

# 다차원 MMCD를 이용한 음성/음악 판별

최무열, 송화전, 박슬한, 김형순  
부산대학교 전자공학과

## Speech/Music Discrimination Using Multi-dimensional MMCD

Mu-yeol Choi, Hwa Jeon Song, Seulhan Park, Hyung Soon Kim  
Department of Electronics Engineering, Pusan National University  
E-mail : mychois@pusan.ac.kr

### Abstract

Discrimination between speech and music is important in many multimedia applications. Previously we proposed a new parameter for speech/music discrimination, the mean of minimum cepstral distances (MMCD), and it outperformed the conventional parameters. One weakness of it is that its performance depends on range of candidate frames to compute the minimum cepstral distance, which requires the optimal selection of the range experimentally. In this paper, to alleviate the problem, we propose a multi-dimensional MMCD parameter which consists of multiple MMCDs with different ranges of candidate frames. Experimental results show that the multi-dimensional MMCD parameter yields an error rate reduction of 22.5% compared with the optimally chosen one-dimensional MMCD parameter.

### I. 서론

디지털 음향매체의 발전과 더불어 오디오 신호를 저장하거나 다루는 일이 늘어남에 따라, 음성과 음악을 자동으로 구별하는 시스템은 여러 분야에서 유용하게 활용된다. 방대한 오디오 데이터에서 원하는 음악과 음성을 검색해 주는 멀티미디어 정보검색, 음성인식

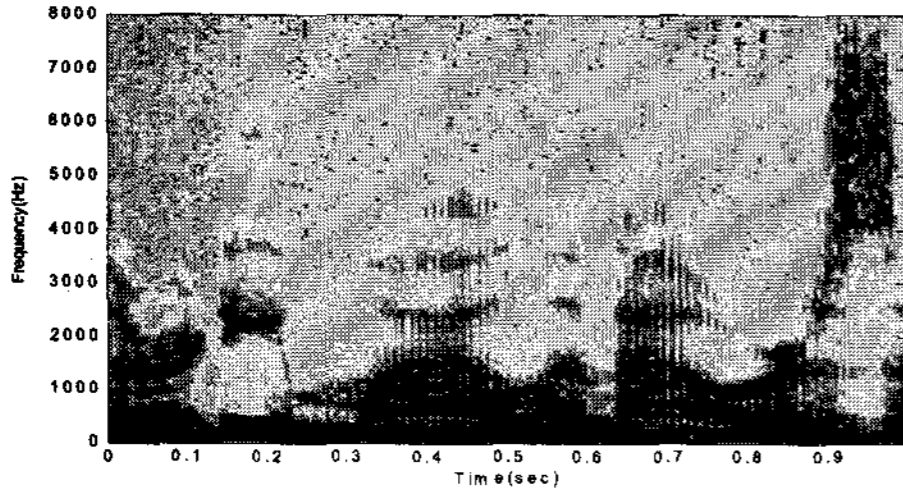
전처리 단계에서 음악 부분을 제외시키는 과정, 그리고 데이터 전송 시 음성과 음악에 적합한 압축방식의 적용 등이 그 예이다. 우수한 음성/음악 판별 성능을 얻기 위해서는 적절한 특징 파라미터를 선택하는 것이 매우 중요하며, 이에 따라 시간 및 주파수 영역에서 음성과 음악의 특성 차이를 나타내는 다양한 파라미터들이 제안되어 왔다[1]-[7].

이들 파라미터 중에서 인접한 여러 프레임들 사이의 캡스트럼 거리의 최소값의 평균을 이용하는 Mean of Minimum Cepstral Distances (MMCD) 파라미터가 성능 면에서 가장 우수한 것으로 보고되었다[7]. 그러나, MMCD의 경우 캡스트럼 거리의 최소값을 구하는 프레임 범위가 성능에 상당한 영향을 미치며, 이 범위를 실험을 통해 최적화해야 하는 단점이 있다. 이 문제의 해결을 위해 본 논문에서는 복수의 프레임 범위에 대한 MMCD들을 함께 사용하는 다차원 MMCD를 제안하며, 실험 결과 앞서 언급한 단점을 해결함과 동시에 추가적인 성능 개선을 얻을 수 있었다.

