JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 6 : 1294~1301 Jun. 2014

유/무성음 구분 및 이종적 특징 파라미터 결합을 이용한 화자인식 성능 개선

강지훈 · 정상배*

Speaker Recognition Performance Improvement by Voiced/Unvoiced Classification and Heterogeneous Feature Combination

Jihoon Kang · Sangbae Jeong*

Department of Electronics Engineering, Gyeongsang National University, Jinju 660-701, Korea

요 약

본 논문에서는 화자 인식의 성능을 개선하기 위해서 유성음 및 무성음에 대한 별도의 확률분포 모델링을 사용하였다. 또한, 종래의 멜-주파수 캡스트럼 계수 이외에 유성음 구간에서 추가적으로 왜도, 첨도, 하모닉 대 잡음비 등을 추출하여 활용하였다. 화자 인식을 위한 스코어는 유성음 및 무성음 확률분포 모델에서 각각 구해지는데 전수 조사 방식에 의해서 최적의 스코어 결합 가중치가 결정되었다. 제안된 방식의 화자인식기의 성능은 종래의 멜-주파수 캡스트럼 계수 및 화자당 하나의 혼합 가우시안 기반 확률분포 모델링을 사용한 방식과 비교되었으며 실험 결과 제안된 방식이 가우시안 혼합의 수가 낮아질수록 더 큰 성능 향상을 얻음을 알 수 있었다.

ABSTRACT

In this paper, separate probabilistic distribution models for voiced and unvoiced speech are estimated and utilized to improve speaker recognition performance. Also, in addition to the conventional mel-frequency cepstral coefficient, skewness, kurtosis, and harmonic-to-noise ratio are extracted and used for voiced speech intervals. Two kinds of scores for voiced and unvoiced speech are linearly fused with the optimal weight found by exhaustive search. The performance of the proposed speaker recognizer is compared with that of the conventional recognizer which uses mel-frequency cepstral coefficient and a unified probabilistic distribution function based on the Gassian mixture model. Experimental results show that the lower the number of Gaussian mixture, the greater the performance improvement by the proposed algorithm.

키워드 : 화자인식, 유/무성음 구분, 왜도, 첨도, 하모닉 대 잡음 비

Key word : speaker recognition, voiced/unvoiced classification, skewness, kurtosis, harmonic-to-noise ratio

접수일자 : 2014. 04. 28 심사완료일자 : 2014. 05. 22 게재확정일자 : 2014. 06. 09

* **Corresponding Author** Sangbae Jeong(E-mail:jeongsb@gnu.ac.kr, Tel:+82-55-772-1727)
Department of Electronics Engineering/ERI, Gyeongsang National University, Jinju 660-701, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.6.1294 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

사회가 발달함에 따라 삶의 편리성을 위해 음성을 기반으로 통제 가능한 인터페이스가 TV, 핸드폰, 자동차 등 많은 분야로 확산·적용되고 있으며, 향후에는 거의 모든 부분에서 보편화될 것으로 예상되고 있다. 그로인해 개인의 인터페이스를 타인으로부터 안전하게 사용하기 위해 인터페이스 보안에 대한 필요성이 대두되고 있다. 현재 음성을 기반으로 한 인터페이스 보안방법으로 화자인식 기술이 사용되고 있다. 화자인식이란 사람에 의해 발성되는 음성이 고유한 특색을 갖는다는 성질을 이용하여 각 화자의 음성신호에서 특징 정보를 추출하여 화자를 확인하는 것을 말한다. 화자인식 기술은 일반적으로 신분을 확인하는 방법인 신분증, ID 카드보다 편리하며, 생체기반의 보안방법이므로 분실할 위험이 없어 매우 안전하다[1]. 기존의 화자인식에 가장 널리 사용되는 특징 파라미터로서 선형 예측 계수(LPC: linear predictive coefficients)와 멜-주파수 캡스트럼 계수(MFCC: mel-frequency cepstral coefficients)특징 파라미터가 있으며, 본 논문에서는 기존의 MFCC 특징 파라미터에 화자의 음성신호를 특징지을 수 있는 왜도, 첨도, 하모닉대 잡음비(HNR: harmonic-to-noise ratio)를 추가적으로 사용하여 화자인식 성능 개선을 도모하였다.

