켑스트럼 거리 기반의 음성/음악 판별 성능 향상*

박슬한(부산대), 최무열(부산대), 김형순(부산대)

<차 례>

<Abstract>

Performance Improvement of Speech/Music Discrimination Based on Cepstral Distance

Seul-Han Park, Mu Yeol Choi, Hyung Soon Kim

Discrimination between speech and music is important in many multimedia applications. In this paper, focusing on the spectral change characteristics of speech and music, we propose a new method of speech/music discrimination based on cepstral distance. Instead of using cepstral distance between the frames with fixed interval, the minimum of cepstral distances among neighbor frames is employed to increase discriminability between fast changing music and speech. And, to prevent misclassification of speech segments including short pause into music, short pause segments are excluded from computing cepstral distance. The experimental results show that proposed method yields the error rate reduction of 68%, in comparison with the conventional approach using cepstral distance.

* Keywords: Speech/music discrimination, Cepstral distance.

* 이 논문은 산업자원부 지원으로 수행하는 21세기 프론티어 연구개발사업(인간기능 생활지원 지능로봇 기술개발사업)의 일환으로 수행됨.

1. 서 론

음성과 음악을 자동으로 구별하는 시스템은 여러 응용 분야에서 유용하게 활용될 수 있다. 예를 들어, 오디오 데이터를 자동 인덱싱하거나 멀티미디어 정보를 검색하는 시스템, 그리고 오디오 압축 시 음성과 음악 각각에 적합한 압축방식을 적용하는데 사용될 수 있다. 또한 방송뉴스 인식 시스템에서는 배경음악 구간을 인식기의 입력에서 제외시킴으로 인식성능을 향상시키는데 기여할 수 있다.

우수한 성능의 음성/음악 판별 시스템을 위해서는 특징 파라미터를 선택하는 문제와 어떤 분류 방법을 이용할 것인지가 중요한데, 그 중에서도 효과적인 특징 파라미터들을 사용하는 것이 성능 향상에 큰 영향을 미친다. 이에 따라 시간과 주파수 영역에서 음성과 음악의 특성 차이를 이용한 다양한 파라미터들이 제안되어 왔다[1]-[5][8][9]. 시간 영역에서의 특징을 이용한 파라미터로는 영교차율(zero crossing rate(ZCR))과 그 변화 특성을 이용한 high ZCR ratio(HZCRR)[1][2], 음성이 음악에 비해 휴지 구간을 많이 포함하고 있음을 이용하여 신호의 단구간의 에너지의 변화를 측정한 low short-time energy ratio(LSTER)[3] 등이 있다. 그리고 스펙트럼 영역 특징을 이용한 파라미터로 스펙트럼의 중심을 이용한 spectral centroid, 스펙트럼의 변화 특성의 차이를 이용한 spectral flux(SF)[1], 켑스트럼 거리(cepstral distance(CD))[8] 및 cepstrum flux(CF)[9], 그리고 음성의 스펙트럼 포락선에 대한 정보를 잘 나타내는 line spectrum pair(LSP)[4] 등이 있다. 그 외에도 음성의 특징을 잘 반영하는 4 Hz modulation energy, 음악의 리듬을 이용한 pulse metric, 그리고 음소인식 결과를 기반으로 하는 entropy와 dynamism[5] 등의 다양한 특징 파라미터들이 제안되었다.

