말소리와 음성과학 제7권 제1호 (2015.03.31)
ISSN 2005-8063, pp. 3~10
www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.1.003

화자인식을 위한 주파수 워핑 기반 특징 및 주파수-시간 특징 평가

Evaluation of Frequency Warping Based Features and Spectro-Temporal Features for Speaker Recognition

최 영 호[1] · 반 성 민[2] · 김 경 화[3] · 김 형 순[4]
Choi, Young Ho · Ban, Sung Min · Kim, Kyung-Wha · Kim, Hyung Soon

ABSTRACT

In this paper, different frequency scales in cepstral feature extraction are evaluated for the text-independent speaker recognition. To this end, mel-frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), and bilinear warped frequency cepstral coefficients (BWFCCs) are applied to the speaker recognition experiment. In addition, the spectro-temporal features extracted by the cepstral-time matrix (CTM) are examined as an alternative to the delta and delta-delta features. Experiments on the NIST speaker recognition evaluation (SRE) 2004 task are carried out using the Gaussian mixture model-universal background model (GMM-UBM) method and the joint factor analysis (JFA) method, both based on the ALIZE 3.0 toolkit. Experimental results using both the methods show that BWFCC with appropriate warping factor yields better performance than MFCC and LFCC. It is also shown that the feature set including the spectro-temporal information based on the CTM outperforms the conventional feature set including the delta and delta-delta features.

Keywords: speaker recognition, GMM-UBM, JFA, MFCC, LFCC, BWFCC, delta feature, cepstral-time matrix

1. 서론

화자인식은 음성으로부터 발성 화자가 누구인지 인식하는 기술을 말하며, 범죄과학수사, PC 및 스마트폰 보안 소프트웨어, 출입통제 시스템 등 여러 분야에서 사용되고 있다. 화자인식은 입력 음성이 등록된 화자의 음성인지를 확인하는 화자확인(speaker verification)과 다수의 등록자 중 누구의 음성인지를 판단하는 화자식별(speaker identification)의 두 가지 부류로 크게 구분된다. 또한 발성 내용에 따라, 정해진 구문의 발성을 대상으로 하는 문장종속(text-dependent) 화자인식과 자유롭게 아무 말을 하더라도 인식이 가능한 문장독립(text-independent) 화자인식으로 나눌 수 있다. 발성 내용에 제한이 없는 문장독립 화자인식은 비협조적인 사용자의 음성으로부터도 인식을 수행할 수 있어 편리하지만, 미리 등록된 음성과 인식 대상 음성의 발성 내용이 다르기 때문에 음성학적 불일치(mismatch)를 극복해야 하는 기술적인 어려움이 따른다. 이에 따라 문장독립 화자인식의 성능을 향상시키기 위해서 많은 연구들이 수행되고 있다[1].

문장독립 화자인식에 사용되는 인식 방식으로는 Gaussian mixture model-universal background model(GMM-UBM) 방식[2], support vector machine(SVM) 방식[3], joint factor analysis(JFA) 방식[4], 그리고 i-vector 방식[5] 등이 있다. 이들 방식 모두 화자 특성을 잘 표현해 주는 특징 추출 과정을 필요로 한다. 음성 신호에는 다양한 정보들이 포함되어 있으며 이러한 다양한 정보 중에서 화자 특성을 잘 나타내는 특징을 찾는 것은 화자인식 성능 향상을 위해 매우 중요하다.

현재 음성인식 및 화자인식 연구에 가장 널리 사용되는 음성 특징은 단구간 음성 스펙트럼 분석에 기반한 특징의 하나

1) 부산대학교, choiyh@pusan.ac.kr
2) 부산대학교, bansungmin@pusan.ac.kr
3) 대검찰청 음성분석실, savoix@spo.go.kr
4) 부산대학교, kimhs@pusan.ac.kr, 교신저자

이 논문은 2014년 대검찰청 연구용역의 일부로 수행되었음.

접수일자: 2015년 1월 12일
수정일자: 2015년 3월 16일
게재결정: 2015년 3월 16일

인 mel-frequency cepstral coefficient(MFCC)이다[6]. 그러나 화자들 사이의 구별을 목적으로 하는 화자인식에서 음소간의 구분을 목적으로 하는 음성인식과 동일한 특징을 사용해야 될 이유는 없다. 실제로 화자별 성도 길이에 따른 화자 특성 차이는 저주파 영역보다 고주파 영역에서 더 강조되는 특성이 있기 때문에, 모든 주파수를 동일하게 대우하는 linear frequency cepstral coefficient(LFCC)가 MFCC보다 화자인식에 더 적합할 수 있다[7]. MFCC와 LFCC 이외에 다른 주파수 스케일을 가지는 켑스트럼 특징을 통해 화자인식 성능을 추가적으로 향상시킬 수 있는지 확인하기 위하여, 본 논문에서는 bilinear 변환을 통해 주파수 워핑(warping)을 거친 bilinear warped frequency cepstral coefficient(BWFCC)를 도입해서 화자인식 성능을 비교 평가하도록 한다.