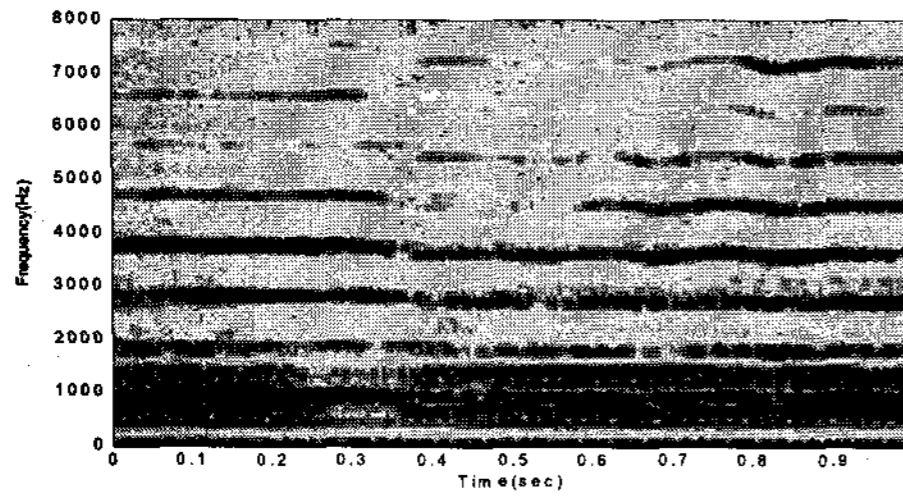
본 논문의 구성은 다음과 같다. 2장에서는 기존의 MMCD 파라미터에 대해 설명하고, 3장에서는 MMCD의 문제점을 개선하기 위한 방안으로 다차원 MMCD 파라미터를 제안한다. 4장에서는 실험 방법 및 결과를 기술하며, 마지막으로 5장에서 결론을 맺는다.

### II. Mean of Minimum Cepstral Distances (MMCD) 파라미터

음성은 다양한 음소들이 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. 이와 같은 특성은 그림 1의 음성과 음악의 스펙트로그램에 잘 나타나있다.



(a) 음성



(b) 음악

그림 1. 음성과 음악의 스펙트로그램 예

신호의 스펙트럼 변화 특성을 이용한 MMCD 파라미터를 설명하기 위해, 우선 스펙트럼 포락선(spectral envelope)을 표현하는 캡스트럼의 저차 성분들을 사용하여 다음 식과 같은 캡스트럼 거리를 정의한다.

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

여기서  $CD(n)$ 은  $n$ 번째 프레임에서의 캡스트럼 거리이고,  $K$ 는 캡스트럼 차수,  $c(n,k)$ 는  $n$ 번째 입력 프레임에 대한  $k$ 차 캡스트럼 계수 값을 나타낸다. [6]에서는 두 프레임 간의 캡스트럼 거리를 그대로 사용하지 않고 다음 식과 같이 일정 구간 동안 캡스트럼 거리의 평균을 음성/음악 판별을 위한 특징 파라미터로 사용하였다.

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

여기서  $N$ 은 평균을 구하기 위한 일정한 구간 내에 있는 프레임 수이다.

캡스트럼 거리를 이용한 파라미터가 일반적인 음성

과 음악의 구별에서 높은 성능을 나타내는 반면, 빠른 템포의 음악인 경우 스펙트럼의 변화가 심하게 되고 이로 인해 캡스트럼 거리가 커지게 되어 인식 성능이 저하되는 문제점이 있다. 이를 해결하기 위해, 현재 프레임에서 일정한 간격만큼 떨어진 프레임과 비교하여 캡스트럼 거리를 구하는 것이 아니라 인접한 여러 프레임들과 비교하여 구한 여러 개의 캡스트럼 거리 중에서 최소값을 그 프레임에서의 캡스트럼 거리로 선택하는 방법을 사용했다[7]. 즉, 인접한 여러 프레임들 중에서 현재 프레임과 가장 스펙트럼 포락선이 닮은 프레임의 캡스트럼 거리의 평균을 파라미터로 이용하는 것이다. 이와 같은 파라미터를 mean of minimum cepstral distance(MMCD) 라고 하고 이는 다음 식과 같이 정의된다.

$$MCD(n) = \min_{d_1 \leq d \leq d_2} \left[ \sqrt{\sum_{k=1}^K (c(n+d,k) - c(n,k))^2} \right] \quad (3)$$

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

여기서  $MCD(n)$ 은  $n$ 번째 프레임과 인접한 프레임들과의 캡스트럼 거리 중에서 최소값을 나타내며,  $d_1$ 과  $d_2$ 는 비교대상 프레임의 범위를 나타낸다.

