멜 켑스트럼 모듈레이션 에너지를 이용한 음성/음악 판별*

김봉완(SiTEC), 최대림(SiTEC), 이용주(원광대)

<차 례>

<Abstract>

Speech/Music Discrimination Using Mel-Cepstrum Modulation Energy

Bong-Wan Kim, Dea-Lim Choi, Yong-Ju Lee

In this paper, we introduce mel-cepstrum modulation energy (MCME) for a feature to discriminate speech and music data. MCME is a mel-cepstrum domain extension of modulation energy (ME). MCME is extracted on the time trajectory of Mel-frequency cepstral coefficients, while ME is based on the spectrum. As cepstral coefficients are mutually uncorrelated, we expect the MCME to perform better than the ME.

To find out the best modulation frequency for MCME, we perform experiments with 4 Hz to 20 Hz modulation frequency. To show effectiveness of the proposed feature, MCME, we compare the discrimination accuracy with the results obtained from the ME and the cepstral flux.

* Keywords: Speech/Music discrimination, Mel-cepstrum modulation energy, Modulation energy.

* 이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (지방연구중심대학육성사업/헬스케어기술개발사업단).

1. 서 론

오디오 데이터에서 음성과 음악을 자동으로 판별하기 위한 시스템은 여러 방면으로 사용될 수 있다. 음성 인식 시스템의 전처리기로서 입력 데이터의 음악 부분을 제거하고 음성 부분만을 입력할 수 있도록 사용될 수 있으며, 오디오 데이터의 자동 인덱싱 및 검색 분야에도 사용될 수 있다.

음성 및 음악 판별을 위한 시스템의 성능을 향상시키기 위해서는 적합한 특징 파라미터의 선택이 매우 중요하며, 많은 특징 파라미터들이 제안되어 왔다. 이들 파라미터들은 시간 영역, 스펙트럼 영역 및 켑스트럼 영역으로 나누어 볼 수 있으며, 주요 파라미터들의 특성과 판별 성능은 [1][2]에서 잘 비교되어 있다. 이를 요약하면 다음과 같다.

시간 영역 파라미터들을 이용한 방법으로는 영교차율을 이용한 방법[3][4] 및 신호의 단구간 에너지를 이용한 방법[5] 등이 있다. 스펙트럼 영역의 파라미터들을 이용한 방법으로는 스펙트럼 파워분포의 균형점(spectral centroid)를 이용한 방법과 프레임간 스펙트럼의 차이(spectral flux, SF)를 이용한 방법이 있다[3]. 또한 음성의 경우 음절 발음의 영향으로 인해 약 4 Hz에서 모듈레이션 에너지(modulation energy, ME)의 피크가 발생한다는 연구 결과[6]에 따라 필터뱅크 출력에서 4 Hz의 모듈레이션 에너지를 구하고 이를 특징으로 사용하는 방법[3]도 제안되어 사용되어 왔다. 켑스트럼 영역의 파라미터들을 이용한 방법으로는 프레임간 켑스트럼의 차이(cepstral flux, CF)를 이용한 방법[7]이 있으며, 이는 SF의 켑스트럼 버전이라고 할 수 있다. 최근 인접한 여러 프레임들 간의 켑스트럼 거리의 최소값의 평균을 이용하는 방법[2][9]으로 좋은 판별 성능을 보임이 보고된 바 있다. 그 외에 음악의 리듬을 이용한 방법[3], 음소 인식 결과를 기반의 엔트로피(entropy)와 다이나미즘(dynamism)을 이용하는 방법[8] 등이 있다.

일반적으로 CF가 SF보다 좋은 성능을 내는 것으로 알려져 있는데, 그 이유는 SF의 경우 프레임간 변화를 측정할 때 모든 DFT 포인트의 값들을 비교하므로 음정 변화에 따른 세밀한 스펙트럼 변화에 민감하기 때문이다. 이에 반해 CF의 경우 스펙트럼 포락선(spectral envelope)를 표현하는 켑스트럼의 저차 성분들만을 이용하기 때문에 SF에 비해 좋은 성능을 보이는 것으로 알려져 있다.