또한 매 프레임 단위로 추출되는 음성 특징들의 시간에 따른 동적 변화 특성을 나타내기 위해 delta 특징[8]이 화자인식 및 음성인식에 널리 사용되고 있다. Delta 특징 이외에 음성의 주파수-시간 특성을 보다 잘 표현해 주고자 하는 시도들 중에 temporal discrete cosine transform(TDCT) 방법이 있다[9]. 이는 음성 특징의 각 차원별 시간열을 discrete cosine transform (DCT)를 통해 표현하는 방법인데, delta 특징들의 시간열마저도 DCT로 표현하다 보니 추출되는 특징의 차원수가 많아지고 정보의 중복에 의한 비효율성 문제가 있다. 본 논문에서는 TDCT의 단점을 극복하기 위해, 정적인 켑스트럼의 시간열에 대해서만 DCT를 통해 동적인 특성을 표현하는 방법으로 켑스트럼-시간 행렬(cepstral-time matrix (CTM))을 사용하였다[10].

본 논문의 구성은 다음과 같다. 2장에서 본 논문에서 사용한 특징들의 추출방식에 대해서 설명하며, 3장에서는 기존 특징들과 제안한 방식의 특징을 사용하여 화자인식 성능을 평가하고, 마지막으로 4장에서 결론을 맺는다.

2. 특징추출 방식

2.1 주파수 워핑 기반 특징

2.1.1 MFCC

MFCC는 사람의 청각기관이 저주파수 대역에서 민감한 반면 고주파수 대역에서 상대적으로 둔감한 특성을 표현한 멜 스케일(mel scale)에 기반한 음성 특징으로서 음성인식과 화자인식 분야에서 모두 널리 사용된다. 멜 스케일은 물리적인 음높이와 청각 인지적인 음높이의 관계를 표현하는 것으로서, Stevens 등에 의해 명명되었다[11]. 식 (1)은 Hz 단위로 표현되는 물리적인 주파수 f를 mel 단위의 청각 인지적 음높이 m으로 변환하는 식이다.

$$m = 1127\log_e\left(1+\frac{f}{700}\right) \quad (1)$$

MFCC 추출과정은 <그림 1>과 같다. 우선 음성 신호로부터 매 프레임 단위로 윈도우 함수를 씌운 다음 discrete Fourier transform(DFT) 과정을 통해 시간 영역에서 주파수 영역으로 변환시키며, 실제 DFT 과정은 연산의 효율성을 위해 fast Fourier transform(FFT)의 형태로 구현된다. 그 다음으로 멜 스케일을 가지도록 식 (1)을 사용하여 주파수 축을 워핑한 다음 이 스케일에서 동일한 대역폭을 가지는 삼각 필터뱅크를 통해 필터뱅크 별 에너지를 구한다. 여기에 로그 함수를 취한 다음 DCT를 통해 최종적인 MFCC 값들을 구하게 된다.

그림 1. MFCC 추출 과정의 구성도
Figure 1. Block diagram of MFCC extraction process

2.1.2 LFCC

MFCC가 사람의 청각기관이 저주파수 대역에서는 민감한 반면 고주파수 대역에서는 둔감한 특성을 반영한 특징이라면, LFCC는 멜 스케일을 대신해서 선형 스케일을 사용하여 모든 주파수 대역에서 동일한 분해능을 가지는 특성을 반영한 특징이다. MFCC와 LFCC는 생성된 삼각 필터뱅크를 보면 차이를 명확히 알 수 있고, <그림 2>에서 MFCC와 LFCC 추출을 위한 삼각 필터뱅크 구성의 차이를 비교해서 보여준다.

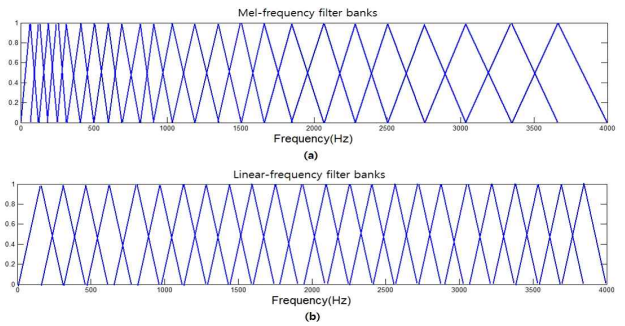

그림 2. MFCC와 LFCC 추출을 위한 필터뱅크 구성 비교
(a) MFCC 추출을 위한 멜-주파수 필터뱅크
(b) LFCC 추출을 위한 선형 주파수 필터뱅크
Figure 2. Comparison of filter bank configurations for extracting MFCC and LFCC
(a) Mel-frequency filter bank for extracting MFCC
(b) Linear frequency filter bank for extracting LFCC

앞서 언급한 바와 같이 화자인식에는 MFCC보다 LFCC가 더 적합할 수 있다는 문제 제기와 이에 따른 실험 결과가 발표된 바 있으며[7], 요약하면 다음과 같다.