### III. 다차원 MMCD 파라미터

MMCD 파라미터를 구하는 과정에서 캡스트럼 거리의 최소값을 구하는 프레임의 범위인  $d_1$ 과  $d_2$  구간을 실험을 통하여 최적화 하였으며, 그 결과 음성/음악 판별 성능에서 기존 파라미터보다 우수한 성능을 보였다. 그러나 이 방법의 문제점은 최소값을 구하는 프레임 범위인  $d_1$ ,  $d_2$ 의 구간별로 음성/음악 판별 성능이 비교적 큰 차이를 보인다는 것이다[7].

그림 2는 기존의 결과 중에서 서로 다른  $d_1$ 과  $d_2$  구간에 의해 구해진 MMCD 파라미터를 사용하여 구한 인식률을 보인 것이다. 그림 2에서 구간A는  $d_1$ ,  $d_2$ 로 각각 10ms, 150ms를, 구간B는 각각 30ms, 250ms를 사용하였다. 구간A에서 구한 MMCD 파라미터는 음성과 음악에서 비교적 큰 성능 차이를 나타내지만 구간B의 MMCD 파라미터는 음성과 음악 모두 안정적으로 높은 성능을 나타낸다. 이와 같이 기존의 MMCD 파라미터는 캡스트럼 거리의 최소값을 구하는 프레임 범위를 실험을 통해 최적화해야 하는 단점이 있을 뿐만 아니라, 그렇게 하더라도 훈련용 데이터에서 관찰되지 않는 특성을 가지는 다양한 음성과 음악 데이터에 대해서 안정적인 성능을 보장할 수 없다는 문제가 있다.