ME의 경우 SF와 마찬가지로 필터 뱅크의 모든 출력을 이용하여 모듈레이션 에너지를 계산한다. 이 경우 각 채널의 출력은 서로 강한 상관을 갖고 있으므로 성능 저하가 발생하리라는 것을 예측할 수 있다. 따라서 본 논문에서는 CF가 SF의 확장 버전인 것처럼, 멜 켑스트럼 영역에서 구한 모듈레이션 에너지(mel-cepstrum modulation energy, MCME)를 음성/음악 판별을 위한 특징으로 사용할 것을 제안한다. MFCC의 계수들이 필터 뱅크의 계수들보다 서로 상관이 적다는 것을 감안할 때, MCME가 ME보다 음성 및 음악 판별에 좋은 성능을 보이리라

고 예상할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 필터 뱅크를 이용한 모듈레이션 에너지 분석 방법을 기술하고, 3장에서는 제안된 멜 켑스트럼 기반의 모듈레이션 에너지 분석 방법 및 그 특성에 대하여 기술한다. 4장에서는 실험 및 검토에 대하여 기술하고 5장에서는 결론을 맺는다.

2. 기존의 모듈레이션 에너지 분석 방법

오디오 신호의 n번째 프레임으로부터 구한 1차 DFT를 $X[n,k]$라고 하면, 다음 식 (1)과 같이 k번째 계수의 프레임별 결과에 대하여 2차 DFT를 취함으로써 크기 모듈레이션 스펙트럼(magnitude modulation spectrum, MMS)을 얻을 수 있다.

$$MMS[n,k,q] = \sum_{p=0}^{P-1} |X[n+p,k]| e^{-j2\pi qp/P} \tag{1}$$

여기에서 n은 프레임 인덱스, k는 1차 DFT결과의 주파수측 인덱스, q는 2차 DFT의 주파수측 인덱스, P는 2차 DFT의 사이즈이다.

모듈레이션 주파수가 낮다는 것은 시간에 따른 스펙트럼의 변화가 느림을 의미하며, 높다는 것은 스펙트럼의 변화가 빠르게 나타난다는 것을 의미한다. 따라서 음성의 경우 유성음과 무성음의 연이은 발음으로 인해 스펙트럼의 변화가 자주 발생하는데 비해 음악의 경우 스펙트럼의 변화가 크지 않은 특징이 있으므로 모듈레이션 스펙트럼은 음성 및 음악의 판별을 위한 특징으로 사용된다.

음성 및 음악 판별을 위해 모듈레이션 주파수 분석을 수행할 때, 대부분 1차 DFT의 결과를 그대로 이용하지 않고, 청각 특성을 반영한 멜(Mel) 주파수 대역으로 분할하는 필터 뱅크 분석 결과를 이용하여 2차 DFT 분석을 수행한다.

또한 음성의 경우 음절 발음의 영향으로 인해 약 4 Hz에서 모듈레이션 에너지의 피크가 발생한다는 연구 결과[4]에 따라, 모듈레이션 에너지 분석에서 모든 주파수를 사용하지 않고 4 Hz (또는 4~8 Hz)의 모듈레이션 에너지만을 계산하여 이를 음성 및 음악 판별을 위한 특징으로 사용하여 왔다[1][10][11][12]. 이 처럼 특정 모듈레이션 주파수 정보만을 사용할 때에는 2차 분석에서 계산량 절감을 위해 모든 주파수에 대하여 DFT 분석을 수행하지 않고 중심 주파수가 4 Hz인 밴드패스 필터(bandpass filter)를 이용하여 에너지를 계산한다. 특징의 차수를 줄이기 위해 필터 뱅크의 각 채널의 4 Hz 모듈레이션 에너지를 합산하여 1차의 특징을 추출하고, 정규화(normalization)를 수행한 후 이를 음성 및 음악 판별을 위한 특징으로 사용한다.