화자 간의 음성 특성 차이에 큰 영향을 미치는 것으로 알려진 성도 길이 차이가 포먼트(formant) 주파수, 즉, 음성 발생 기관의 공명 주파수에 미치는 영향을 도식화하면 <그림 3>과 같다. 이처럼 매우 단순화된 모델에 의하면, 그림에서 보는 바와

같이 성도 길이 L이 $\triangle L$만큼 증가할 때 포먼트 주파수는 $\triangle L/L$의 비율만큼 줄어들게 된다. 모든 포먼트 주파수가 동일 비율로 변화될 경우 낮은 차수의 포먼트에 비해서 높은 차수의 포먼트 주파수의 차이가 더 두드러지게 되고, 결과적으로 이에 따른 화자 특성은 저주파 영역보다 고주파 영역에서 더 큰 차이를 나타내게 된다. 따라서 저주파 영역에 비해 고주파 영역의 주파수 정밀도가 떨어지게 되는 MFCC보다는 모든 주파수를 동등하게 대우하는 LFCC가 성도 길이 차이에 따른 화자 특성 차이를 보다 잘 표현할 가능성이 높아지게 된다.

그러나 음성 발생 기관의 실제 구조는 상기 모델보다 더 복잡하여 성도 길이와 포먼트 주파수의 관계가 모든 모음에 대해 일정한 수식으로 표현하는 것이 보장되지 않으며, 화자에 따른 포먼트 주파수는 성도 길이 이외에도 발음 습득 과정 등 후천적 요인에 의해서도 많은 영향을 받는다. 그리고 현재 사용되는 대부분의 화자인식 시스템들이 화자 발성의 음소 공간적 특성을 GMM으로 표현하는 것을 기반으로 하고 있으며, 음소 공간의 표현에는 음성인식에 효과적인 MFCC가 LFCC보다 유리하기 때문에, <그림 3>으로 표현되는 논리만 가지고 화자인식에 LFCC가 MFCC보다 유리하다고 단정하기에는 어려움이 따른다. 그러나 화자인식에 MFCC보다 LFCC가 유리할 수 있는 개연성은 충분히 높고, 일부 화자인식 실험에서 LFCC가 MFCC보다 성능 면에서 우수하다는 결과들도 보고되고 있다[7]. 따라서 본 논문에서는 MFCC와 LFCC를 포함하여 다양한 주파수 스케일에 대한 화자인식 실험을 통해 이들의 성능을 비교해 보기로 한다.

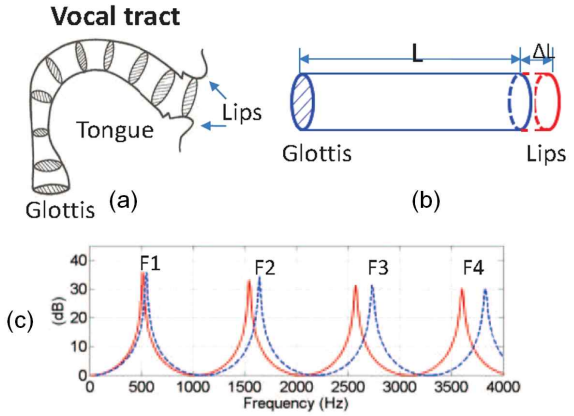

그림 3. 성도길이 차이가 포먼트 주파수에 미치는 영향[7]
(a) 성도모양 (b) 성도를 단순화 시킨 음향 튜브 모델
(c) 성도 길이에 따른 공명 주파수의 변화

Figure 3. The effect of formant frequency according to vocal tract length difference[7]
(a) Schematic of vocal tract (b) Simple acoustic tube model for vocal tract (c) Variation of formant frequency according to vocal tract length

2.1.3 BWFCC

화자인식에 효과적인 주파수 스케일을 찾기 위해서 본 연구에서는 bilinear 변환을 사용하여 다양한 주파수 워핑을 구현하고 이에 따른 켑스트럼 특징을 BWFCC라고 명명하였다.

Bilinear 변환은 식 (2)와 같이 all-pass filter를 사용하여 주파수 워핑을 수행한다[12].

$$e^{-j\tilde{\omega}} = \frac{e^{-j\omega}-\alpha}{1-\alpha e^{-j\omega}}, \quad |\alpha|<1 \tag{2}$$

여기서 ω와 $\tilde{\omega}$는 각각 워핑 이전과 이후의 주파수 성분을 의미하며, α는 warping factor이고, 식 (2)로부터 주파수 변환 관계식은 식 (3)과 같이 유도된다.

$$\tilde{\omega} = 2\arctan\left[\frac{1+\alpha}{1-\alpha}tan\left(\frac{\omega}{2}\right)\right] \tag{3}$$

<그림 4>는 bilinear 변환에 의한 warping factor별 주파수 워핑 특성을 나타낸다. 그림에 나타난 것 같이 $\alpha=0$일 때 선형 스케일이 되고, 샘플링 주파수가 8 kHz일 경우 $\alpha=0.36$에서 멜 스케일에 근접한다[12]. 따라서 α값을 조절하여 선형 및 멜 스케일 이외에 다양한 주파수 스케일을 가지는 특징을 구할

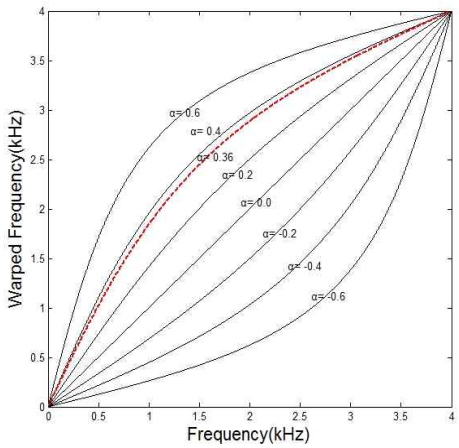

그림 4. Bilinear 변환에 의한 주파수 워핑 특성[13]

Figure 4. Frequency warping property according to bilinear transform[13]

수 있게 된다. BWFCC는 식 (3)에 적절한 α값을 대입한 후 워핑된 주파수 스케일에 따른 필터뱅크 출력 값을 이용하여 켑스트럼 계수를 구하며, 주파수 스케일 변환을 제외하고는 MFCC 및 LFCC 추출 과정과 동일하다.