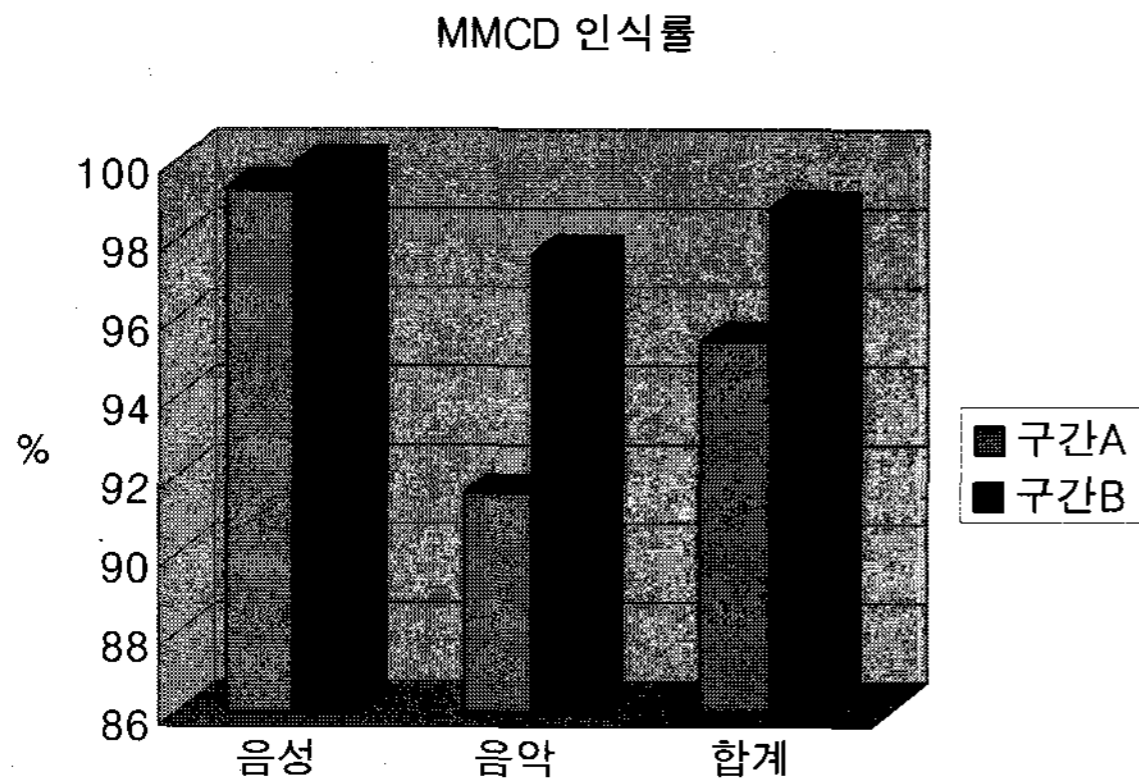


그림 2. 기존 MMCD 파라미터의 인식률 차이

본 논문에서는 기존의 MMCD를 사용함에 있어서 성능에 큰 영향을 미치는  $d_1$ 과  $d_2$  구간의 최적화 문제를 해결하기 위해 다차원 MMCD 파라미터를 제안한다.

$p$ 차원 MMCD 파라미터를  $x = [x_1 x_2 x_3 \dots x_p]$ 라고 하면, 각 차원의  $x_i$ 는 서로 다른  $d_1$ 과  $d_2$ 를 갖는 MMCD 파라미터로 구성된다. 이렇게 구성된 다차원 파라미터의 분포를 Gaussian mixture model(GMM)로 표현하면 다음과 같다.

$$p(x|\theta) = \sum_{i=1}^m \alpha_i p_i(x|\theta_i) \quad (5)$$

$$p_i(x|\theta_i) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]$$

여기서  $\mu$ 는  $p$ 차원 MMCD 파라미터의 평균벡터이고,  $\Sigma$ 는  $p \times p$ 차원의 공분산 행렬이다.

다차원 MMCD의 구성은 선행 연구[7]에서 비교적 높은 인식 성능을 보인  $d_1, d_2$  값들에 의거하여,  $d_1$ 은 10, 30, 50ms의 값을, 그리고  $d_2$ 는 150, 250ms 값을 가지도록 하여, 총 6가지 구간에 대한 MMCD 파라미터들로 구성하였다. 이를 supervector로 만들어 식 (5)와 같이 Gaussian 분포로 모델링하였다. 식 (5)의 공분산 행렬  $\Sigma$ 는 다차원 MMCD인 경우 벡터의 차원 간 연관관계가 크므로 full covariance를 사용하였다.

그림 3은 다차원 MMCD에서 특정한 두 차원 파라미터를 선택하여 음성과 음악에 대한 scatter plot을 그려본 것이다. 이 그림으로부터 차원간 연관관계가 높다는 점과 더불어, 단일 차원 MMCD에 비해 다차원 MMCD를 통해 음성과 음악의 판별을 더 효과적으로 할 수 있다는 가능성을 확인할 수 있다.

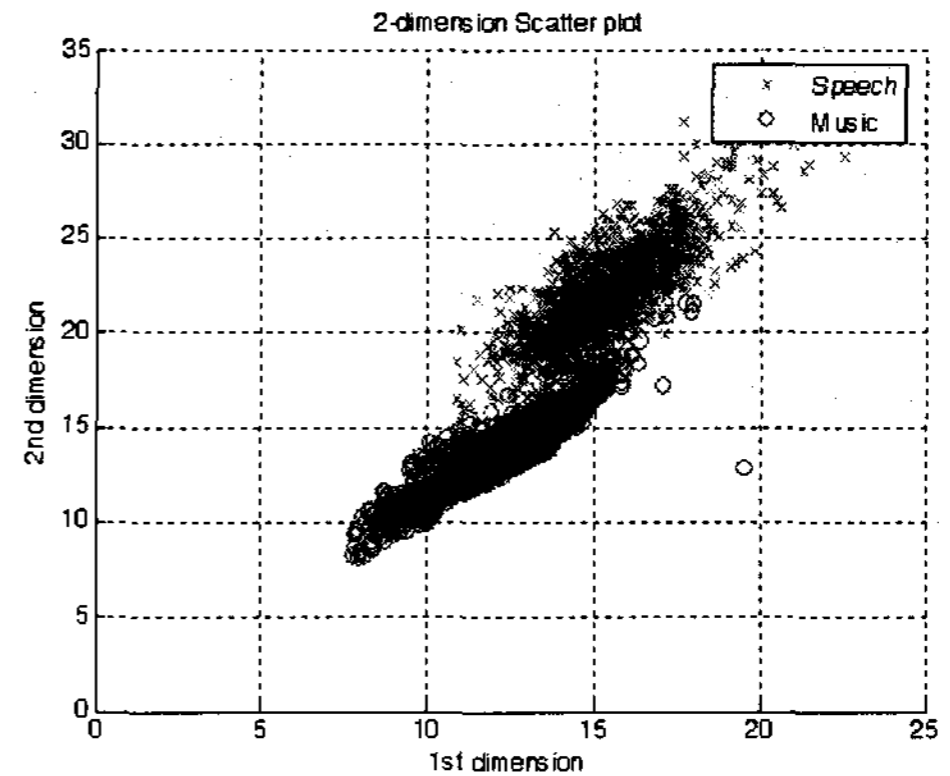


그림 3. 다차원 MMCD의 특정 두 차원 파라미터에 의한 scatter plot의 예

## IV. 인식실험 및 결과

### 4.1. 실험환경

본 논문에서 제안한 방법의 성능 평가를 위한 훈련과 테스트에 사용된 음성 데이터는 선행 연구와의 동등한 비교를 위해 국어공학센터에서 제공하는 음성 DB를 사용하였고 음악 데이터도 선행 연구와 동일한 것으로서 다양한 장르의 클래식 음악을 사용 하였다 [7].

음성/음악 판별을 위한 분류기도 선행 연구와 동일한 GMM 분류기를 사용하였다. 켄스트럼 거리 측정을 위해 MFCC를 이용하여 켄스트럼 거리를 계산하였고, 이 때 프레임의 크기는 25ms, shift size는 10ms로 하였다. 이렇게 구한 MFCC를 이용하여 1초 구간마다 MMCD를 계산하여 파라미터로 사용하였다.

### 4.2. 실험결과

기존의 MMCD 파라미터를 이용한 음성/음악 판별 성능을 표 1에 나타내었다. 켄스트럼 거리의 최소값을 구하는 범위별로 음성과 음악의 평균 인식률을 Gaussian mixture 개수에 대해 비교한 결과이다.

표 1. 단일 MMCD를 이용한 분류 결과(%)

범위 (ms) mix#	10 ~	30 ~	50 ~	10 ~	30 ~	50 ~
	150	150	150	250	250	250
1	95.44	99.11	99.03	95.53	98.81	99.08
2	95.08	99.11	99.03	95.00	98.83	99.08

본 논문의 표 1의 결과와 동일한 파라미터를 이용한 선행 연구 결과[7]와는 약간의 성능 차이가 있는데, 이는 MMCD 파라미터 추출과정에서 다른 계산 방법을 적용하였기 때문이며, 기존의 결과 특성과 큰 변화가 없음을 확인 하였다.

표 1에서 보는 바와 같이 단일 MMCD를 사용하는 경우  $d_1$ 이 30ms이고  $d_2$ 가 150ms인 구간에서 최고 성능을 보였다. 반면에  $d_2$ 에 관계없이  $d_1$ 이 10ms인 두 구간은 모두 상대적으로 저조한 성능을 보였는데 이는 음성의 경우에도 매우 가까운 프레임 사이에서는 스펙트럼 포락선이 비슷하여 캡스트럼 거리의 최소값이 작게 되므로, 일반적으로 작은 MMCD값을 갖는 음악과의 변별력이 떨어지기 때문이다. 또한  $d_2$ 가 150ms인 경우  $d_1$ 구간의 변화에 의해 최대성과 최저성이 나타났는데 이는 MMCD를 구하는 범위 선정이 성능에 상당한 영향을 미치는 것을 보여주는 예이다.

본 논문에서 제안한 다차원 MMCD의 음성/음악 판별 성능을 표 1에서 보인 최고 성능의 결과와 비교하여 그림 5에 나타냈다.