본 논문에서 사용된 모듈레이션 에너지의 정의는 식 (2)와 같다. 여기에서 $FMS[n,m,q]$는 필터뱅크의 출력의 n번째 프레임으로부터 구한 모듈레이션 스펙트럼으로 1차 DFT의 결과열로부터 구한 크기 모듈레이션 스펙트럼 MMS와 구분하기 위하여 FMS로 표기하였다. m은 필터 뱅크 계수의 인덱스, M은 필터뱅크의 차수 그리고 $E[n]$은 소스 시그날의 n번째 프레임의 단구간 에너지를 의미한다.

$$ME[n,\ q] = \frac{\frac{1}{M}\sum_{m=0}^{M-1}|FMS[n,m,\ q]|^2}{\frac{1}{P}\sum_{p=0}^{P-1}\log(E[n+p])} \tag{2}$$

제안된 방법은 정규화를 위한 분모가 기존의 문헌[3][4]와 다소 다르다. 기존 방법의 경우 4 Hz 모듈레이션 에너지와 전체 모듈레이션 에너지의 합과의 비율을 계산하고, 이들을 전체 필터뱅크 밴드에 대해 더하는 방법을 사용한다. 제안된 방법은 모듈레이션 에너지의 다이나믹 레인지(dynamic range)를 줄이기 위해 신호의 단구간 로그 에너지의 평균을 사용한다. 사전 실험을 통하여 제안된 방법이 기존 방법에 비해 다소 좋은 성능을 보이는 것을 확인하였다. 이는 기존의 방법이 모듈레이션 스펙트럼의 변화에 민감하게 반응하기 때문으로 해석된다. 또한 기존의 방법은 비율을 계산하기 위해 매 프레임마다 모든 모듈레이션 에너지를 계산하여야 하므로 계산량이 많지만 제안된 방법은 신호의 단구간 에너지를 사용하므로 필터뱅크 계산시 한번만 계산하면 되므로 계산량이 감소하는 잇점이 있다.

3. 멜 켑스트럼 모듈레이션 에너지

$C[n,l]$을 1차 DFT $X[n,k]$의 실수 켑스트럼이라고 하면, 이는 다음과 같이 구할 수 있다.

$$C[n,l] = \frac{1}{K}\sum_{k=0}^{K-1}\log(|X[n,k]|)e^{+j2\pi kl/K} \tag{3}$$

$C[n,l]$이 대칭을 이루는 실수열(real symmetric sequence)이므로, 저차의 $C[n,l]$로부터 쿼프런시 리프터(quefrency lifter)를 통하여 평탄화된 스펙트럼에 대한 추정(cepstrally smoothed estimate of spectrum) $S[n,k]$를 다음과 같이 얻을 수 있다.

$$\log S[n,k] = C[n,0] + \sum_{l=1}^{L-1}2C[n,l]\cos(2\pi lk/K) \tag{4}$$

위의 식 (3)과 식 (4)을 이용하면 다음과 같이 MMS에 대한 추정을 얻을 수 있다.

$$\begin{aligned} MMS'[n,k,q] &= \sum_{p=0}^{P-1} \log(S[n+p,k]) e^{-j2\pi pq/P} \\ &= \sum_{p=0}^{P-1} C[n+p,0] e^{-j2\pi pq/P} \\ &\quad + \sum_{l=1}^{L-1} \cos(2\pi kl/K) \sum_{p=0}^{P-1} 2C[n+p,l] e^{-j2\pi pq/P} \end{aligned} \tag{5}$$

식 (5)의 마지막 부분을 이용하여 멜 켑스트럼 모듈레이션 스펙트럼 (mel-cepstrum modulation spectrum, MCMS)을 다음과 같이 정의한다.