Bilinear 변환에 따른 주파수 워핑을 화자인식 연구에 적용한 기존의 사례에서는 단지 5명의 화자로부터 구한 3개의 모음만으로 화자인식 실험을 수행하였다[14]. 따라서 본 연구에서는 보다 큰 규모의 데이터를 통해 주파수 스케일에 따른 화자인식 성능을 평가하고자 한다.

2.2 주파수-시간 특징

2.2.1 Delta 특징

주파수-시간 특징은 프레임 단위로 구해진 음성 스펙트럼 또는 켑스트럼의 시간에 따른 변화 특성을 음성 특징 파라미터로 사용하는 방식으로서, 음성 천이 구간에서의 포먼트 천이와 같은 화자 특징적인 정보를 표현할 수 있다. 가장 단순한 형태의 주파수-시간 특징은 음성인식이나 화자인식에 이미 널리 사용되고 있는 켑스트럼 계수들의 1차 미분 (또는 delta) 계수 및 2차 미분 (또는 delta-delta) 계수이다[8]. Delta 특징은 식 (4)와 같이 각 프레임의 앞과 뒤에 있는 프레임들의 켑스트럼 계수 벡터의 가중합으로 계산된다.

$$\triangle c_t(n)=\frac{\sum_{k=-K}^{K} k c_{t+k}(n)}{\sum_{k=-K}^{K} k^2} \tag{4}$$

여기서 $\triangle c_t(n)$은 t번째 프레임에서 n번째 켑스트럼 계수의 1차 미분계수이고, $c_{t+k}(n)$은 $t+k$번째 프레임에서 정적인 켑스트럼 계수 벡터의 n번째 계수이며, 이는 현재 프레임 앞뒤의 K개 프레임까지 총 $2K+1$개의 켑스트럼 계수들의 가중합 형태이다.

2차 미분계수는 정적인 켑스트럼 계수 대신 1차 미분계수들의 가중합으로 식 (5)와 같이 계산된다.

$$\triangle\triangle c_t(n)=\frac{\sum_{k=-K}^{K} k \triangle c_{t+k}(n)}{\sum_{k=-K}^{K} k^2} \tag{5}$$

여기서 $\triangle\triangle c_t(n)$은 t번째 프레임에서 n번째 2차 미분계수이다.

2.2.2 TDCT를 이용한 특징

Temporal DCT(TDCT)는 프레임 별로 구해진 MFCC + delta + delta-delta 특징벡터의 프레임에 따른 변화 특성을 Hamming 윈도우와 DCT 변환을 통해 적은 차수의 특징 파라미터로 변환하는 방법으로 그 추출과정은 <그림 5>에 나타나 있으며[9], TDCT 특징은 식 (6)과 같이 계산된다.

$$\begin{aligned} &TDCT_t(m\times K+k) \\ &=\sum_{l=-L}^{L} w(L+l)c_{t+l}(m)\cos\left[\frac{\pi k}{2L+1}\left(l+L+\frac{1}{2}\right)\right], \\ &\qquad 0\le m\le M-1,\ 1\le k\le K \end{aligned} \tag{6}$$

여기서 $TDCT_t(n)$는 t번째 프레임에서 $M\times K$차원의 TDCT 벡터의 n번째 성분이고, M은 MFCC와 delta 및 delta-delta특징의 전체 차수이며, K는 특징으로 사용하는 DCT 계수의 개수, $2L+1$은 블록의 크기이다. 그리고 $w(i)$는 Hamming 윈도우 함수이며, 다음 식과 같다.

$$w(i)=0.54-0.46\cos\left(\frac{i\pi}{L}\right),\ 0\le i\le 2L \tag{7}$$

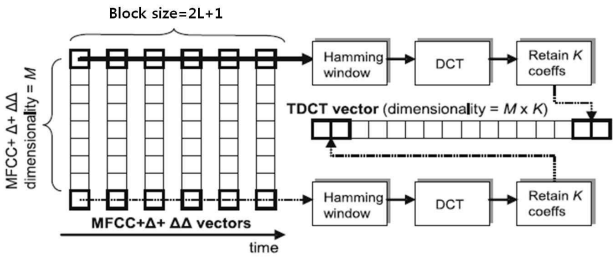

그림 5. TDCT 특징 추출 과정[9]
Figure 5. Extraction process of TDCT feature[9]

이때 대부분의 에너지를 포함하는 낮은 차수의 DCT 계수만 사용함으로써 TDCT 벡터의 차원을 줄일 수 있지만, MFCC 특징뿐 아니라 delta, delta-delta에 대해서도 TDCT를 적용하기 때문에 매우 큰 차원의 특징 파라미터를 추출하게 된다. 20차 MFCC 특징과 그 delta 및 delta-delta 특징들로 총 60차 특징을 사용할 경우, TDCT에 사용하는 DCT 계수가 3개라고 하면 ($K=3$) TDCT 특징의 차원은 $60\times 3=180$이 된다.

