Journal of the Korea Academia-Industrial cooperation Society
Vol. 21, No. 6 pp. 679-686, 2020

https://doi.org/10.5762/KAIS.2020.21.6.679
ISSN 1975-4701 / eISSN 2288-4688

MCE기반의 다중 특징 파라미터 스코어의 결합을 통한 화자인식 성능 향상

강지훈*, 김보람, 김규영, 이상훈
국방기술품질원

Performance Improvement of Speaker Recognition by MCE-based Score Combination of Multiple Feature Parameters

Ji Hoon Kang*, Bo Ram Kim, Kyu Young Kim, Sang Hoon Lee
Defense Agency for Technology and Quality, Jinju, Korea

요 약 본 논문에서는 화자인식 성능 향상을 위해 음원에서 개선된 특징추출 방식과 최소 분류 오차 기반의 다중 특징 벡터 스코어에 대한 가중치 추정을 사용하여 스코어 결합을 제안하였다. 제안한 특징 벡터는 Glottal Flow에서 무의미한 정보구간인 평탄한 스펙트럼 구간을 제거하기 위하여 저역통과 필터를 수행한 신호에서 인지적 선형 예측 캡스트럼 계수, 왜도, 첨도를 추출하여 구성하였다. 제안한 특징 벡터는 종래의 음원에서 멜-주파수 캡스트럼 계수, 인지적 선형 예측 캡스트럼 계수를 추출하여 가우시안 혼합 모델로 모델링한 화자인식 시스템을 개선하기 위해 사용된다. 또한, 스코어 추정과정의 신뢰성을 높이기 위하여 기존의 스코어의 확률 분포를 사용하여 가중치를 추정하는 대신 제안한 특징 벡터에서 평가된 점수와 종래의 특징 벡터에서 평가된 점수에 대하여 최소 분류 오차 기법으로 가중치를 추정하여 스코어를 결합함으로써 최적의 화자를 찾는다. 실험 결과 제안한 특징 벡터가 화자를 인식하는데 유효한 정보를 포함하고 있는 것을 확인하였다. 또한, 최소 분류 오차 기반의 다중 특징 파라미터 스코어를 결합하여 화자인식을 수행하였을 때, 종래의 화자인식 성능보다 더 우수한 성능을 나타내는 것을 확인할 수 있으며, 특히 가우시안 혼합 모델이 낮을 때 더 높은 성능향상을 보였다.

Abstract In this thesis, an enhanced method for the feature extraction of vocal source signals and score combination using an MCE-Based weight estimation of the score of multiple feature vectors are proposed for the performance improvement of speaker recognition systems. The proposed feature vector is composed of perceptual linear predictive cepstral coefficients, skewness, and kurtosis extracted with lowpass filtered glottal flow signals to eliminate the flat spectrum region, which is a meaningless information section. The proposed feature was used to improve the conventional speaker recognition system utilizing the mel-frequency cepstral coefficients and the perceptual linear predictive cepstral coefficients extracted with the speech signals and Gaussian mixture models. In addition, to increase the reliability of the estimated scores, instead of estimating the weight using the probability distribution of the convectional score, the scores evaluated by the conventional vocal tract, and the proposed feature are fused by the MCE-Based score combination method to find the optimal speaker. The experimental results showed that the proposed feature vectors contained valid information to recognize the speaker. In addition, when speaker recognition is performed by combining the MCE-based multiple feature parameter scores, the recognition system outperformed the conventional one, particularly in low Gaussian mixture cases.

Keywords : Speaker Recognition, GMM, Glottal Flow, MCE, Score Combination

*Corresponding Author : Ji-Hoon Kang(Defense Agency for Technology and Quality)
email: jh1989@dtaq.re.kr
Received April 10, 2020 Revised May 7, 2020
Accepted June 5, 2020 Published June 30, 2020

1. 서론

최근 정보통신 네트워크의 보안 취약성으로 인해 개인 정보 유출이 심각한 사회적 문제로 인식되어 개인의 정보를 타인으로부터 안전하게 보호하기 위한 정보보안에 대한 필요성이 크게 대두되고 있다. 이러한 정보통신망에서의 정보보안의 일환으로 생체기반의 보안방법이 화두가 되고 있으며, 생체인식의 한 가지 방법으로 사람이 발성하는 음성신호를 이용한 화자인식 기술이 사용되고 있다.