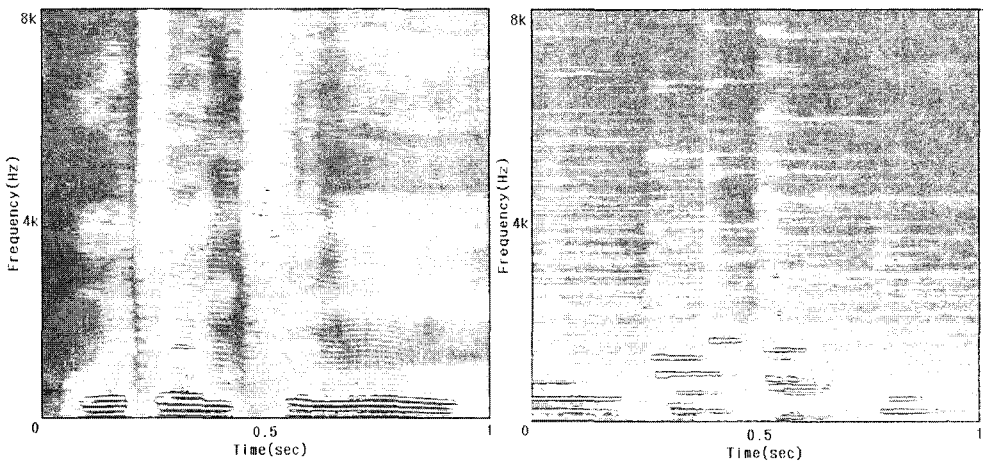
본 논문에서는 기존에 제안된 파라미터들 중에서 시간에 따른 스펙트럼의 변화 특성을 이용하여 우수한 음성/음악 판별 성능을 갖는 켑스트럼 거리의 평균을 이용하는 방식[8]을 기반으로 하되, 이 방식에서 음성과 빠르게 변화하는 음악 사이의 변별 성능이 떨어지는 문제를 극복하기 위해, 고정된 프레임 간격으로 CD를 구하는 대신에 일정 범위의 여러 프레임 사이의 CD의 최소값을 사용하는 방식을 제안한다. 또한 음성이 휴지 구간을 많이 포함하고 있음에 기인해 에너지를 이용하여 간단히 휴지 구간을 추정하고 그 구간을 CD의 평균을 구할 때 제외함으로써 판별 성능을 추가적으로 향상시키는 방식에 대해서도 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 스펙트럼 변화를 이용한 특징 파라미터에 대해 설명하고, 3장에서는 본 논문에서 기반으로 한 켑스트럼 거리의 문제점과 이를 개선한 제안 방식에 대해서 설명한다. 4장에서는 실험에 사용된 데이터 베이스에 대해 언급한 후 기존 방식과 제안 방식에 의한 실험 내용 및 결과를 기술하며, 마지막으로 5장에서 결론을 맺는다.

2. 스펙트럼 변화를 이용한 기존의 파라미터

2.1. 음성과 음악의 스펙트로그램 상에서의 특징

음성은 평균적으로 1초에 4음절 정도 발성하는 것으로 알려져 있다[6]. 또한 음성은 모음의 공명주파수를 나타내는 포먼트가 스펙트로그램 상에서 잘 나타나며, 빈번하게 나타나는 음소 전이 구간에서 포먼트의 전이 현상이 관찰된다. 음성은 대개 유성음과 무성음이 번갈아 나타나므로 짧은 구간에서도 스펙트럼 포락선의 변화가 자주 일어난다. 이에 비해 음악의 경우 동일한 악기 그룹의 연주가 계속되는 동안에는 빠른 템포로 음정 변환이 진행되더라도 스펙트럼 포락선은 비교적 유사한 형태를 유지한다. <그림 1>은 음성과 음악의 전형적인 스펙트로그램을 보여준다. <그림 1>의 (a)는 음성의 스펙트럼이 1초 동안 어떻게 변하는지를 보여준다. 그림에서 보듯이 음성 구간에서는 짧은 구간에서도 스펙트럼 포락선의 변화가 자주 일어난다. 동일한 시간 동안에 음악의 스펙트럼 변화를 나타내는 <그림 1>의 (b)에서 음악은 여러 프레임을 이동하여 스펙트럼 포락선을 비교해 보아도 음성에 비해 상대적으로 비슷한 모양을 가진다.