2.2.3 켑스트럼-시간 행렬을 이용한 특징

본 논문에서는 켑스트럼-시간 행렬을 이용하여 추출한 주파수-시간 특징을 MFCC와 같은 정적(static) 특징과 함께 사용하여 화자인식을 수행함으로써 TDCT 특징의 차원 수가 큰 문제점을 보완한다. 켑스트럼-시간 행렬은 $2L+1$크기의 연속적인 정적 특징 벡터들에 DCT를 적용하여 다음 식과 같이 구한다[10].

$$C_t(n,m)=\sum_{l=-L}^{L} c_{t+l}(n)\cos\left[\frac{\pi m}{2L+1}\left(l+L+\frac{1}{2}\right)\right] \tag{8}$$

여기서 $C_t(n,m)$은 t번째 프레임에서의 켑스트럼-시간 행렬의 요소로서, 이 행렬의 행을 표현하는 n은 정적 특징 벡터에서 몇 번째 계수인지를 나타내며, 열을 표현하는 m은 DCT의 몇 번째 계수인지를 나타낸다.

2.2.1절에서 식 (4)와 (5)로 표현된 켑스트럼 계수의 delta와 delta-delta 특징은 다음 식 (9)와 (10)에서 보는 바와 같이 켑스트럼-시간 행렬에서 (0번째 열을 제외한) 첫 번째 및 두 번째 열벡터의 성분들과 그 성격이 유사하다.

$$C_t(n,1)=\sum_{l=-L}^{L} c_{t+l}(n)\cos\left[\frac{\pi}{2L+1}\left(l+L+\frac{1}{2}\right)\right] \tag{9}$$

$$C_t(n,2)=\sum_{l=-L}^{L} c_{t+l}(n)\cos\left[\frac{2\pi}{2L+1}\left(l+L+\frac{1}{2}\right)\right] \tag{10}$$

여기서 $C_t(n,1)$과 $C_t(n,2)$은 각각 켑스트럼-시간 행렬에서 0번째 열을 제외한 첫 번째 및 두 번째 열벡터의 n번째 계수를

나타낸다. <그림 6>에서 보는 바와 같이 delta 특징과 켑스트럼-시간 행렬로부터 구한 특징 모두 적절한 기저함수를 이용하여 정적인 켑스트럼 벡터들의 시간에 따른 가중합으로 표현되는 것임을 알 수 있다.

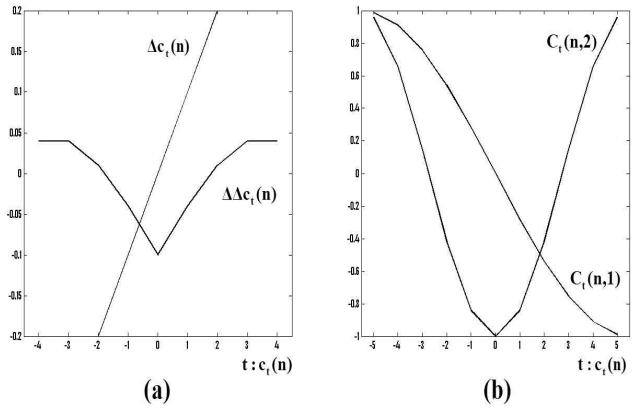

그림 6. Delta 및 DCT 기저함수[10]
(a) Delta 기저함수($K=2$) (b) DCT 기저함수($L=5$)
Figure 6. Delta and DCT basis functions[10]
(a) Delta basis function($K=2$) (b) DCT basis function($L=5$)

Delta 특징과 켑스트럼-시간 행렬에서 추출한 저차(1차 및 2차)특징은 둘 다 인접 프레임들 사이의 켑스트럼 계수의 시간에 따른 동적인 변화 특성을 표현한다는 면에서 공통점을 가지나, 전자는 동적인 변화를 단순히 1차 및 2차 미분 값을 통해 나타낸 것이고, 후자는 시간에 따른 변화 곡선 자체를 저차 DCT 계수로 근사화시킨 것이다. 참고로 DCT가 가지는 에너지 압축(energy compaction)성질로 인해 저차의 DCT 계수들만으로도 동적인 변화의 주요 특성을 잘 나타낼 수 있다. Delta 특징과 켑스트럼-시간 행렬에서 구한 저차 특징의 또 하나의 차이는, <그림 6>에서 보는 바와 같이 전자는 차수에 따라 사용되는 인접 프레임의 개수가 달라지는 반면에, 후자의 경우 차수에 관계없이 일정한 개수의 인접 프레임들이 사용된다는 점이다.

유사 연구로 참고문헌 [15]에서도 켑스트럼-시간 행렬로부터 추출한 여러 가지 형태의 특징들을 화자인식에 적용했으나, 기존 연구에서는 정적 특징으로부터 2D-DCT하여 구한 켑스트럼-시간 행렬의 성분들만으로 특징벡터를 구성하였지만, 본 연구에서는 동일하게 구한 켑스트럼-시간 행렬의 첫 번째 열벡터와 두 번째 열벡터를 정적 특징과 함께 사용했다는 점에서 근본적인 차이가 있다.