화자인식은 사람이 발성하는 음성신호에서 고유한 특색을 추출하여 화자를 확인하는 것을 말한다. 화자인식 기술은 생체기반의 보안방법이므로 편리하고 분실위험이 없으며, 기존의 보안카드, 주민등록증, 도장, 서명 등의 보안방법에 비해 도난이나 위조의 문제가 적으므로 매우 안전하다[4]. 기존의 화자인식의 방법으로 음성신호를 멜-주파수 캡스트럼 계수(MFCC: Mel-Frequency Cepstral Coefficients) 또는 선형 예측 캡스트럼 계수(LPCC: linear predictive cepstral coefficients)를 이용하여 특징을 추출하여 벡터 양자화(VQ: Vector Quantization), 은닉 마코브 모델(HMM: Hidden Markov Model), 가우시안 혼합 모델(GMM: Gaussian Mixture Model), 서포트 벡터 머신(SVM.: Support Vector Machine), 동적 시간 굽힘(DTW: Dynamic Time Warping), 인공신경망(ANN: Artificial neural networks)등으로 화자를 모델링하여 인식하는 방법이 널리 사용되어왔다[5]. Table 1에 기존 연구에 대하여 조사하였다. 기존 연구 방식에서는 화자의 형태학적 부분까지 모델링하고 있지 않다. 또한 기존의 화자인식 과정에서 파라미터간의 결합 방식을 살펴보았을 때, 파라미터의 가중치를 부여하는 관점에서 최적 가중치를 찾기 위하여 전수조사로써 무수한 반복으로 계산량의 증가와 비과학적인 가중치 부여 방식으로 화자인식률이 떨어질 수 있다는 문제점을 가지고 있다. 이러한 단점을 보완하기 위하여 본 논문에서는 MFCC, 인지적 선형 에측 캡스트럼 계수(PLPCC: Perceptual Linear Predictive Cepstral Coefficients)와 함께 입력음성의 형태학적 특성을 포함하고 있는 Glottal Flow에서의 PLPCC 추출 및 왜도, 첨도 특징을 추가하여 화자인식에 사용하였다. 또한, 특징 벡터를 결합할 때 각 파라미터의 스코어에 대한 가중치를 부여하는 방법으로 최소 분류 오차(MCE: Minimum Classification Error)기법 기반의 스코어 가중치 부여 기법을 이용하여 화자인식률 개선을 도모하였다.

Table 1. Different techniques of speaker recognition

Title	Feature Extraction	Classifiers
MFCC and Its Applications in Speaker Recognition[1]	LPC, MFCC	VQ
Feature Extraction And Classification for Automatic Speaker Recognition System-A Review[2]	LPC, LPCC, MFCC	VQ, GMM, SVM, DTW, HMM
Feature Extraction And Classification Techniques for Speaker Recognition: Review[3]	LPC, MFCC	GMM, ANN

본 논문의 구성은 다음과 같다. 2장에서는 기존 화자인식과 관련된 연구에 대하여 설명하고, 3장에서는 제안하는 화자인식 시스템에 대하여 기술한다. 4장에서는 제안한 방법을 적용한 화자인식 시스템의 실험 결과를 분석하고 마지막으로 5장에서 본 논문의 결론을 기술한다.

2. 관련 연구

2.1 MFCC 및 PLPCC

MFCC는 주파수 축에서의 응답의 인지적 변화도를 나타내는 계수로서 잡음환경에 강하며 사람의 청각모델의 비선형성에 기인하여 비 균일한 스케일로 주파수를 나눌 수 있다는 장점을 갖고 있다[6].

PLPCC는 MFCC와 동일하게 저주파에서는 주파수 변화에 따른 음의 강도를 심하게 느끼는 반면 고주파로 갈수록 주파수 변화에 따른 음의 강도를 잘 느끼지 못하게 되는 인간의 청각모델을 고려하여 음성신호를 분석하는 방법으로서 음성신호에 대한 주파수 강도를 조절하고 역변환을 통해 시간에 따른 음성신호의 스펙트럼을 분석하여 성도의 모양을 모델링하는 특징을 갖고 있다[7].

이 두 계수는 음성 신호를 분석하는데 가장 대표적으로 사용되고 있다.