화자가 갖는 파라미터를 확률적으로 모델링하기 위하여 가우시안 혼합 모델(GMM: Gaussian mixture model)이 가장 많이 사용되고 있는데, 본 연구에서는 스코어의 변별력을 높이기 위해서 음성신호를 유성음과 무성으로 분류한 후 각각의 파라미터를 GMM으로 모델링 하였다. 입력된 음성에서 계산되는 유성음 및 무성음에 관한 평균 스코어는 실험적으로 최적화된 가중치를 이용하여 선형 결합되어 최종적인 후보 화자의 스코어가 계산된다. 제안된 방법의 성능을 확인하기 위하여 기존의 MFCC를 이용한 GMM 기반의 화자인식 성능과 비교하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 GMM 기반의 특징 파라미터를 이용한 화자인식과 관련된 기존의 연구에 대하여 간략하게 설명하고, 제 3장에서는 본 논문에서 제안하는 방법에 대하여 소개한다. 제 4장에서는 제안된 새로운 방법을 적용한 실험 결과를 분석하고 마지막으로 제 5장에서 본 논문의 결론을 맺는다.

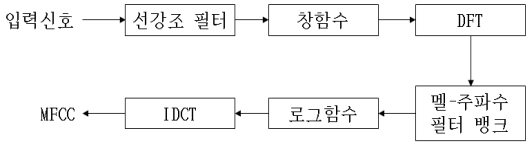


그림 1. MFCC 추출 과정
Fig. 1 Process of MFCC extraction

Ⅱ. 관련 연구

2.1. MFCC 특징 파라미터 추출

주파수 축에서의 응답의 인지적 변화도를 나타내는 MFCC는 음성 신호 분석을 위한 대표적인 특징 파라미터이다. MFCC 파라미터는 잡음환경에서 강하며 비 균일한 스케일로 주파수를 나눌 수 있다는 장점으로 화자인식에 널리 사용되고 있다. MFCC 특징 벡터를 얻기 위한 일련의 과정은 그림 1의 블록다이어그램과 같다. 입력 신호를 식 (1)로 주어지는 선강조(pre-emphasis) 필터를 이용하여 고주파 성분을 높여주고 창함수를 씌워서 신호를 프레임 단위로 나누어 준다.

$$H(z) = 1 - az^{-1}, \qquad 0.9 \le a \le 1 \tag{1}$$

나누어진 분석 프레임은 각각 이산 푸리에 변환(DFT: discrete Fourier transform)을 이용하여 주파수 영역의 응답으로 변환되며 응답의 절대치를 멜-주파수 필터뱅크로 적분하여 인간의 청각특성을 반영하도록 한다. 즉, 멜-주파수 필터뱅크는 인간의 청각 시스템이 갖는 비선형적 특성을 반영하기 위한 블록이다. 식 (2)에서 보편적으로 사용되고 있는 멜-주파수 변환을 나타내었다.

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{2}$$

식 (2)에 의해서 변환되는 선형 주파수에 대한 멜 응답을 그림 2에 나타내었다. 그 후, 로그를 취한 응답에 대해서 역 이산 코사인 변환(IDCT: inverse discrete cosine transform)를 취하여 최종적으로 MFCC 특징 파라미터를 추출한다[2].

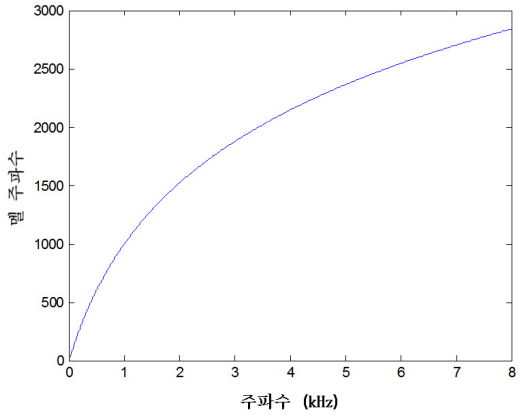


그림 2. 멜 주파수 변환
Fig. 2 Mel-frequency transform

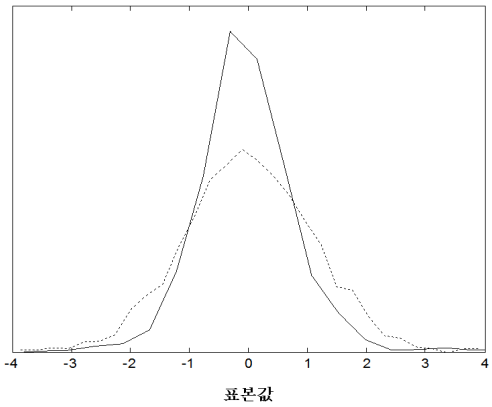


그림 3. 유성음과 무성음의 분포 특성 예시(실선: 유성음(/ah/), 파선: 무성음(/sh/))
Fig. 3 Examples of sample distributions of voiced and unvoiced speech (solid: voiced(/ah/), dashed: unvoiced (/sh/))