(a) 음성 (b) 음악

<그림 1> 음성과 음악의 스펙트로그램 예

2.2. 시간에 따른 스펙트럼 변화를 이용한 파라미터

시간에 따른 음성과 음악의 스펙트럼 변화의 차이를 이용하여 제안된 SF는 인접한 프레임간의 스펙트럼 변화의 크기를 나타내며, 다음 식과 같이 정의된다.

$$SF = \frac{1}{(N-1)F}\sum_{n=1}^{N-1}\sum_{f=1}^{F}[\log(A(n,f)+\delta)-\log(A(n-1,f)+\delta)]^2 \quad (1)$$

여기서 A(n,f)는 n번째 입력 프레임의 f번째 bin에 대한 DFT 크기 값이다. 그리고, N은 SF를 구하는 기본 시간 단위 안에 포함되는 총 프레임 수이고, F는 DFT point 수, δ는 계산상 log 함수에 영(0)이 들어가는 것을 막기 위한 작은 상수이다. 앞서 언급했듯이 음성의 경우 프레임 간 스펙트럼의 변화가 크기 때문에 음악에 비해 SF의 값이 큰 쪽에 분포하게 된다[1]. 그러나 SF는 프레임간의 변화를 측정할 때 모든 DFT point의 값들을 비교하기 때문에, 피치 하모닉 성분 등 음정 변화에 따른 세밀한 스펙트럼 변화에 민감한 문제점이 있다.

스펙트럼 변화를 이용한 또 다른 파라미터는 스펙트럼 포락선(spectral envelope)을 표현하는 켑스트럼의 저차 성분들을 이용한 켑스트럼 거리(cepstral distance(CD))이다. CD는 스펙트럼 포락선의 변화 특성만을 비교하기 때문에 SF보다 더 나은 성능을 갖는다. 켑스트럼 거리를 구하는 식은 다음과 같다.

$$CD(n) = \sqrt{\sum_{k=1}^{K}(c(n+d,k)-c(n,k))^2} \quad (2)$$

여기서 CD(n)은 n번째 프레임에서의 CD 값이고, K는 켑스트럼 차수, c(n,k)는 n번째 입력 프레임에 대한 k차 켑스트럼 계수 값을 나타낸다. d는 CD 계산 시 얼마나 떨어진 프레임과 비교할 것인지를 가리킨다.

[9]에서는 두 프레임 간의 CD를 그대로 사용하지 않고 다음 식과 같이 일정 구간 동안 CD 값들의 평균과 분산을 음성/음악 판별을 위한 특징 파라미터로 사용하였다.

$$\mu_{CD} = \frac{1}{N-d}\sum_{n=1}^{N-d} CD(n)$$
$$\sigma_{CD}^2 = \frac{1}{N-d}\sum_{n=1}^{N-d}(CD(n)-\mu_{CD})^2 \quad (3)$$

여기서 N은 평균과 분산을 구하기 위한 일정한 구간 내에 있는 프레임 수이다. CD의 평균은 음악보다 음성에서 큰 값을 가지며, 음악의 경우 음성에 비해 스펙트럼 포락선의 변화가 작기 때문에 CD의 분산도 음악보다 음성에서 큰 값을 가진다.

CD는 일정한 간격만큼 떨어진 한 개의 프레임과 켑스트럼을 비교하는 반면에

[8]에서 제안된 cepstrum flux(CF)는 인접한 프레임에서부터 다소 멀리 떨어진 프레임까지 여러 프레임의 켑스트럼과 비교하였다. CF는 여러 프레임과 비교하여 구한 켑스트럼 거리들의 평균으로 정의된다.

$$D(n)=\frac{1}{J}\sum_{j=1}^{J}\sqrt{\sum_{k=1}^{K}(c(n+j,k)-c(n,k))^2} \tag{4}$$