2.2 왜도 및 첨도

표본의 분포 특성을 통계적으로 분석하기 위하여 왜도와 첨도가 사용된다. 왜도는 분포의 비대칭도를 나타내는 통계량으로 표본 집합의 분포가 대칭일 경우에 0, 오른쪽으로 치우칠 때 양수, 왼쪽으로 치우칠 때 음수 값을 갖는다. 첨도는 표본 집합의 분포가 얼마나 뾰족한지를 나타내는 통계량으로 정규분포일 경우에

0의 값을 가지며 분포가 상대적으로 뾰족해질수록 더 큰 양의 값을 갖는다. 왜도와 첨도의 정도에 관한 파라미터는 식 (1), (2)와 같이 각각 3차 혹은 4차 모멘트로부터 추정 가능하다[8].

$$S = \frac{\sum_{n=0}^{N-1}(x(n)-\mu)^3}{(N-1)\sigma^3} \quad (1)$$

$$K = \frac{\sum_{n=0}^{N-1}(x(n)-\mu)^4}{(N-1)\sigma^4} - 3 \quad (2)$$

여기서, N 은 표본의 총 개수, μ는 주어진 표본 집합의 평균, $x(n)$ 은 n 번째 표본의 값, σ 는 표본 집합이 갖는 표준편차이다.

2.3 Glottal Flow 추정

Glottal Flow란 인간의 음성 생성 메커니즘에서 폐를 통해 방출되는 공기의 흐름이 성대의 진동에 의해 토막내져 만들어지는 작은 공기 파열로 음성의 형태학적 부분을 반영하고 있다. Glottal Flow는 자기 상관 선형 예측 분석에서 얻은 잔차신호의 근사적 추정 또는 폐쇄 위상 공분산 분석(Closed-phase Covariance Analysis), 반복 적응형 역 필터(IAIF: Iterative Adaptive Inverse Filtering)을 통해 추정된다[9].

2.4 GMM 모델링

가우시안 혼합 모델은 통계적 모델링 방법으로서 구조가 간단하고 광범위한 음향학적 특성을 모델링할 수 있다는 장점이 있어서 화자를 모델링하는 가장 효과적인 방법으로 사용되고 있다. 가우시안 혼합 모델은 식 (3)과 같이 M개의 가우시안 확률 분포들의 가중된 합으로 구성된다[10].

$$P(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} w_i b_i(\boldsymbol{x}) \quad (3)$$

여기서, $\boldsymbol{x}$ 는 D차 특징 벡터, w_i 는 i 번째 가우시안 혼합의 가중치로서 가중치의 총 합은 항상 1을 만족한다. M은 가우시안 혼합의 개수, $b_i(\boldsymbol{x})$ 는 D차원의 가우시안 분포로서 식 (4)에 의해 구할 수 있다.

$$b_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad (4)$$

여기서 μ_i 는 i 번째 가우시안 분포의 평균벡터, Σ_i 는 i 번째 가우시안 분포의 공분산행렬이다. $P(\boldsymbol{x}|\boldsymbol{\lambda})$ 는 각각의 가우시안 분포의 평균벡터, 공분산행렬, 가중치에 관한 함수로 식 (5)와 같이 표현된다.

$$\boldsymbol{\lambda} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\},\ i = 1,...,m \quad (5)$$

가우시안 혼합 모델은 최대우도함수(MK: Maximum Likelihood)추정방법을 사용하여 가우시안 분포가 최대가 되는 $\boldsymbol{\lambda}$를 추정한다.

2.5 스코어 측정

각 화자의 발성에서 추출된 특징 파라미터를 사용하여 훈련된 가우시안 혼합 모델에 대하여 유사도를 측정하고 가장 큰 유사도를 갖는 모델을 발성화자로 인식한다. 유사도 측정은 식 (6)과 같이 계산된다.

$$Score(i) = \sum_{t=0}^{T-1} \log P(\boldsymbol{c_t}|\boldsymbol{S}_i) \quad (6)$$

여기서, $\boldsymbol{c_t}$ 는 t 번째 단구간 분석 프레임에서 추출된 특징 벡터, T는 입력된 발성에서 추출된 특징벡터의 총 개수, $\boldsymbol{S}_i$ 는 i 번째 화자의 모델 파라미터, $P(\boldsymbol{c_t}|\boldsymbol{S}_i)$ 는 i 번째 화자가 $\boldsymbol{c_t}$ 를 발생시킬 확률이며 식 (7)과 같이 구해진다.