$$MCMS[n,l,q] = \sum_{p=0}^{P-1} C[n+p,l] e^{-j2\pi pq/P} \tag{6}$$

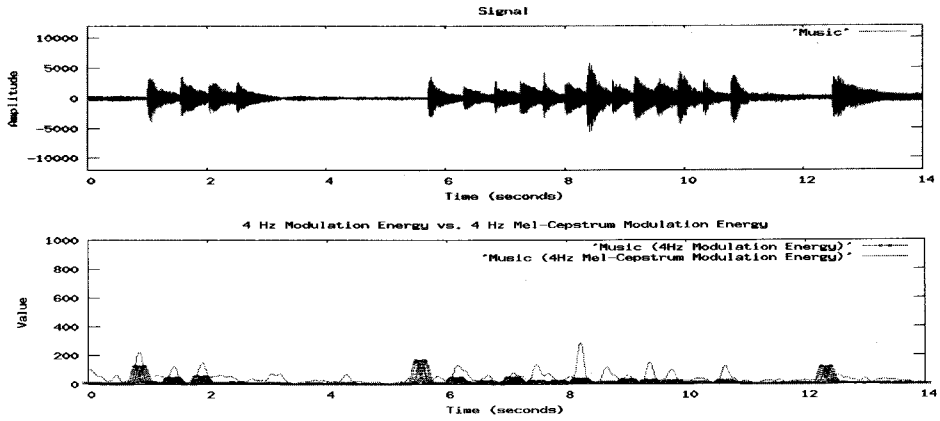
Tyagi 등은 모듈레이션 스펙트럼은 잡음이 부가된다 하여도 변화가 크지 않다는 사실[13]에 착안하여 MCMS를 MFCC의 다이나믹 특징(dynamic feature)으로 사용할 경우, 차분 파라미터 및 RASTA PLP와 비교하여 잡음 환경에서 음성 인식 시스템의 성능을 향상 시킬 수 있음을 보였다[14][15].

식 (5)를 통하여, $MMS'[n,k,q]$는 $MCMS[n,l,q]$의 선형 변환이라는 것을 알 수 있다. MFCC의 계수들이 필터 뱅크의 계수들보다 서로 상관이 적다는 것을 감안할 때, MCMS가 MMS보다 음성 인식에 있어 더 좋은 성능을 보이리라고 예상할 수 있다. 즉 MMS가 전체 스펙트럼의 시간에 따른 변화를 모델링하는 데 비하여, MCMS의 경우 스펙트럼 포락선에 영향을 미치는 요소들의 변화만을 모델링한다는 차이가 있다. 그러나 MCMS의 경우에는 MFCC의 각 계수별 스펙트럼을 모두 별도로 취급함으로써 특징 차수가 커지는 단점 (즉, 오디오 신호에서 추출된 일련의 MFCC 열에서 MFCC의 각 계수별로 주파수 분석을 수행하고 그 결과를 모두 특징으로 사용한다. 따라서 B개의 대역 필터를 통하여 MCMS를 계산할 경우 추출되는 특징 벡터의 차수는 $B \times$ MFCC 차수가 된다.)을 갖고 있다.