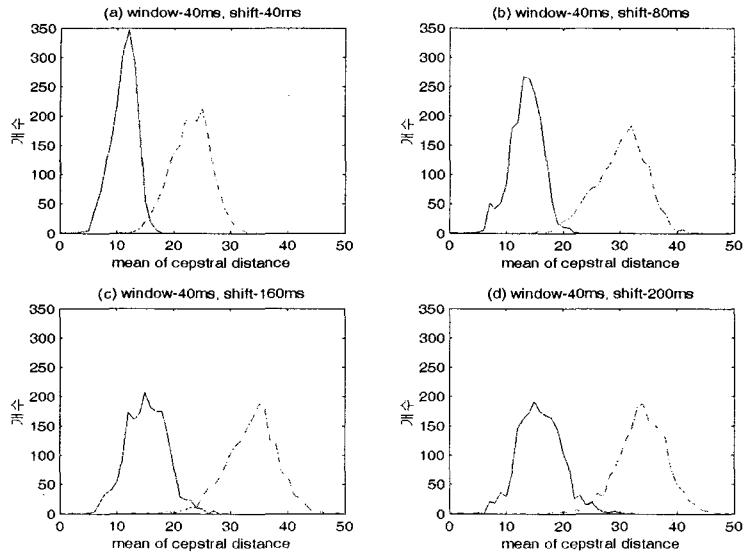
여기서 D(n)은 n번째 프레임에서의 CF 값이고, J는 켑스트럼을 비교할 프레임의 수, K는 켑스트럼 차수이며, c(n,k)은 n번째 입력 프레임에 대한 k차 LPC 켑스트럼 계수 값을 나타낸다. 음성에서 스펙트럼의 변화가 더 크기 때문에 LPC 켑스트럼의 변화를 이용한 CF는 음악보다 음성에서 큰 값을 가진다.

CF를 기반하여 보다 성능을 개선시키기 위해 Block Cepstrum Flux(BCF)가 제안되었다[8]. BCF는 일정한 구간 내에서 CF의 평균을 구한 것이다.

$$B(n)=\frac{1}{W}\sum_{i=0}^{W-1}D(n-i) \tag{5}$$

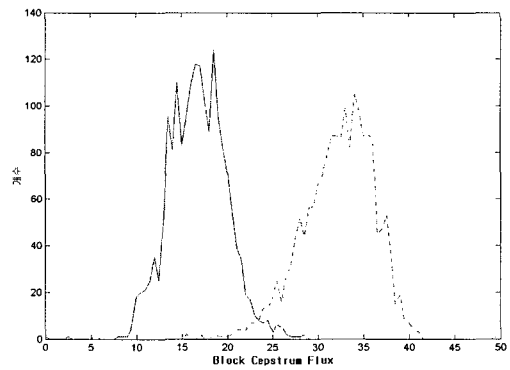
여기서 W는 CF의 평균을 구하기 위한 시간 블록에 속하는 프레임의 개수이다. 그리고 B(n)은 n번째 시간에서 구한 BCF의 값이다. BCF의 값도 음악보다 음성에서 큰 값을 가진다.

CD의 평균에 대한 히스토그램을 <그림 2>에 나타내었다. 켑스트럼으로는 음성인식에 주로 사용되는 파라미터인 Mel-Frequency Cepstrum Coefficient (MFCC)를 사용하였다. 이 때 켑스트럼을 구하는 윈도우의 크기는 40msec로 고정하였고 켑스트럼을 비교하는 프레임과의 거리를 40msec, 80msec, 160msec, 200msec로 변경하여 히스토그램을 구하였다. 이미 언급한 것처럼 실선으로 표시된 음악은 스펙트럼 포락선의 변화가 작기 때문에 CD의 평균이 작은 값들에 분포하며 그 분산 또한 작은 값을 갖는다. 음성은 CD의 평균이 넓게 분포하여 스펙트럼 포락선의 변화가 크다는 것을 알 수 있다. 육안으로 관찰하기 쉽지는 않지만, <그림 2>의 (a), (b), (c)를 통해 CD를 구하는 프레임 사이의 거리가 멀어질수록 음성과 음악의 겹치는 부분이 점점 적어지는 경향을 보인다. 이는 음성의 경우 CD를 구하는 프레임 간의 간격이 길어질수록 스펙트럼이 크게 변하는 반면, 음악은 음성에 비해 상대적으로 작게 변하기 때문이다. 켑스트럼을 비교하는 프레임 사이의 거리가 너무 멀면 (d)처럼 오히려 겹치는 부분이 증가되는데, 이는 음악의 경우에도 CD를 구하는 시간 간격이 멀어지면 스펙트럼의 변화가 커지기 때문이다. 따라서 적절한 거리에서 켑스트럼을 비교하는 것이 효과적이다.