2.2. 왜도 및 첨도

인간의 음성은 성대의 떨림 여부에 따라서 유성음 및 무성음으로 구분된다. 무성음의 경우 성대가 열린 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 시간 영역에서의 분포는 정규분포보다 뾰족할 것으로 예상할 수 있다. 그림 3에서 유성음과 무성음의 시간영역에서의 분포를 예시하였다. 그림 3에서와 같이 무성음의 표본 값은 평균치가 0에 가깝고 어떤 분산 값을 갖는 정규분포에 가까움을 알 수 있다[3]. 일반적으로 분산의 의미는 신호의 크기 정보에 불과하므로 화자를 구분할 수 있는 특징파라미터로 사용하기에는 적합하지 않을 수 있다. 이에 반해서 유성음의 분포는 그림 3에서 예시하였듯이 무성음에 비해서 조금 더 뾰족하며 왼쪽 혹은 오른쪽으로 상대적으로 치우쳐 있다고 알려져 있다[4]. 이러한 유성음의 분포 특성은 성대의 떨림 특성과 관련이 있겠고 개개인 목소리의 음색을 결정짓는 중요한 요소로 판단할 수 있으므로 화자 인식을 위해서 유용하게 활용될 수 있을 것이다. 앞서 언급한 분포의 왜도 및 첨도의 정도에 관한 파라미터는 각각 3차 혹은 4차 모멘트로부터 추정이 가능하다[5]. 식 (3), (4)에서 주어진 표본 집합의 분포가 갖는 왜도와 첨도의 측정 방법을 나타내었다.

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

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

여기서, N 은 표본의 총 개수, μ는 주어진 표본 집합의 평균, $x(n)$은 n 번째 표본의 값, σ는 표본 집합이 갖는 표준편차이다. 식 (3)의 왜도는 평균을 기준으로 표본 집합의 분포가 대칭일 경우에 0, 오른쪽으로 치우칠 때 양수, 왼쪽으로 치우칠 때 음수 값을 갖는다. 식 (4)의 첨도는 표본 집합의 분포가 얼마나 뾰족한지를 나타낸다. 정규분포일 경우에 0의 값을 가지며 분포가 상대적으로 뾰족해질수록 더 큰 양의 값을 가진다.

2.3. 하모닉 대 잡음비

2.2절에서도 언급했듯이 사람의 성대는 화자마다 고유한 특색을 가지고 있다. 그러므로 성대가 진동하는 주기도 다르며, 음성신호에 포함되어있는 잡음의 비율도 다르다. 이 때, 잡음이란 주기성이 없는 신호를 말한다. 인간이 유성음의 경우 완벽한 주기성을 갖는다고 말하기 어려우므로 이러한 현상은 잡음이 섞여 있는 것으로 간주할 수 있다. 즉, 이러한 주적인 신호에서 잡음

성분이 얼마만큼 포함되어 있느냐를 나타내는 파라미터로서 HNR이 있다[6]. HNR의 측정은 신호의 자기상관도 함수의 계산에서 시작된다. 식 (5)에 자기 상관도를 구하는 방법을 나타내었다.

$$r_x(\tau) = E[x(n)x(n+\tau)] \tag{5}$$

여기서, $x(n)$은 입력 신호, τ는 자기상관도 지연을 나타낸다. $r_x(\tau)$는 입력 신호의 특성에 상관없이 τ가 0일 때 전역 최대이다. 만약, τ가 0이 아닌 곳에서 $r_x(0)$에 근접하는 값이 존재한다면 신호가 유성음이라고 말할 수 있으며 그 때의 자기상관도 지연치를 피치(pitch) 값 T_0, 혹은 $F_0 = 1/T_0$로 두어 피치 주파수로 정의한다. 주기 신호의 HNR 측정을 위한 수식을 유도하기 앞서서 유성음을 식 (6)과 같이 표현한다.

$$x(n) = x_H(n) + x_q(n) \tag{6}$$

여기서, $x_H(n)$은 주기 T_0를 갖는 완벽한 주기 신호이며 $x_q(n)$은 $x_H(n)$과 상관없는 백색 잡음으로 간주한다. 식 (6)을 식 (5)에 대입하여 식 (7)을 얻어낸다.

$$\begin{aligned} r_x(\tau) &= E[(x_H(n)+x_q(n))((x_H(n+\tau)+x_q(n+\tau))] \\ &= E[x_H(n)x_H(n+\tau)] + E[x_q(n)x_q(n+\tau)] \end{aligned} \tag{7}$$