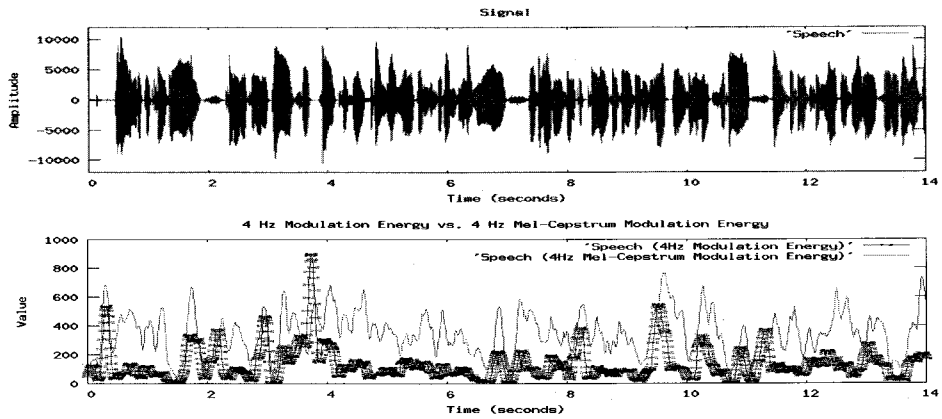
따라서, 본 논문에서는 MCMS의 장점을 취하면서 특징의 차수를 줄이기 위해 다음과 같이 멜-켑스트럼 모듈레이션 에너지(mel-cepstrum modulation energy, MCME)를 정의하고 이를 음성/음악 판별을 위한 특징으로 사용하고자 한다.

$$MCME[n,q] = \frac{\frac{1}{L}\sum_{l=0}^{L-1} |MCMS[n,l,q]|^2}{\frac{1}{P}\sum_{p=0}^{P-1} \log(E[n+p])} \tag{7}$$

다음 <그림 1>은 음악 및 음성 신호에서 추출한 4 Hz ME와 4 Hz MCME를 비교한 그림이다. <그림 1>(a)와 <그림 1>(b)에서 위의 창은 오디오 신호의 웨이브



(a) 음악 신호의 경우

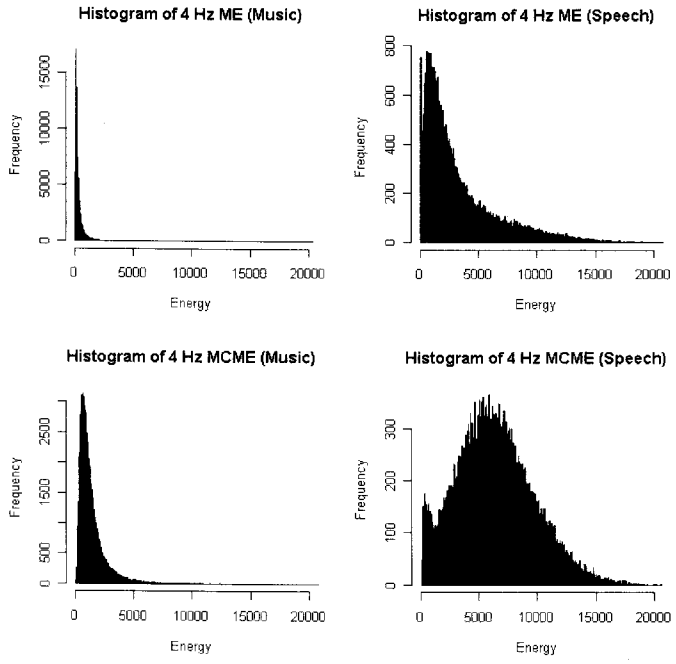


(b) 음성 신호의 경우

<그림 1> 4 Hz ME(—✕—)와 4 Hz MCME(———)의 비교

폼을, 그리고 아래의 창은 오디오신호로부터 추출한 특징들이며 굵은 선이 4 Hz ME, 가는 선이 4 Hz MCME이다. <그림 1>(a)와 <그림 1>(b)를 비교해 보면 ME와 MCME 모두 음악보다 음성에서 큰 값을 갖는 것을 볼 수 있다. 또한 MCME의 경우 ME보다 대체로 큰 값을 갖는데 음성데이터에서 보다 큰 값을 갖게 되며, 음악과의 격차가 더욱 두드러짐을 볼 수 있다.