$$\begin{aligned} P\left(\boldsymbol{S}_i|\{\boldsymbol{c_t}\}_0^{T-1}\right) &= \frac{P\left(\{\boldsymbol{c_t}\}_0^{T-1}|\boldsymbol{S}_i\right)P(\boldsymbol{S}_i)}{P\left(\{\boldsymbol{c_t}\}_0^{T-1}\right)} \\ &= P\left(\{\boldsymbol{c_t}\}_0^{T-1}|\boldsymbol{S}_i\right)\frac{P(\boldsymbol{S}_i)}{P\left(\{\boldsymbol{c_t}\}_0^{T-1}\right)} \\ &= cP\left(\{\boldsymbol{c_t}\}_0^{T-1}|\boldsymbol{S}_i\right) \end{aligned} \quad (7)$$

여기서, $P(\boldsymbol{S}_i)$ 는 화자인식 시스템에서 특정 화자의 입력이 들어올 확률이므로 일정하다고 간주할 수 있고, $P\left(\{\boldsymbol{c}_t\}_0^{T-1}\right)$ 는 $\{\boldsymbol{c}_0, ..., \boldsymbol{c}_{T-1}\}$ 를 수신할 확률이므로 화자인식 스코어의 순위에는 영향을 주지 않는 값이다. 따라서 그 비를 상수 $\boldsymbol{c}$로 두었다.

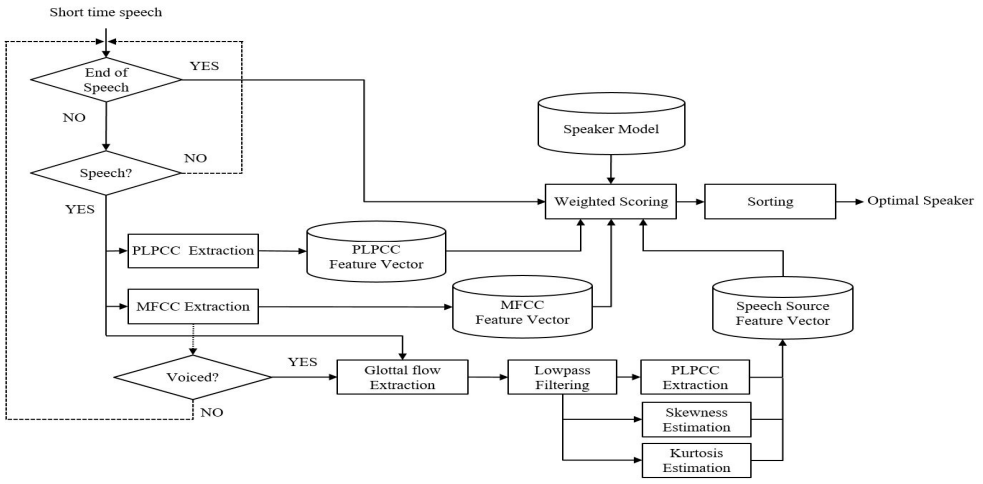

Fig. 1. Block diagram of the proposed speaker recognition system

3. 제안한 방법

본 논문에서 제안한 화자인식 시스템의 블록다이어그램을 Fig. 1에 나타내었으며, 주요 블록에 대한 설명은 아래와 같다.

3.1 GMM기반 유/무성음 분류

Glottal Flow에서의 특징 추출은 성대의 떨림으로 발생하는 유성음 구간에서 유의미한 정보를 포함하고 있으므로 해당 음성이 유성음인지 혹은 무성음 인지를 분류하여야 한다. 본 논문에서는 주기적인 신호와 비주기적인 신호의 구간에서 추출된 MFCC을 사용하여 가우시안 혼합 모델을 생성하여 유/무성음을 분류하였다. 음성으로부터 추출한 MFCC가 입력되었을 시 가우시안 혼합 모델을 이용한 유/무성음 분류는 식 (8)과 같이 수행된다.

$$\log P(\boldsymbol{c_t}|\lambda_V) \underset{unvoiced}{\overset{voiced}{\gtrless}} \log P(\boldsymbol{c_t}|\lambda_U) \tag{8}$$

여기서, $\boldsymbol{c_t}$ 는 입력된 MFCC, λ_V 는 유성음에 대한 가우시안 혼합 모델, λ_U 는 무성음에 대한 GMM이다.