<그림 2> CD의 평균의 히스토그램
(실선 : 음악, 점선 : 음성)

<그림 3>은 BCF의 히스토그램이다. MFCC로부터 구한 CD의 평균과 비교하기 위해 BCF의 경우에도 LPC 켑스트럼 대신에 MFCC를 사용하였다. 인접한 프레임에서부터 300msec 떨어진 프레임까지 비교하여 그 평균으로 CF를 구했고, 1초 구간 동안 구한 CF들의 평균을 취해 BCF를 구했다. CF를 구할 때 짧은 거리에서부터 비교적 먼 거리의 프레임까지 비교 프레임의 개수를 늘려가며 비교하여 보았지만 이에 따라 음성/음악 판별 성능은 별로 변하지 않았다. 또한 CD의 평균과 비교했을 때 판별 성능은 비교 프레임 개수에 따라 비슷하거나 조금 못하였기 때문에 이후의 실험에서는 적용하지 않았다.



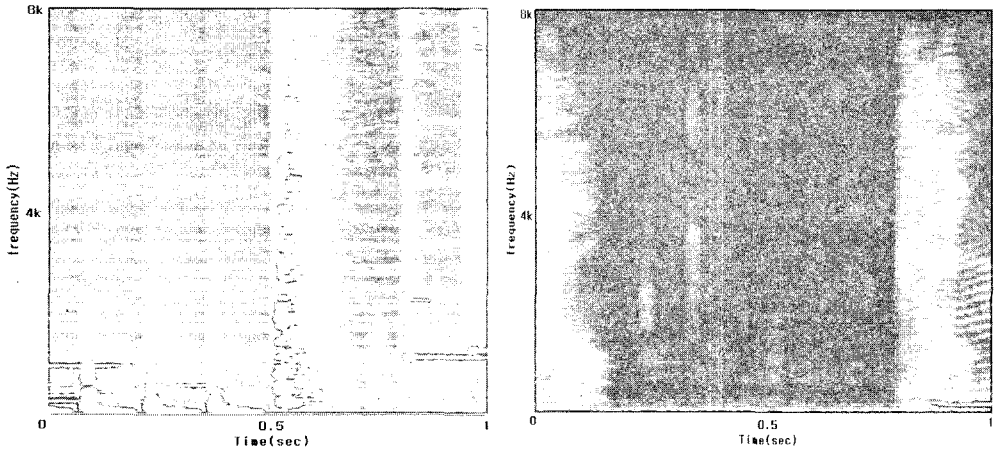
<그림 3> BCF의 히스토그램
(실선 : 음악, 점선 : 음성)

3. 켑스트럼 거리 기반의 제안 방식

3.1. 기존 켑스트럼 거리의 문제점

앞서 설명한 바와 같이 음악에 비해 음성의 CD가 더 큰 값을 가지는 경향이 있다. 그런데 음악의 경우 유난히 빠른 템포이거나 새로운 악기가 짧은 시간 여러 차례 등장하게 되면 스펙트럼의 변화가 자주 일어나고, 이에 따라 CD가 커지게 된다. 이런 음악은 일정한 프레임 간격만큼 떨어진 프레임과 비교하여 구한 CD의 평균을 파라미터로 이용하면 음성으로 오인식될 가능성이 높아진다. <그림 4>의 (a)는 변화가 심한 음악의 스펙트로그램의 예를 보여준다.

그리고 음성은 음악에 비해 휴지 구간을 많이 포함하고 있는데 이 구간은 변화가 거의 없는 구간이기 때문에 매우 작은 CD 값을 가진다. CD의 평균을 구하는 일정 구간(1초) 내에 휴지 구간의 비율이 어느 이상 된다면 음성의 CD 평균이 상대적으로 작은 값을 갖게 되고, 음성이 음악으로 오인식되는 결과를 초래할 수 있다. 긴 휴지 구간을 포함하는 음성의 스펙트로그램의 예가 <그림 4>의 (b)에 나타나 있다.