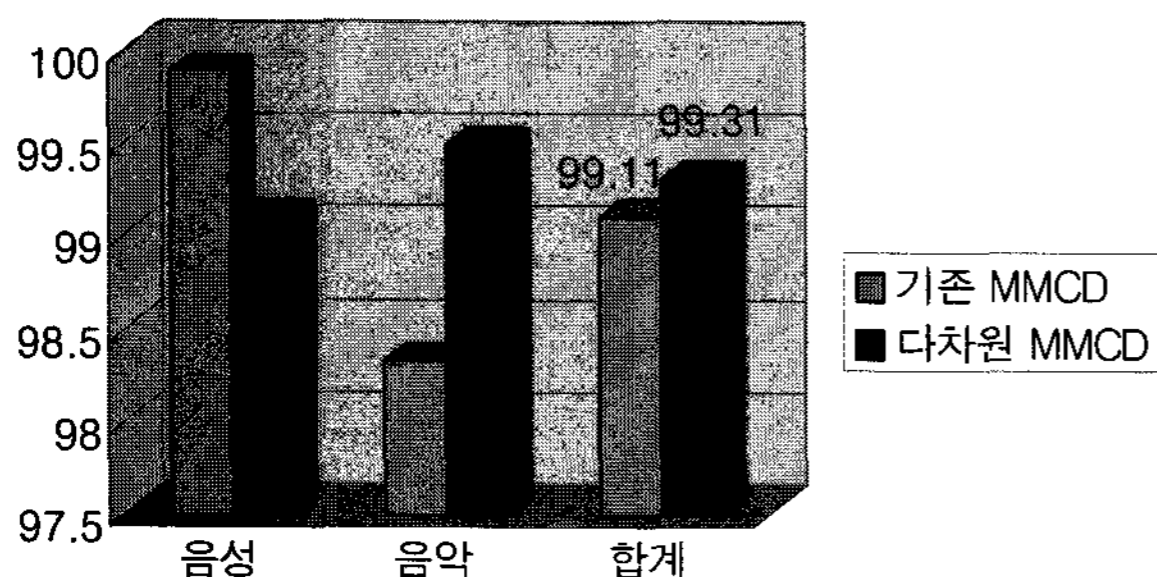


그림 5. 기존 파라미터와 제안한 파라미터와의 성능 비교

그림 5에서 보는 바와 같이 기존 MMCD는 음성에서 특히 높은 성능을 보이거나 음악에서는 비교적 낮은 성능을 보인다. 이는 최소값을 구하는  $d_1$ 간격의 변별력이 다양하게 변하는 음악의 리듬을 반영하지 못한다는 결과이다. 그에 비해 다차원 MMCD의 경우 음성에서는 기존 MMCD 보다 낮은 성능을 보이거나 평균적인 성능에서 22.5%의 인식 오류를 감소를 나타냈다. 이는 결과적인 성능 향상에 대한 의의도 있겠으나 MMCD 파라미터를 실험적으로 최적화해야 하는 문제를 해결한 점에서 의미가 더 크다고 할 수 있다. 이로 보건데 다차원 MMCD는 다양한 발성과 음악의 많은 변화의 차이에서도 변별력이 높은 파라미터라고 할 수 있겠다.

## V. 결론

본 논문에서는 음성과 음악의 판별을 위한 효과적인 특징 파라미터인 MMCD의 단점, 즉, 캡스트럼 거리의 최소값을 구하는 범위를 실험적으로 최적화해야 하는 문제를 극복하기 위하여, 복수의 프레임 범위에 대해 각각 구한 MMCD들을 함께 사용하는 다차원 MMCD 파라미터를 제안하였다. 실험 결과, 제안된 MMCD 파라미터가 프레임 범위 최적화를 통한 기존의 MMCD 파라미터의 최고 성능에 비해서도 22.5%의 오류감소를 얻을 수 있었다. 이러한 성능향상도 의미가 있지만, 최소 캡스트럼 거리를 구하는 범위의 최적화 문제로부터 자유롭기 때문에, 훈련용 데이터에서 관찰되지 않는 특성을 가지는 다양한 음성과 음악 데이터에 대해 안정적으로 우수한 성능을 나타낼 수 있다는 점이 더 큰 의미를 갖는다고 판단된다.

## 참고문헌

- [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] 박슬한, 김형순, "캡스트럼 거리를 이용한 음성/음악 판별 성능 향상", 제 18회 신호처리합동학술대회 논문집, vol. 18, no. 1, pp. 1, 2005년 9월.
- [7] 박슬한, 최무열, 김형순, "변형된 캡스트럼 거리를 이용한 음성/음악 판별", *말소리*, vol. 56, pp. 195-206, 2005년 12월.