3.2 Glottal Flow신호의 저역통과 필터링

Glottal Flow의 스펙트럼 중 고주파 대역에서 스펙트럼이 평탄한 부분을 갖는다. 이러한 평탄한 스펙트럼 부분에서의 정보 추출은 무의미하므로, 저역통과 필터링을 통하여 제거하여 준다. 이때, 차단주파수가 낮을경우 스펙트럼의 주기적인 부분이 포함되므로 화자에 대한 정보가 손실될 우려가 있으며, 차단주파수가 높을 경우 평탄한 스펙트럼의 제거가 원활히 이루어지지 않을 수 있다. 그러므로 저역통과 필터링의 차단주파수는 전수조사를 통하여 적절한 차단주파수를 결정한다. 저역통과 필터링을 거친 스펙트럼은 Glottal Flow의 스펙트럼 중 주기적인 성분만을 포함하고 있다.

3.3 Glottal Flow에서의 왜도 및 첨도 추출

인간의 음성은 성대의 떨림 여부에 따라서 유성음 및 무성음으로 구분된다. 무성음의 경우 성대가열린 상태에서 진동 없이 발성기관을 통과하면서 소리가 발생된다. 따라서 무성음의 경우에는 시간 영역에서 신호의 분포가 정규분포에 가깝다고 알려져 있다. 유성음의 경우 기본적으로 성대의 진동음이 발성 기관을 통과하여 소리가 발생한다. 즉, 유성음의 근원에 해당하는 음파는 주기적인 펄스 형태로 예상할 수 있으므로 생성되는 유성음은 시간영역에서의 분포는 정규분포보다 뾰족할 것으로 예상할 수 있다. Glottal Flow 또한 인간의 성대의 떨림 여부에 따라서 발생하는 신호이므로 음성에서의 특성과 같다. Fig. 2에서 음성의 유성음과 무성음부분에서 추정된 Glottal Flow의 시간영역에서의 분포를 예시하였다. Fig. 2에서와 같이 무성음에서 추정된 glottal flow의 표본 값은 평균치가 0에 가깝고 어떤 분산 값을 갖는 정규분포에 가까움을 알 수 있다.

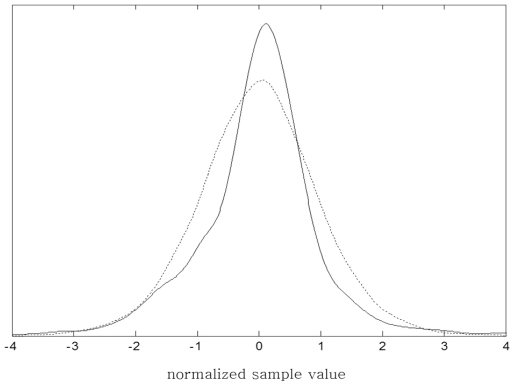

Fig. 2. Illustrating distribution characteristics of the estimated glottal flow from voiced and unvoiced (solid line: voiced(/ah/), dashed line: unvoiced(/sh/))

일반적으로 분산의 의미는 신호의 크기 정보에 불과하므로 화자를 구분할 수 있는 특징 파라미터로 사용하기에는 적합하지 않을 수 있다. 이에 반해서 유성음에서 추정된 Glottal Flow의 분포는 Fig. 2에서 예시하였듯이 무성음에 비해서 조금 더 뾰족하며 왼쪽 혹은 오른쪽으로 상대적으로 치우쳐 있다고 알려져 있다. 이러한 유성음에서 추정된 Glottal Flow의 분포 특성은 성대의 떨림 특성과 관련이 있겠고 개개인 목소리의 음색을 결정짓는 중요한 요소로 판단할 수 있으므로 화자 인식을 위해서 유용하게 활용될 수 있을 것이다.

3.4 MCE기반 스코어 가중치 추정

최소 분류 오류 기반의 스코어 가중치 훈련 기법은 손실함수를 최소화하여 가중치를 추정하는 방법으로써 손실함수는 감소시킬 파라미터에 대하여 미분 가능하여야 하며, 손실함수를 최소화 하는 것이 오차를 최소화하는데 영향을 주어야 한다는 조건을 만족하여야 한다[11].