(a) 시간에 따른 변화가 심한 음악 (b) 휴지 구간을 많이 포함하는 음성

<그림 4> 켑스트럼 거리로 판별할 경우 오인식되는 스펙트로그램의 예

3.2. 제안 방식

음악의 스펙트럼의 변화가 심할 때 큰 CD 값을 가지는 문제점을 해결하기 위해, 본 논문에서는 현재 프레임에서 일정한 간격만큼 떨어진 프레임과 비교하여

CD를 구하는 대신에, 인접한 여러 프레임들과 비교하여 구한 여러 개의 CD 값들 중에서 최소값을 그 프레임에서의 CD 값으로 선택하는 방법을 제안한다. 즉, 인근의 프레임 중에서 현재 프레임과 가장 스펙트럼 포락선이 닮은 CD 값을 이용하겠다는 것이다. 최소값을 이용한 CD와 그 평균은 다음과 같이 정의된다.

$$MCD(n) = \min_{d_1 \le d \le d_2} \left[\sqrt{\sum_{k=1}^{K} (c(n+d,k) - c(n,k))^2} \right] \tag{6}$$

$$\mu_{MCD} = \frac{1}{N-d_2} \sum_{n=1}^{N-d_2} MCD(n) \tag{7}$$

여기서 MCD(n)은 n번째 프레임과 인접한 프레임들과의 CD 값의 최소값이고, d1과 d2는 비교대상 프레임의 범위를 나타낸다. 즉, d1-1은 매우 인접하여 비교대상에서 제외되는 프레임 수를 의미한다.

이와 같이 CD의 최소값을 이용하면 음악의 경우 빠른 템포나 새로운 악기가 등장하여 스펙트럼의 변화가 크더라도 리듬이 있기 때문에 인접한 프레임들 중에 동일한 패턴이 있는 곳(스펙트럼 포락선이 가장 닮은 곳)을 찾게 되고, 그 때의 CD 값을 이용함으로써 변화가 심한 음악이라 할지라도 CD 값이 줄어들어 음성보다 작은 값을 가지게 된다. 그러나 음성은 이러한 특성이 없기 때문에 여러 프레임을 비교하더라도 일반적으로 가장 인접한 프레임이 가장 닮은 프레임인 경우가 대부분이다. 또한 음성의 경우에도 바로 인접한 프레임에서는 스펙트럼 포락선이 비슷할 수 있으므로, 매우 인접한 프레임들은 제외시킴으로써(d1 > 1) 음성과 음악 사이의 변별력을 높이도록 하였다.

음성에 많이 존재하는 휴지 구간은 음성의 CD 값을 감소시키지만, 이 구간은 에너지만으로도 간단하게 찾을 수 있다. 그러므로, 보다 나은 음성/음악 판별 성능을 위해 에너지를 통해 휴지 구간을 찾은 후 CD의 평균을 구할 때 이들 구간을 제외시킴으로써, 휴지 구간으로 인해 낮게 구해진 음성의 CD 값을 높이도록 하였다.

4. 실험 및 결과

4.1. 실험환경

본 논문에서 제안한 방법의 성능 평가를 위해 [7] 및 [9]의 결과를 베이스라인으로 하였다. 이 때 훈련을 위해 사용된 음성 DB는 국어공학센터에서 구축한 PBS 589 문장에 대한 남녀 50명분의 발성 데이터 약 13시간 분량을 사용하였다. 그리

고 음악 DB는 클래식 음악 음반으로부터 42곡을 16kHz로 다운샘플링하고 16bit로 양자화하여 음성 DB와 동일한 조건이 되도록 만들었다. 테스트 DB는 훈련 시 사용되지 않고 clean 환경에서 녹음한 음성 데이터와 다양한 장르의 음악을 15초씩 번갈아 나타나도록 구성하였다. 여기서 사용한 음악은 클래식은 아니지만 사람의 목소리가 포함되지 않는 연주곡으로 구성하였다.

음성판별을 위한 분류기로는 주로 Gaussian Mixture Model(GMM), Hidden Markov Model(HMM), k-Nearest Neighbors(k-NN) 분류기 등이 있는데, 이전 연구에 따르면 그 성능은 거의 비슷한 것으로 알려져 있다[7]. 본 논문에서는 특징 벡터의 분포를 몇 개의 Gaussian 분포들의 가중합으로 표현하는 GMM 분류기를 사용하였으며 GMM에서의 mixture 수를 늘려가면서 실험을 수행하였다.