이때, $E[x_H(n)x_q(n+\tau)]=0$으로 간주하였다. 자기상관도 지연 τ 가 0일 때에는 식 (7)은 식 (8)과 같이 표현이 가능하다.

$$r_x(0) = r_H(0) + r_q(0) \tag{8}$$

여기서, $r_H(0)$ 및 $r_q(0)$는 $x_H(n)$ 및 $x_q(n)$의 전력과 같다. 만약, 지연 τ 가 T_0 가 된다면 식 (7)은 식(9)로 표현이 가능하다.

$$r_x(T_0) = r_H(T_0) + r_q(T_0) = r_H(T_0) = r_H(0) \tag{9}$$

여기서, $x_q(n)$은 백색 잡음으로 간주하였으므로 $r_q(T_0)$ 를 0으로 간주하였고, $x_H(n)$의 주기는 T_0라 하였으므로 $r_H(T_0) = r_H(0)$로 두었다. 식 (8) 및 (9)를 이용하여 유성음의 주기 성분의 전력은 $r_x(T_0)$, 잡음 성분의 전력은 $r_x(0) - r_x(T_0)$로 표현됨을 알 수 있다. 이때, T_0 는 유성음 내에서 잡음의 전력이 주기 신호의 전력에 비해서 낮다는 가정 하에 식 (5)를 최대화시키는 지연 $\tau_{\max}$를 구하여 추정한다. 이러한 배경 지식을 바탕으로 HNR을 식 (10)과 같이 표현할 수 있다.

$$HNR(dB) = 10 \cdot \log_{10} \frac{r_x(\tau_{\max})}{r_x(0) - r_x(\tau_{\max})} \tag{10}$$

2.4. 가우시안 혼합모델

가우시안 혼합 모델(GMM: Gaussian mixture model)은 구조가 간단하고 광범위한 음향학적 특성을 모델링할 수 있다는 장점이 있어서 화자를 모델링하는 가장 효과적인 방법으로 사용되고 있다. GMM은 식 (11)와 같이 M개의 가우시안 확률 분포들의 가중된 합으로 구성된다[7].

$$p(\vec{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} w_i b_i(\vec{x}) \tag{11}$$

$\boldsymbol{\lambda}$는 확률분포를 구성하는 파라미터 집합, $\vec{x}$는 D차 특징 벡터, w_i는 i번째 가우시안 혼합의 가중치, M은 가우시안 혼합의 개수를 의미하며, $b_i(\vec{x})$는 D차원의 가우시안 분포로서 식 (12)에 의해 구할 수 있다.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T \boldsymbol{\Sigma}_i^{-1}(\vec{x}-\vec{\mu_i})\right\}} \tag{12}$$

$\vec{\mu_i}$는 i번째 가우시안 분포의 평균벡터, Σ_i는 i번째 가우시안 분포의 공분산행렬이다. 각 화자를 모델링하기 위하여 사용된 GMM의 훈련방법으로 최대 우도함수 추정(Maximum Likelihood)방법을 사용하여 가우시안 분포가 최대화되는 $\boldsymbol{\lambda}$를 추정한다. 이 때, 가우시안 분포가 최대화되는 매개변수를 추정하기위해 GMM의 파라미터를 벡터 양자화를 이용하여 추출된 파라미터를 초기치로 입력하여 EM(Expectation-Maximization) 알고리즘에 의해 매개변수가 수렴할 때까지 반복하여

수행한다. EM 알고리즘은 초기 모델 $\boldsymbol{\lambda}$를 설정하고 새로운 모델 $\overline{\boldsymbol{\lambda}}$를 $P(\vec{x}|\overline{\boldsymbol{\lambda}}) \geq P(\vec{x}|\boldsymbol{\lambda})$를 이용하여 추정한다. 새로운 모델은 다음 반복의 초기 모델이 현재모델과 초기모델의 차가 특정 수렴 값에 도달하거나 최대 반복 횟수를 만족 할 때까지 반복된다[8].