본 논문에서는 기존의 전수조사를 통해 가중치를 추정하는 방법 대신 최소 분류 오류 기반의 스코어 가중치 훈련 기법을 사용함으로써 화자에 대한 모든 스코어 데이터를 사용하여 화자인식 결과가 최대가 되는 방향으로 가중치를 추정하였다. 이는 가중치에 대한 신뢰성을 향상시키고, 가중치 전수조사로 인한 계산량 증가에 대한 문제를 해결할 수 있다. 훈련에 사용되는 분류 오류 함수는 식 (9)와 같다.

$$d(i) = \alpha \cdot \Psi_{mfcc}(i) + \beta \cdot \Psi_{plpcc}(i) + \gamma \cdot \Psi_{smfcc}(i) \quad (9)$$

여기서, α, β, γ는 각각의 스코어에 대한 가중치이며, 가중치가 양수 값을 갖도록 하기 위하여 가중치에 지수함수를 취하여 계산한다. 각 파라미터에 대한 $\Psi(i)$ 는 식 (10)로 계산된다.

$$\Psi(i) = S_{i,ans}^{avg} - S_{i,cmp}^{avg} \quad (10)$$

여기서, $\Psi(i)$ 는 i 번째 훈련 데이터베이스에 대한 정답 화자의 스코어 평균과 최대 경쟁 화자 스코어 평균 간의 차이다. 손실함수는 식 (11)와 같이 S자모양의 함수(sigmoid)를 사용하여 분류 오류 함수가 0~1사이의 값을 갖도록 하였다.

$$l = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{1+\exp(-\Lambda d(i))} \quad (11)$$

여기서, N은 훈련 데이터베이스의 개수, Λ는 S자모양의 함수의 기울기를 조절하는 상수 값이다. 가중치를 정규화하기 위하여 가중치의 합은 항상 1을 만족하도록 하였으며 가중치의 업데이트는 식 (12), (13), (14)를 이용하여 수행한다.

$$\alpha_{n+1} = \alpha_n - \mu\nabla_\alpha l \quad (12)$$

$$\beta_{n+1} = \beta_n - \mu\nabla_\beta l \quad (13)$$

$$\gamma_{n+1} = \gamma_n - \mu\nabla_\gamma l \quad (14)$$

여기서, μ는 훈련에 사용되는 학습율, $\nabla_\alpha l$, $\nabla_\beta l$, $\nabla_\gamma l$는 손실함수를 각각의 가중치 α, β, γ로 편미분한 것으로서 식 (15), (16), (17)에 나타내었다.

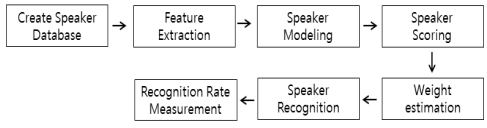

Fig. 3. Block diagram of experimental procedure

$$\nabla_\alpha l = \frac{1}{N}\sum_{i=0}^{N-1}\frac{\exp(-\Lambda d(i))\cdot\alpha\cdot\Psi_{mfcc}}{\{1+\exp(-\Lambda d(i))\}^2} \quad (15)$$

$$\nabla_\beta l = \frac{1}{N}\sum_{i=0}^{N-1}\frac{\exp(-\Lambda d(i))\cdot\beta\cdot\Psi_{plpcc}}{\{1+\exp(-\Lambda d(i))\}^2} \quad (16)$$

$$\nabla_\gamma l = \frac{1}{N}\sum_{i=0}^{N-1}\frac{\exp(-\Lambda d(i))\cdot\gamma\cdot\Psi_{smfcc}}{\{1+\exp(-\Lambda d(i))\}^2} \quad (17)$$

4. 실험 및 결과

본 논문에서 제안한 방식에 대한 절차를 Fig. 3에 나타내었다. 각 화자의 실험 데이터베이스에 대하여 Fig. 1에 따라 특징을 추출 및 GMM으로 모델링을 수행한다. 다음으로 각 화자에 대하여 스코어를 측정한다. 여기서 화자 훈련 데이터베이스에 대한 스코어를 사용하여 MCE 기반의 가중치를 추정하여 화자 테스트 데이터베이스에 대한 특징을 결합할 때 사용한다. 결합된 스코어의 값을 통해 군중모델 기반 화자 인식결과를 도출하고 최종적으로 입력화자에 대한 인식결과의 적/부판정에 따라 인식률을 측정하였다. 시험 수행과 관련된 개발소프트웨어는 C언어를 사용하였다. 실험 데이터베이스 및 실험 조건에 대한 자세한 설명은 아래와 같다.