3. 실험 및 결과

3.1 실험 환경

본 논문에서는 기존에 화자인식에서 널리 사용되는 GMM-UBM 방식과 최신의 화자인식 방식 중 하나인 JFA 방식을 사용하여 화자인식 실험을 수행하였으며, 이들 두 방식 모두 화자인식용 open-source toolkit인 ALIZE 3.0을 사용하였다[16]. 평가용 DB로는 미국 National Institute of Standard and Technology(NIST)의 speaker recognition evaluation(SRE) 2004 DB[17] 중 301 명 화자로 구성된 핵심 테스트(core test)용 DB를 사용하였다. 그 중에서 화자모델 훈련에는 약 5분 길이의 발화 616개를 사용하였고, 테스트에는 약 5분 길이의 발화 1174개를 사용하였다. 또한 UBM 훈련용 DB로는 기존의 NIST SRE 2004 평가와는 달리 mixer 6 DB[18]를 사용하였으며, 그 중 GMM-UBM 방식의 화자인식에서는 약 10분 길이의 발화 500개를 사용하였고, JFA 방식의 화자인식에서는 약 10분 길이의 발화 1500개를 사용하여 1024개의 혼합(mixture) 수를 가지는 UBM을 구성하였다.

화자인식 실험에는 MFCC, LFCC, 그리고 warping factor 값을 0.1 간격으로 -0.1에서 +0.4까지를 사용한 BWFCC 특징들을 사용하였다. 이 특징들은 25 ms 크기의 Hamming 윈도우를 씌운 음성 프레임을 10 ms씩 이동시키면서 추출하였고, 추출된 켑스트럼 특징벡터는 에너지 정보를 포함한 20차 특징에 delta와 delta-delta 특징을 추가하여 총 60차의 특징벡터로 구성하였다. 특징 추출 후 에너지 기반의 음성검출(voice activity detection(VAD))을 적용하여 음성구간에 대해서만 화자인식 실험에 사용하였으며, 채널 왜곡을 보상하기 위해서 cepstral mean variance normalization(CMVN)을 적용하였다. 또한 추가적으로 화자인식 성능을 향상시키기 위해서 주파수-시간 특징을 켑스트럼-시간 행렬을 이용하여 추출한 후 기존에 주파수-시간 특징으로 사용했던 delta 특징을 대신해서 사용하여 화자인식 성능을 평가하였다.

GMM-UBM 방식과 JFA 방식 모두 다양한 스코어 정규화 방식을 사용하여 성능향상을 도모하였으며, 이때 사용된 방식은 t-norm, z-norm 및 zt-norm 방식이다[1].

성능평가 척도로는 사칭자를 등록자로 잘못 인식하는 오검출율(false alarm probability)과 등록자를 사칭자로 잘못 인식하는 누락율(miss probability)의 trade-off 관계를 표현하는 receiver operating characteristics(ROC) 곡선에서 이들 두 확률이 동일한 값을 가질 때의 오류확률인 equal error rate(EER)을 측정하여 사용하였다.

3.2 주파수 워핑 기반 특징에 따른 화자인식 결과

GMM-UBM 방식에서 주파수 워핑 기반 특징에 따른 화자인식 성능을 EER로 표현한 결과는 <표 1>과 같고, JFA 방식에서의 결과는 <표 2>와 같다. GMM-UBM 방식에서의 실험결과를 보면 일관성 있게 기존에 화자인식에 널리 사용되는 MFCC보다 LFCC의 성능이 우수함을 확인할 수 있고, 주파수 스케일을 적절히 워핑한 BWFCC가 LFCC보다도 더 나은 성능

을 나타내는 것을 알 수 있다. 그 중에서도 정규화 방식으로 t-norm을 사용하고, warping factor가 0.1인 BWFCC의 성능이 12.57%의 EER로 가장 우수한 성능을 나타내었다. JFA 방식에서의 실험결과를 보면 GMM-UBM 방식보다 화자인식 성능이 더 우수한 것을 볼 수 있고, GMM-UBM 방식에서와는 달리 MFCC가 LFCC보다 성능이 우수하지만, 주파수 스케일을 적절히 워핑한 BWFCC가 가장 우수한 성능을 나타낸다는 점은 동일하였다. JFA 방식의 경우 정규화 방식으로 z-norm을 사용하고, warping factor가 0.2일 때의 BWFCC의 성능이 9.64%의 EER로 가장 우수한 성능을 나타내었다.

3.3 주파수-시간 특징에 따른 화자인식 결과

주파수-시간 특징에 따른 화자인식 실험에서는 MFCC, LFCC, BWFCC의 정적인 특징 20차와 켑스트럼-시간 행렬(CTM)을 통해 추출한 주파수-시간 특징 40차를 합한 총 60차 특징과 기존 방식인 MFCC, LFCC, BWFCC의 정적인 특징 20차에 delta 및 delta-delta 특징 40차를 합한 총 60차 특징을 사용하여 화자인식 성능평가를 수행하였다. GMM-UBM 방식에서는 앞선 실험에서 성능이 가장 우수했던 스코어 정규화 방식인 t-norm을 적용하였고, JFA 방식에서는 역시 앞선 실험에서 성능이 가장 우수했던 z-norm을 적용하였다. 이 실험에 앞서 우선 GMM-UBM 방식에서 MFCC에 대한 TDCT 특징(180차원), MFCC + delta + delta-delta 특징(60차원), 본 논문에서 제안하는 MFCC + CTM 특징(60차원), 그리고 참고문헌 [15]와 같이 CTM만을 이용하는 특징(60차원)에 대해서 화자인식 성능비교를 수행하였고, 인식성능을 EER로 표현한 결과를 <표 3>에 나타내었다. 실험 결과 TDCT 특징은 MFCC + delta + delta-delta 특징 및 MFCC + CTM 특징과 비교해서 차원 수는 더 많음에도 불구하고 성능은 오히려 저조함을 확인할 수 있었고, CTM 특징만을 사용했을 경우에는 MFCC + delta + delta-delta 특징보다 정규화 이전(No norm.)에는 약간 우수하나 t-norm 적용시 오히려 성능이 약간 떨어지는 것을 확인할 수 있다. 이에 따라 이후 실험에서는 TDCT 특징과 CTM 특징만을 사용하는 경우는 제외하였다.

GMM-UBM 방식에서 주파수-시간 특징에 따른 화자인식 성능을 EER로 표현한 결과는 <표 4>와 같고, JFA 방식의 결과는 <표 5>와 같다. 실험 결과 두 방식 모두에서 기존 주파수-시간 특징으로 사용한 delta 특징보다 켑스트럼-시간 행렬을 통해 추출한 주파수-시간 특징을 사용했을 때 성능이 일관성 있게 더 우수하게 나옴을 확인할 수 있다.

GMM-UBM 방식에서 정규화 방식으로 t-norm을 사용하고, warping factor가 0.1일 때의 BWFCC + CTM 특징이 12.07%의 EER로 가장 좋은 성능을 나타내었고, JFA 방식에서는 정규화 방식으로 z-norm을 사용하고 warping factor가 0.1일 때의 BWFCC + CTM 특징이 9.26%의 EER로 가장 좋은 성능을 나타내었다.

표 1. GMM-UBM 방식에서의 주파수 워핑 기반 특징에 따른 화자인식 성능 (EER %)

Table 1. Speaker recognition performance according to frequency warping based features in GMM-UBM method (EER %)

Feature parameters		Score normalization method			
		No norm.	z-norm	t-norm	zt-norm
MFCC		16.68	16.73	15.59	15.81
LFCC		14.25	14.75	12.62	13.46
BWFCC	α=-0.1	14.67	15.21	13.51	13.70
	α=0.1	14.42	14.71	**12.57**	**12.95**
	α=0.2	**14.00**	**14.59**	12.84	13.18
	α=0.3	15.00	15.72	13.91	14.21
	α=0.4	16.60	16.64	15.42	15.67

표 2. JFA 방식에서의 주파수 워핑 기반 특징에 따른 화자인식 성능 (EER %)

Table 2. Speaker recognition performance according to frequency warping based features in JFA method (EER %)

Feature parameters		Score normalization method			
		No norm.	z-norm	t-norm	zt-norm
MFCC		10.83	10.39	11.59	10.98
LFCC		11.38	10.46	11.88	11.27
BWFCC	α=0.1	10.94	10.09	11.25	**10.56**
	α=0.2	**10.41**	**9.64**	**11.06**	**10.56**

표 3. 여러 가지 주파수-시간 특징에 따른 화자인식 성능 (EER %)

Table 3. Speaker recognition performance according to various spectro-temporal features (EER %)

Feature parameters	# dim.	Score normalization method	
		No norm.	t-norm
TDCT [9]	180	16.76	16.41
MFCC+△+△△	60	16.68	15.59
MFCC + CTM	60	**15.57**	**14.79**
CTM only [15]	60	16.32	16.26

표 4. GMM-UBM 방식에서의 주파수-시간 특징에 따른 화자인식 성능 (EER %)

Table 4. Speaker recognition performance according to spectro-temporal features in GMM-UBM method (EER %)

Feature parameters		Score normalization method	
		No norm.	t-norm
MFCC	+ △ + △△	16.68	15.59
	+ CTM	**15.57**	**14.79**
LFCC	+ △ + △△	14.25	12.62
	+ CTM	**13.33**	**12.43**
BWFCC (α=0.1)	+ △ + △△	14.42	12.57
	+ CTM	**13.26**	**12.07**
BWFCC (α=0.2)	+ △ + △△	14.00	12.84
	+ CTM	**13.22**	**12.43**

표 5. JFA 방식에서의 주파수-시간 특징에 따른 화자인식 성능 (EER %)
Table 5. Speaker recognition performance according to spectro-temporal features in JFA method (EER %)

Feature parameters		Score normalization method	
		No norm.	z-norm
MFCC	+ △ + △△	10.83	10.39
	+ CTM	**10.30**	**9.71**
LFCC	+ △ + △△	11.38	10.46
	+ CTM	**10.41**	**9.33**
BWFCC (α=0.1)	+ △ + △△	10.94	10.09
	+ CTM	**10.25**	**9.26**
BWFCC (α=0.2)	+ △ + △△	10.41	9.64
	+ CTM	**9.97**	**9.45**

4. 결론

본 논문에서는 기존의 화자인식에서 주로 사용된 특징인 MFCC와 모든 주파수를 동일하게 대우하는 LFCC, 그리고 bilinear 변환을 이용하여 warping factor에 따른 주파수 워핑을 적용한 BWFCC를 추출하여 화자인식 성능평가를 실시하였고, 또한 켑스트럼-시간 행렬을 이용한 주파수-시간 특징을 추출하여 켑스트럼 계수들의 delta 및 delta-delta 특징을 대신해서 사용하여 화자인식 성능을 평가하였다.