켑스트럼 거리 측정을 위해 음성인식에 널리 사용되는 MFCC를 이용하여 CD를 계산하였고, 이 때 프레임의 크기는 25ms, shift size는 10ms로 하였다. 앞서 언급한 것처럼 1초 구간에서 CD의 평균을 파라미터로 사용했다.

4.2. 실험결과

2.2절에서 본 음성과 음악의 히스토그램은 음성과 음악의 분포가 한 개의 Gaussian 함수만으로도 충분히 잘 모델링될 수 있음을 알 수 있다. 그러므로 mixture를 늘려가며 한 실험에서는 CD의 평균과 분산을 이용할 때 성능이 비슷하거나 over-estimation에 의해 오히려 성능이 조금 떨어지는 결과를 보인다. 그러나 다른 파라미터들의 경우 mixture를 늘림에 따라 조금 더 높은 성능을 보이는 것도 있다.

일정한 프레임 간격으로 구한 CD 값의 평균을 이용하였을 때와 여러 프레임 중 최소값을 이용하였을 때의 음성/음악 판별성능 결과를 <표 1>과 <표 2>에 나타내었다. <표 1>에서 간격은 식 (2)에서의 d를 나타내고, <표 2>에서 범위는 식 (6)에서의 d1과 d2를 나타낸다.

일정한 간격으로 CD를 구한 방법의 분류 결과 중 가까운 프레임을 비교한 것보다 140ms와 같이 여러 프레임 더 떨어진 프레임과 비교한 것이 성능이 더 높게 나타났다. 이것은 음성의 경우에도 매우 가까운 프레임 사이에서는 스펙트럼 포락선이 비슷하기 때문이다. 이 결과는 본 논문에서 제안한 방법, 즉 CD의 최소값을 이용할 때도 나타나는데, <표 2>에서 바로 인접한 프레임과 비교한 CD를 포함하여 최소값을 구한 것보다는 인접한 프레임은 배제하고 여러 프레임 떨어진 프레임들만 비교한 것이 성능 면에서 더 우수하였다. <표 1>과 <표 2>를 비교해 볼 때, 본 논문에서 제안한 CD의 최소값을 이용하는 방법이 기존 방식에 비해 성능이 개선되는 것을 확인할 수 있었다.

<표 1> 일정 간격의 CD를 이용한 분류 결과(%)

간격 / mix #	10ms	30ms	60ms	120ms	140ms	160ms
1	94.25	97.47	97.42	97.44	97.97	97.69
2	93.39	97.03	97.42	97.36	**98.00**	97.50

<표 2> CD의 최소값을 이용한 분류 결과(%)

범위 (ms) / mix #	10 ~ 150	30 ~ 150	50 ~ 150	10 ~ 250	30 ~ 250	50 ~ 250
1	95.64	97.78	97.47	95.39	**99.20**	98.00
2	95.44	97.61	97.64	94.92	99.08	97.92

<표 3>과 <표 4>는 에너지로부터 간단히 추정된 휴지 구간을 CD의 평균 계산 시 제외하여 실험한 결과이다. <표 1>과 <표 3>을 비교해 보면 휴지 구간을 제외한 것만으로도 기존 방법에서 최고성능이 개선되었다. CD의 최소값을 사용하고 휴지구간을 제외함으로써 더 많은 성능 향상을 얻었고, <표 1>과 <표 4>의 최고 성능을 비교할 때 오류율이 2.00%에서 0.64%로 줄어들어 68%의 오류감소율을 얻을 수 있었다. 가장 우수한 성능을 얻기 위해 최소값을 선택할 프레임의 범위를 어떻게 선정할 것인가에 관해서는 앞으로 추가 연구가 더 필요하다고 판단된다.