4.1 실험 데이터베이스

실험에 사용한 음성 데이터베이스는 미국 국립 표준기술 연구소에서 구성한 TIMIT 데이터베이스로서 음성 및 화자 인식에 널리 사용되고 있다. 본 실험에는 TIMIT 데이터베이스를 가공하지 않고 사용하였다. TIMIT 데이터베이스는 미국 주요 8개 지역의 남성 438명 여성 192명으로 총 630명의 화자가 2~4초 동안의 노이즈가 포함되지 않은 상황에서 각기 다른 10문장을 표준어 또는 방언으로 발성하여 총 6300개의 음성데이터로 구성되었으며, 발성은 16 kHz/16 bit로 녹음되었다. 화자모델 훈련을 위하여 화자 당 임의의 8개의 음성 데이터를 사용하였으며, 모델 훈련에 사용되지 않은 2개의 음성 데이터는 테스트 데이터로 사용되었다.

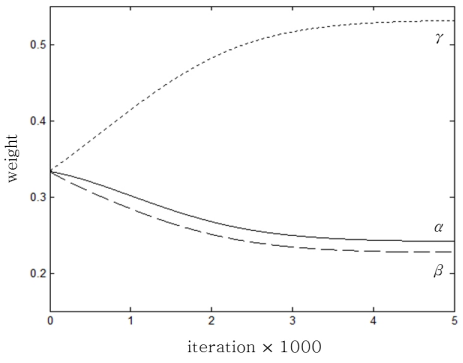

Fig. 4. Learning curves of score weights

4.2 실험 조건

단구간 분석 프레임에서 MFCC, PLPCC는 캡스트럼 계수 12차, delta-cepstral 12차, delta-delta-cepstral 12차를 포함한 총 36차의 음성 특징 파라미터를 추출하였다. 본 논문에서 제안한 특징 파라미터는 Glottal Flow 전수조사를 통하여 최적 대역 차단 주파수인 6 kHz로 저역통과 필터링을 취한 신호의 스펙트럼에 대하여 보간법을 취하여 위와 동일한 방법으로 MFCC, PLPCC와 동일하게 36차를 추출하고 왜도, 첨도를 추출하여 총 38차로 구성하였고, 제안된 특징 파라미터는 음원 특징 벡터(SFV: Source Feature Vectors)라 명시하였다.

유/무성음을 분류하는데 사용된 가우시안 혼합수는 16이며, 가우시안 혼합수 2, 4, 8, 16개를 사용하여 종래의 화자인식 시스템과 제안된 화자인식 시스템의 인식률을 비교하였다.

MCE기반의 가중치 추정은 최초 특징 파라미터의 상관없이 가중치를 0.3으로 하여 5000회 반복하였으며 Fig. 4에 가중치 수렴과정을 나타내었으며, 가중치 α, β, γ는 각각 0.24, 0.23, 0.53로 수렴하였다.

4.3 실험 결과

Table 2에 GMM 개수에 따른 종래의 특징 파라미터인 MFCC, PLPCC와 제안한 특징 파라미터인 SFV에 대한 각각의 인식률과 각 특징 파라미터를 MCE 기반으로 추정된 가중치에 의하여 결합한 인식률을 나타내었다. 본 논문에서 제안한 특징 파라미터는 종래 MFCC, PLPCC에 비하여 낮은 인식률을 보였지만, 형태학적인 부분에서의 특징을 추출하였다는 것에 의미가 있다. 종래의 특징 파라미터와 본 논문에서 제안한 특징 파라미터를 MCE 기반으로 추정된 가중치에 의하여 결합하였을 때, 각각의 특징 파라미터들로 얻어진 최대 인식률보다 낮게는 6.7 %, 높게는 25.8 %로의 개선된 인식률을 보이는 것을 확인할 수 있었다.

Table 2. Speaker recognition rate(%)

Number of Mixures	MFCC	PLPCC	SFV	Proposed fusion
2	58.17	56.90	46.59	78.49
4	72.54	72.30	56.12	90.79
8	84.13	83.96	63.81	94.60
16	88.42	88.96	65.79	95.40

5. 결론

본 논문에서는 기존의 화자인식 시스템을 보완하기 위하여 개선된 특징 추출방식과 MCE기반으로 추정된 가중치를 통해 스코어를 결합함으로서 화자인식 성능을 향상시킬 수 있는 방법에 대하여 기술하였다. 실험 결과 본 논문에서 제안한 화자인식 시스템이 종래 화자인식 시스템보다 우수한 성능을 보이는 것을 확인 하였으며, 특히 GMM 개수가 낮을수록 개선율이 높게 측정되었다.

향후 해당 논문에 관한 결과를 토대로 최근 화자인식에 널리 사용되고 있는 심층신경망(DNN: Deep Neural Network), 합성곱신경망(CNN: Convolution Neural Network)을 GMM을 대신하여 적용한다면 더우수한 화자인식 성능을 보일 것이라 판단된다.

References

[1] V. Tiwari, "MFCC and its applications in speaker recognition," IEEE International Journal on Emerging Technologies, vol. 1, no. 7, pp. 33-37, May 2013.

[2] K. Kau and N. Jain, "Feature Extraction and Classification for Automatic Speaker Recognition System - A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, pp. 1-6, January 2015.

[3] K. Dhameliya and N. Bhatt, "Feature Extraction And Classification Techniques for Speaker Recognition: A Review," IEEE International Conference on Electrical, Electronics, Signal, Communication and Optimization (EESCO), pp. 1-4, January 2015.
DOI:http://dx.doi.org/10.1109/EESCO.2015.7253831

[4] 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, January 2010.
DOI: http://dx.doi.org/10.1016/j.specom.2009.08.009

[5] Sonali T. Saste1 and Prof. S. M. Jagdale, "Comparative Study of Different Techniques in Speaker Recognition: Review," International Journal of Advanced Engineering, Management and Science (IJAEMS), vol. 3, no. 3, pp. 284-287, March 2017.
DOI: https://dx.doi.org/10.24001/ijaems.3.3.25

[6] B. Putra and Suyanto, "Implementation of secure speaker verification at web login page using Mel Frequency Cepstral coefficient-Gaussian Mixture Model (MFCC-GMM)," International Conference on Instrumentation Control and Automation (ICA), pp. 358-363, November 2011.
DOI: http://dx.doi.org/10.1109/ICA.2011.6130187

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738- 1752, April 1990.
DOI: http://dx.doi.org/10.1121/1.399423

[8] C. L. Nikias, "Higher-Order Spectral Analysis," Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Societ, pp. 319-319, October 1993.
DOI: http://dx.doi.org/10.1109/IEMBS.1993.978564

[9] T. Kinnunen and P. Alku, "On separation glottal source and vocal tract information in telephony speaker verification," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , pp. 4545-4548, April 2009.
DOI: https://doi.org/10.1109/ICASSP.2009.4960641

[10] D. Raynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, January 1995.
DOI: https://doi.org/10.1109/89.365379

[11] P. Salmela, K. Laurila, M. Lehtokangas and J. Saarinen, "On string level MCE training in MLP/HMM speech recognition system," IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 165-171, October 1999.
DOI: https://doi.org/10.1109/ICSMC.1999.825227

강 지 훈(Ji Hoon Kang) [정회원]

- 2013년 2월 : 경상대학교 전자공학과 (학사)
- 2015년 8월 : 경상대학교 전자공학과 (석사)
- 2016년 8월 ~ 2019년 7월 : 한국산업기술시험원(KTL) 연구원
- 2019년 8월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
신호처리, 음성신호처리, 화자인식, 유도무기

김 보 람(Bo Ram Kim) [정회원]

- 2008년 2월 : 서울시립대학교 전자전기컴퓨터공학과 (학사)
- 2008년 2월 ~ 2012년 4월 : 삼성전자 LCD사업부 선행기술개발팀 연구원
- 2012년 4월 ~ 2014년 1월 : 삼성디스플레이 개발팀 선임연구원
- 2018년 12월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
전자공학, 음성신호처리, 유도무기

김 규 영(Kyu Young Kim) [정회원]

- 2013년 2월 : 고려대학교 기계공학과 (학사)
- 2013년 3월 ~ 2018년 7월 : (주)한화/방산 대리
- 2018년 12월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
기계공학, 음성신호처리, 유도무기

이 상 훈(Sang Hoon Lee) [정회원]

- 2016년 8월 : 한국과학기술원 전기 및 전자공학부 (학사)
- 2016년 7월 ~ 2019년 7월 : 한화시스템 연구원
- 2019년 8월 ~ 현재 : 국방기술품질원(DTaQ) 연구원

〈관심분야〉
전자공학, 음성신호처리, 유도무기