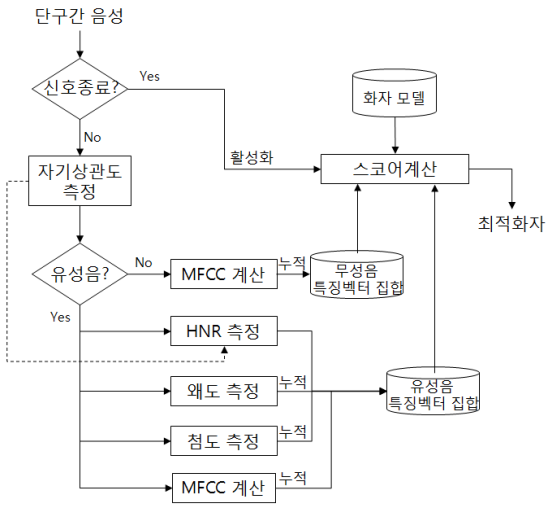


그림 4. 제안된 화자 인식 시스템의 블록다이어그램
Fig. 4 Block diagram of the proposed speaker recognition system

Ⅲ. 제안된 화자 인식 시스템

본 연구에서 제안하는 화자 인식 시스템에서는 스코어의 변별력을 높이기 위해서 각 화자를 모델링할 때 유성음 및 무성음에 대한 확률 분포 모델을 별도로 만든다. 또한, 유성음 구간에 대해서는 제 II장에서 설명한 왜도 및 첨도에 관한 파라미터를 추가적으로 추출하여 화자 인식에 활용하여 화자 인식 성능을 개선하는 것이다. 그림 4에서 본 논문에서 제안하는 화자인식 시스템의 블록다이어그램을 나타내었다. 주요 블록에 대한 설명은 다음과 같다.

3.1. 유/무성음의 분류

유성음과 무성음은 식 (8)과 식 (9)의 자기상관도 영역에서 구분된다. 구분을 위해서 $R = r_x(\tau_{\max})/r_x(0)$로 주어지는 비를 생각해본다면, 유성음 구간의 값이 무성음 구간의 값 보다 상대적으로 높은 값을 가질 수 있음을 예상할 수 있다. 즉, 어떤 단구간 신호의 자기 상관도 값이 갖는 전역 최대값의 위치 $\tau_{\max}$가 정해진 피치 주기 구간 내의 값일 때, R 값이 어떤 임계치 R_{TH} 높으면 유성음으로 선언한다. 유무성음 분리를 위해서 영교차율, 입력 신호대 선형예측 잔차 신호 에너지 비 등의 방법을 활용할 수 있다. 그러나, 본 연구에서는 HNR 측정을 위한 과정에서 유/무성음의 정보를 파악할 수 있으므로 상기와 같이 수행하였다.

3.2. 특징 파라미터의 구성

특징 파라미터의 구성은 유성음과 무성음에 대하여 다르게 구성되어있다. 3.1절을 토대로 신호가 무성음일 경우 특징 파라미터는 2.1절의 MFCC 특징 파라미터 추출 과정에서 추출된 파라미터와 2.2절의 식 (3),(4)을 이용하여 구한 왜도와 첨도를 파라미터로 추가 구성된다. 신호가 유성음일 경우 특징 파라미터는 무성음과 같이 MFCC 파라미터와 왜도, 첨도를 특징 파라미터로 사용하고 별도로 2.3절의 식 (10)을 이용하여 구한 HNR 파라미터가 추가되어 구성된다.

3.3. 스코어 계산

각 화자의 발성으로 만들어진 GMM에 대하여 임의의 화자의 발성에서 구해진 특징 파라미터를 이용하여 스코어가 식 (13)과 같이 계산된다.

$$Score(i) = \sum_{t=0}^{T-1} \log P(\vec{c_t}|\boldsymbol{S}_i) \tag{13}$$

$\vec{c_t}$는 t 번째 단구간 분석 프레임에서 추출된 특징 벡터, T는 입력된 발성에서 추출된 특징벡터의 총 개수, $\boldsymbol{S}_i$는 i번째 화자의 모델 파라미터, $P(\vec{c_t}|\boldsymbol{S}_i)$ 는 i번째 화자가 $\vec{c_t}$를 발생시킬 확률이며 식 (14)와 같이 구해진다.

$$P(\boldsymbol{S}_i|\{\vec{c_t}\}_0^{T-1}) = \frac{P(\{\vec{c_t}\}_0^{T-1}|\boldsymbol{S}_i)P(\boldsymbol{S}_i)}{P(\{\vec{c_t}\}_0^{T-1})} = \frac{P(\{\vec{c_t}\}_0^{T-1}|\boldsymbol{S}_i)P(\boldsymbol{S}_i)}{P(\{\vec{c_t}\}_0^{T-1})} \tag{14}$$

$$= P(\{\vec{c_t}\}_0^{T-1}|\boldsymbol{S}_i)\frac{P(\boldsymbol{S}_i)}{P(\{\vec{c_t}\}_0^{T-1})}$$
$$= cP(\{\vec{c_t}\}_0^{T-1}|\boldsymbol{S}_i)$$

이 때, $P(\boldsymbol{S}_i)$는 화자인식 시스템에서 특정 화자의 입력이 들어올 확률이므로 일정하다고 간주할 수 있고, $P(\{\vec{c_t}\}_0^{T-1})$는 $\{\vec{c}_0, ..., \vec{c}_{T-1}\}$를 수신할 확률이므로 화자인식 스코어의 순위에는 영향을 주지 않는 값이다. 따라서 그 비를 상수 c 로 두었다. 제안된 방식에서는 화자의 유성음 성분에 대한 스코어와 무성음 성분에 대한 스코어를 별도로 구하고 있으므로 그것을 반영하여 식 (14)를 식 (15)와 같이 변형한다. 이 때, 유성음 및 무성음 특징 파라미터는 상호간에 독립이라고 가정한다.

$$P(\boldsymbol{S}_i|\{\vec{c_t^V}\}_0^{T_V-1}, \{\vec{c_t^U}\}_0^{T_U-1}) = \frac{P(\{\vec{c_t^V}\}_0^{T_V-1}, \{\vec{c_t^U}\}_0^{T_U-1}|\boldsymbol{S}_i)P(\boldsymbol{S}_i)}{P(\{\vec{c_t^V}\}_0^{T_V-1}, \{\vec{c_t^U}\}_0^{T_U-1})}$$
$$= \frac{P(\{\vec{c_t^V}\}_0^{T_V-1}|\boldsymbol{S}_i)P(\{\vec{c_t^U}\}_0^{T_U-1}|\boldsymbol{S}_i)P(\boldsymbol{S}_i)}{P(\{\vec{c_t^V}\}_0^{T_V-1})P(\{\vec{c_t^U}\}_0^{T_U-1})}$$
$$= \frac{P(\{\vec{c_t^V}\}_0^{T_V-1}|\boldsymbol{S^V}_i)P(\{\vec{c_t^U}\}_0^{T_U-1}|\boldsymbol{S^U}_i)P(\boldsymbol{S}_i)}{P(\{\vec{c_t^V}\}_0^{T_V-1})P(\{\vec{c_t^U}\}_0^{T_U-1})}$$
$$= c' P(\{\vec{c_t^V}\}_0^{T_V-1}|\boldsymbol{S}_i^V)P(\{\vec{c_t^U}\}_0^{T_U-1}|\boldsymbol{S}_i^U) \quad (15)$$

각 기호에 대한 설명은 식 (14)의 것과 동일하다. 다만, 위첨자 '$\boldsymbol{V}$'는 유성음, '$\boldsymbol{U}$'는 무성음에 관한 특징벡터 및 모델 파라미터를 의미한다. c'은 식 (14)의 경우와 마찬가지로 화자 입력에 대한 확률과 모델 파라미터의 비로서 상수값을 의미한다. 식 (15)의 물리적 의미는 유성음과 무성음 특징 벡터의 스코어를 로그 영역에서 단순 가산하여 최종적인 스코어를 얻어낸다는 것이다. 그러나, 음성인식이나 화자 인식의 경우 무성음보다 유성음에 더 많은 정보를 담고 있다고 간주하는 것이 옳으므로 단순 가산에 의한 스코어링은 적절하지 않다고 볼 수 있다. 이러한 문제점을 해결하기 위하여 최종적으로 식 (16)로 주어지는 유성음 및 무성음 스코어의 결합 방식을 제안하였다.

$$Score(i) = \alpha \sum_{t=0}^{T_V-1} \log P(\vec{c_t^V}|\boldsymbol{S}_i^V) + (1-\alpha) \sum_{t=0}^{T_U-1} \log P(\vec{c_t^U}|\boldsymbol{S}_i^U) \quad (16)$$

Ⅳ. 실험 및 결과

4.1. 실험 데이터베이스 및 실험 조건

본 실험에서 사용된 음성 데이터베이스는 화자 당 16 kHz/16 bit로 녹음된 고립어 160개이며, 화자는 남자 15명, 여자 15명로 구성되어있다.화자는 20대 ~30대의 한국인으로 구성되었으며 발성당 지속시간은 1~2초 사이의 3~5음절 고립어였다. 화자와 고품질 마이크간의 거리는 30cm 미만의 근거리였으며 36m^2 크기의 일반 사무실환경에서 DB 수집이 수행되었다. 화자 당 120개의 음성 데이터베이스는 모델을 훈련하는데 사용되었으며, 모델을 훈련하는데 사용되지 않은 40개의 음성 데이터베이스는 테스트 DB로 사용 되었다. 본 논문에서 유성음과 무성음을 분류하는 기준인 R_{TH} 는 0.4로 두었다. 분류된 무성음 유성음 파라미터는 공통적으로 음성 데이터베이스를 2.1절의 식 (1)에서 a = 0.97로 프리 엠파시스 필터를 통과시켜 고주파 성분을 높여주고, 30 msec 프레임 크기의 해밍 윈도우를 사용하여 10 msec 씩 이동하며 MFCC 12차, 밴드 에너지의 평균값 1차, delta-cepstral 13차, delta-delta-cepstral 13차, 왜도, 첨도를 포함한 총 41차의 음성 특징 파라미터를 사용하고, 유성음일 경우에만 HNR을 포함한 총 42차의 음성 특징 파라미터를 사용하였다. 이종적 파라미터는 별도의 정규화 과정 없이 본 연구에서 제시한 수식 그대로 추정하여 사용되었다. 베이스라인 화자인식 시스템은 MFCC 12차, 밴드 에너지의 평균값 1차, delta-cepstral 13차, delta-delta-cepstral 13차 총 39차의 특징 파라미터를 이용하였다. 화자 모델링을 위하여 GMM 혼합수 8, 16, 32, 64개를 사용하여 비교하였으며, GMM은 2.4절에서 언급하였듯이 modified K-means(MKM) 군집화 알고리즘으로 초기치를 입력하고 EM 훈련의 반복횟수를 10번으로 하여 최적 모델 파라미터를 추정하였다.

4.2. 성능 평가

전체 GMM 가우시안 혼합 개수에 대하여 유성음과 무성음의 가우시안 혼합 비율을 달리하여 화자인식률을 비교하였다. 화자인식률을 비교한 결과는 표 1에 나타내었다. 이때, 화자 훈련 DB에서의 유/무성음 분리 성능은 평균적으로 77.2%를 얻을 수 있었다. 표 1의 결과를 도출하기 위하여 주어진 총 GMM 혼합 수를 유지하면서 유성음 GMM 혼합수의 비를 50%, 60%, 70%,

80%, 90%에 근사화 되도록 증가시켰으며 각 조합에 대해서 식 (17)에서 유/무성음 스코어 결합을 위해서 제안한 α값을 0에서 1사이에서 0.01 단위로 전수검사하여 최적 인식률을 측정하였다.

표 1. 유성음과 무성음의 가우시안 혼합 수에 따른 최적 화자 인식률
Table. 1 Optimal speaker recognition rate according to the number of voiced/unvoiced Gausian mixtures

가우시안 혼합수	유/무성음 혼합수	α	인식률 (%)
8	(4, 4)	0.87	85.91
	(5, 3)	0.96	87.17
	(6, 2)	0.97	85.25
	(7, 1)	0.74	**91.83**
	(베이스라인) 8		54.25
16	(8, 8)	0.68	92.41
	(10, 6)	0.75	93.83
	(11, 5)	0.83	**94.41**
	(13, 3)	0.99	93.67
	(14, 2)	0.99	93.58
	(베이스라인) 16		92.67
32	(16, 16)	0.62	96.83
	(19, 13)	0.65	96.50
	(22, 10)	0.67	**97.08**
	(25, 7)	0.75	96.92
	(28, 4)	0.98	97.00
	(베이스라인) 32		95.33
64	**(32, 32)**	0.57	**98.92**
	(38, 26)	0.62	98.42
	(45, 19)	0.61	97.92
	(51, 13)	0.59	98.42
	(58, 6)	0.63	97.75
	(베이스라인) 64		98.23

표 1을 통해서 알 수 있는 것은 첫 째, 유성음 파라미터만을 이용하여 화자인식을 하였을 때 보다 유성음과 무성음 파라미터를 모두 사용하여 화자인식을 할 때 인식률이 더욱 상승한다는 것이다. 둘 째, 전체 GMM 가우시안 혼합 개수가 많아질수록 무성음에 가해지는 가중치가 전반적으로 올라가는 것을 볼 수 있다. 이는 가우시안 혼합의 개수가 많아지면 유성음 파라미터만으로 화자를 모델링하는데 한계가 있음을 의미한다. 가우시안 혼합 개수에 따른 제안된 방법과 베이스 라인의 최대 인식률을 그림 5에서 나타내었다.