<표 3> 일정 간격의 CD를 이용한 분류 결과(%, silence 제외)

간격 / mix#	10ms	30ms	60ms	120ms	140ms	160ms
1	89.83	97.75	97.92	98.06	98.28	97.83
2	89.89	96.86	97.92	98.00	**98.39**	97.53

<표 4> CD의 최소값을 이용한 분류 결과(%, silence 제외)

범위 (ms) / mix#	10 ~ 150	30 ~ 150	50 ~ 150	10 ~ 250	30 ~ 250	50 ~ 250
1	95.36	**99.36**	99.11	94.78	98.50	98.94
2	95.31	98.78	99.11	94.56	98.36	98.89

마지막으로 여러 가지 특징 파라미터에 의한 음성/음악 판별 성능의 비교결과

를 <표 5>에 나타내었다. 이 표에서 Hn_Dn은 기존의 방법 중 우수한 성능을 나타내는 것으로 알려진 entropy와 dynamism을 함께 사용한 것이고[5][7], μ_{CD}는 일정 간격으로 구한 CD의 평균 파라미터[9], μ_{MCD}는 본 논문에서 제안한 최소값을 이용한 CD의 평균 파라미터이다. SF와 Hn_DN은 shfit size를 25ms로 하여 추출하였다. 표에서 보는 바와 같이 본 논문에서 제안한 CD의 최소값을 이용한 방식이 기존의 SF나 entropy와 dynamism보다도 더 높은 성능을 나타내었다.

<표 5> 다양한 파라미터에 의한 음성/음악 판별 성능 비교(%)

파라미터	SF	Hn_Dn	μ_{CD}	μ_{MCD}	μ_{MCD} (silence 제외)
인식률	91.58	95.08	98.00	99.20	**99.36**

5. 결 론

본 논문에서는 음성과 빠르게 변화하는 음악의 판별 성능을 높이기 위해 인접한 프레임들의 켑스트럼 거리 중 최소값을 이용하는 방식을 제안하였고, 이와 더불어 켑스트럼 거리 계산 시 휴지구간을 제외함으로써 휴지기간이 많이 포함된 음성을 음악과 구별하는 성능을 높이는 방식을 제안하였다. 실험 결과 기존의 여러 특징 파라미터에 비해 제안된 방식이 높은 성능 향상을 보였으며, 일정한 간격으로 구했던 켑스트럼 거리와 비교해 볼 때, 68%의 오류감소율을 얻었다. 제안된 방식은 음성인식에 널리 사용되는 MFCC를 이용하기 때문에 음성/음악 판별에 뒤이어 음성인식에 적용하는데 유리한 장점이 있다. 음성신호에 배경음악이 섞여있는 오디오 신호로부터 제안된 방식을 이용하여 음성 구간을 검출해 내는 방법에 대한 연구가 현재 진행 중에 있다.

참 고 문 헌

[1] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature music/speech discrimination", *Proc. ICASSP97*, Vol.2, pp. 1331-1334, 1997.

[2] L. Lu, H. Jiang, H. J. Zhang, "A robust audio classification and segmentation method", *Proc. 9th ACM Multimedia*, pp.203-211, 2001.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music", *Proc. ICASSP96*, Vol.2, pp.993-996, 1996.

[4] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, "Speech/music discrimination for multimedia application", *Proc. ICASSP00*, Vol.4, pp.2445-2449, 2000.

[5] J. Ajmera, I. McCowan, H. Bourlard, "Speech/music discrimination using entropy and dynamism features in a HMM classification framework", *Speech Communication*, Vol. 40, Issue 3, pp.259-430, 2003.

[6] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acustica*, Vol.28, pp.66-73, 1973.

[7] 김수미, 김형순, "음성/음악 판별을 위한 특징 파라미터와 분류기의 성능 비교", *말소리*, Vol. 46, pp.37-50, 2003.

[8] S. Takeuchi, T. Uchida et al., "Optimization of voice/music detection in sound data", *CRAC workshop*, 2001.

[9] 박슬한, 김형순, "켑스트럼 거리를 이용한 음성/음악 판별 성능 향상", *제 18회 신호처리 합동학술대회 논문집*, Vol .18, no. 1, pp.1, 2005.

접수일자 : 2005년 11월 28일
게재결정 : 2005년 12월 14일

▶ 박슬한(Seul-Han Park)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: seulhany@pusan.ac.kr

▶ 최무열(Mu Yeol Choi)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 516-4279
E-mail: mychois@pusan.ac.kr

▶ 김형순(Hyung Soon Kim) : 교신저자
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr