Kennedy, J. 1995. "A new optimizer using particle swarm theory." *Proc. of Sixth International Symposium on Micro Machine and Human Science*, 39-43.

접수일자: 2007. 4. 27
게재결정: 2007. 5. 30

▶ 민소희
광주광역시 북구 용봉동 300 번지 전남대학교(우: 500-757)
전남대학교 일반대학원 전자공학과 박사과정
Tel: +82-62-530-0370
E-mail: minsh@chonnam.ac.kr

▶ 김진영: 교신저자
광주광역시 북구 용봉동 300 번지 전남대학교 (우: 500-757)
전남대학교 공과대학 전자컴퓨터공학부 교수
Tel: +82-62-530-1757
E-mail: beyondi@chonnam.ac.kr

▶ 송민규
광주광역시 북구 용봉동 300 번지 전남대학교 (우: 500-757)
전남대학교 일반대학원 전자공학과 박사과정
Tel: +82-62-530-0472
E-mail: smg686@lycos.co.kr

▶ 나승유
광주광역시 북구 용봉동 300 번지 전남대학교 (우: 500-757)
전남대학교 공과대학 전자컴퓨터공학부 교수
Tel: +82-62-530-1753
E-mail: syna@chonnam.ac.kr