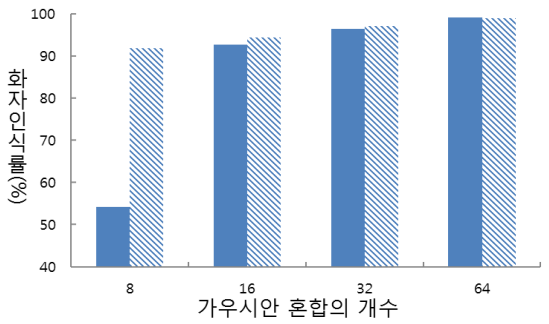


그림 5. 베이스라인 시스템과 제안된 방식(최적 인식률 조건)의 성능 비교(채움: 베이스라인, 빗금: 제안된 방식)
Fig. 5 Comparative performance between the baseline and the proposed method at best recognition rate condition (filled: baseline, slashed: proposed)

제안된 방법은 그림 5에서 볼 수 있듯이 가우시안 혼합 8, 16, 32, 64일 때, 제안된 방식은 베이스라인 보다 37.58%, 1.74%, 1.75%, 0.69% 더 높은 인식률을 나타내었다. 가우시안 혼합 개수가 일정개수 이상 증가하면 제안된 방법과 베이스 라인의 화자인식률이 거의 99%에 도달하므로 더 이상 혼합개수의 증가에 대한 실험은 수행하지 않았다. 제안된 방식에서 왜도, 첨도, HNR의 파라미터를 사용하지 않고 단순 유성음 및 무성음 분리만을 활용하였을 때의 제안된 방식에 의한 인식률은 가우시안 혼합수가 각각 8, 16, 32, 64 일 때, 84.48%, 94.05%, 96.67%, 98.78% 였다. 이 때, 표 1에서 구한 최적의 유/무성음 혼합수를 변동없이 활용하였다. 즉, 제안된 방식의 화자 인식 시스템에서는 이종적 특징 파라미터가 인식 성능 향상에 다소 영향을 준 것은 사실이지만, 유/무성음 분리 기법이 더욱 효과적이었음을 알 수 있다.

V. 결 론

본 논문에서는 화자인식의 성능을 개선하기 위하여 음성의 특징 파라미터를 유성음과 무성음의 파라미터로 분류하고, 사람의 성대가 갖는 성질을 이용하여 특징 파라미터로 왜도, 첨도, HNR을 추가하여 파라미터를 구성 방법을 제안하였다. 유/무성음 스코어의 최적 가중치를 살펴봤을 때 혼합의 수가 낮을수록 유성음의 역할이 중요함을 알 수 있었다. 화자인식률 측면에서

가우시안 혼합의 수 8일 때, 즉 낮은 혼합일 때, 제안된 방식은 베이스라인에 비해서 최대 37.58% 더 높은 성능을 보였다. 향후, 제안된 방법을 이용하여 100명 이상의 화자에 대한 화자인식과 임베디드 시스템에서의 화자인식 연구를 계획하고 있다.

REFERENCES

[1] 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.

[2] N. Ahmed, "How I came up with the discrete cosine transform," *Digital Signal Processing* Vol. 1, No. 1, pp. 4-9, 1991.

[3] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, Prentice Hall, 2010.

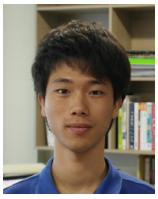
[4] W. Kleijin and K. Paliwal, *Speech Coding and Synthesis*, 2nd ed., Elsevier, 1998.

[5] C. Nikias and A. Petropulu, Higher-Order Spectra Analysis, *Prentice Hall*, 1993.

[6] C. Ferrand, "Harmonics-to-Noise Ratio: An Index of Vocal Aging," *Journal of Voice*, Vol. 16, No. 4, pp. 480-477, 2002.

[7] D. Raynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Proc.*, Vol. 3, No. 1, pp. 72-83, 1995.

[8] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.



강지훈(Jihoon Kang)

2013년 경상대학교 전자공학과(학사)
2013년 - 현재 경상대학교 전자공학과(석사과정)
※관심분야 : 음성신호처리, 화자인식

정상배(Sangbae Jeong)

1997년 부산대학교 전자공학과(학사)
1999년 한국과학기술원 전기 및 전자공학과(석사)
2002년 한국과학기술원 정보통신공학과(박사)
2002년 - 2006년 삼성종합기술원 컴퓨팅랩(책임연구원)
2006년 - 2009년 한국과학기술원 디지털미디어랩(연구조교수)
2009년 - 현재 경상대학교 전자공학과/공학연구원(부교수)
※관심분야 : 음성신호처리, 음성오디오 부호화