NIST SRE 2004 DB를 이용한 실험 결과, MFCC와 LFCC보다 주파수 스케일을 적절히 워핑한 BWFCC가 더 우수한 화자인식 성능을 나타내는 것을 확인하였고, 또한 켑스트럼-시간 행렬을 통해 추출한 주파수-시간 특징을 사용함으로써 delta 및 delta-delta 특징을 사용할 때 보다 화자인식 성능이 향상되는 것을 확인할 수 있었다.

참고문헌

[1] Kinnunen, T. & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun,* Vol. 52, No. 1, 12-40.

[2] Reynolds, D., Quatieri, T., Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Process,* Vol. 10, No. 1, 19-41.

[3] Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language,* Vol. 20, No. 2-3, 210-229.

[4] Kenny, P. (2006). Joint factor analysis of speaker and session variability: Theory and algorithms. http://www.crim.ca/perso/patrick.kenny/

[5] Senoussaoui, M., Kenny, P., Dehak, N., Dumouchel, P. (2010). An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. *Proc. Odyssey Speaker and Language Recognition Workshop,* 28-33.

[6] Davis, S., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech Signal Process,* Vol. 28, No. 4, 357-366.

[7] Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., Shamma, S. (2011). Linear versus mel frequency cepstral coefficients for speaker recognition. *Proc. ASRU Workshop,* 559-564.

[8] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoustics, Speech Signal Process,* Vol. 29, No. 2, 254-272.

[9] Kinnunen, T., Koh, C., Wang, L., Li, H., Chng, E. (2006). Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. *Proc. ISCSLP,* 547-558.

[10] Milner, B. P., Vaseghi, S. V. (1995). An analysis of cepstral-time feature matrices for noise and channel robust speech recognition. *Proc. Eurospeech,* 519-522.

[11] Stevens, S., Volkman, J., Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America,* Vol. 8, No. 3, 185-190.

[12] Wolfel, M., McDonough, J., Waibel, A. (2003). Warping and scaling of the minimum variance distortionless response. *Proc. ASRU Workshop,* 387-392.

[13] Choi, Y. H., Ban, S. M., Lee, G. H., Kim, K. H. Kim, H. S. (2014). Performance comparison of different frequency scales in feature extraction for speaker recognition. *Proceedings of 2014 Fall Conference of Korean Society of Speech Sciences,* 195-196. (최영호, 반성민, 이가희, 김경화, 김형순 (2014). 화자인식 특징추출을 위한 주파수 스케일 성능 비교. *2014 한국음성학회 가을 학술대회 발표 논문집,* 195-196.)

[14] Kumar, P., Rao, P. (2004). A study of frequency-scale warping for speaker recognition. *Proc. NCC 2004,* 203–207.

[15] Zhang, W. Q., Deng, Y., He, L., Liu, J. (2010). Variant time-frequency cepstral features for speaker recognition. *Proc. Interspeech,* 2122–2125.

[16] Larcher, A., Bonastre, J. F., Fauve, B., Lee, K. A., Lévy, C., Li, H., Mason, J. S., Parfait, J. Y. (2013). ALIZE 3.0 – open source toolkit for state-of-the-art speaker recognition. *Proc. Interspeech,* 2768–2773.

[17] The evaluation plan of NIST 2004 speaker recognition evaluation campaign. http://www.itl.nist.gov/iad/mig/tests/spk

/2004/SRE-04_evalplan-v1a.pdf.

[18] Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., Neely, A. (2010). The mixer 6 corpus: Resources for cross-channel and text independent speaker recognition. *Proc. LREC 2010,* 2441-2444.

• **최영호 (Choi, Young Ho)**
부산대학교 전자전기컴퓨터공학과
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-1704 Fax: 051-510-4279
Email: choiyh@pusan.ac.kr
관심분야: 음성합성, 화자인식

• **반성민 (Ban, Sung Min)**
부산대학교 전자전기컴퓨터공학과
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-1704 Fax: 051-510-4279
Email: bansungmin@pusan.ac.kr
관심분야: 음성인식, 음성 전처리

• **김경화 (Kim, Kyung-Wha)**
대검찰청 과학수사담당관실
서울시 서초구 반포대로 157
Tel: 02-3480-2150 Fax: 02-3480-2707
Email: savoix@spo.go.kr
관심분야: 법음성학, 화자식별

• **김형순 (Kim, Hyung Soon)** 교신저자
부산대학교 전자공학과
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-2452
Email: kimhs@pusan.ac.kr
관심분야: 음성인식, 음성합성, 음성신호처리