<그림 2>는 학습데이터의 음성 및 음악 파일에서 구한 4 Hz ME와 4 Hz MCME의 히스토그램이다. ME의 경우 음성의 에너지가 음악에 비해 크긴 하지만 그 중심이 음악에 가깝게 위치함을 볼 수 있다. 그러나 MCME의 경우 음악의 에너지는 큰 변화가 없는데 반해 음성의 경우 중심이 에너지가 높은 쪽으로 이동하여 음악과 음성의 분리가 좀 더 분명하게 이루어지고 있음을 볼 수 있다.



<그림 2> 4 Hz ME와 4 Hz MCME의 히스토그램 비교

4. 실험 및 검토

4.1. 데이터베이스

본 논문에서 제안한 MCME의 성능을 평가하기 위해 다음과 같은 데이터베이스를 사용하였다.

4.1.1. 낭독 문장 및 CD 음악 데이터 (RDCD) 세트

RDCD세트는 자유발화가 아닌 낭독체 음성 데이터와 CD로부터 추출된 음악 데이터로 구성되어 있다. 음성 데이터를 한국어/영어/중국어 문장으로 구성한 것은 MCME 및 ME가 음절 발음율과 관련이 있는 특징이므로 각 언어별 발성 특징이 음성/음악 판별 성능에 영향을 미칠 수 있기 때문이다. 따라서 학습 데이터에 포함되지 않은 다른 언어의 음성을 평가 데이터에 포함함으로써, 언어가 미치는 영향을 살펴보기 위해 영어 및 중국어 문장을 추가하였다.

▪ 한국어 문장

- CleanSent01 DB[16] 중 남,녀 각 10명분, 총 2,120문장 (약 189분 분량).

▪ 영어 문장
- TIMIT The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[17] 중 Test set에 포함된 문장만을 선정, 총 1,680문장 (약 86분 분량).

▪ 중국어 문장
- Chinese03 DB[16] 중 남,녀 각 5명분, 총 1,000문장 (약 62분 분량).

▪ RWC - Music Database : Music Genre[18]
- 10개의 장르에 총 33개의 서브 카테고리로 구성.
- 추가로 아카펠라 1장르를 포함하여 총 100곡 (약 420분 분량)
- 약 50%의 음악이 가수의 노래가 포함되어 있음.

▪ RWC - Music Database : Popular Music[18]
- 다수의 일본 유명곡과 일부 서양의 유명곡, 총 100곡 (약 406분).
- 대부분 가수의 노래가 포함되어 있음.

4.1.2. 방송 데이터 (BRDC) 세트

RDCD세트로 학습 데이터를 구성했을 경우, 동일한 환경의 음성 및 음악에 대해서는 좋은 판별 성능을 나타낸다고 하여도, 자유 발화와 같은 실제의 음성과 채널 노이즈가 가미된 음악데이터처럼 학습 환경과 다른 오디오 데이터에서는 성능 저하가 발생할 수 밖에 없다. 따라서 이러한 환경에서의 성능을 검증하기 위하여 인터넷 방송에서 추출된 오디오 데이터를 테스트 데이터로 포함하였다. KBS에서 제공되는 '콩3'[19] 프로그램을 이용하여 실시간 방송되는 FM 라디오 내용을 PC에서 녹음하고 이를 테스트 데이터로 사용하였다. 데이터에 포함된 프로그램은 '강수정의 뮤직쇼' 2006년 12월 21일 방송분 일부, '임백천의 골든 팝스' 2006년 12월 26일 방송분 일부, '김구라의 가요 광장' 2006년 12월 26일 방송분 일부이다.

프로그램의 중간에 방송되는 광고 방송은 데이터에서 제거하였으며, 진행자의 발성이 이루어지는 사이에 배경음악이 포함되는 것은 그대로 음성 데이터로 포함시켰다. 프로그램의 특성상 프로그램 진행자와 손님과의 자유로운 담화, 웃음 소리, 리듬이 있는 발성들도 포함되어 있다. 방송 프로그램으로부터 얻어진 음성 데이터는 약 70분 분량, 음악 데이터는 약 87분 분량이다.

4.1.3. 학습 및 평가 세트의 구성

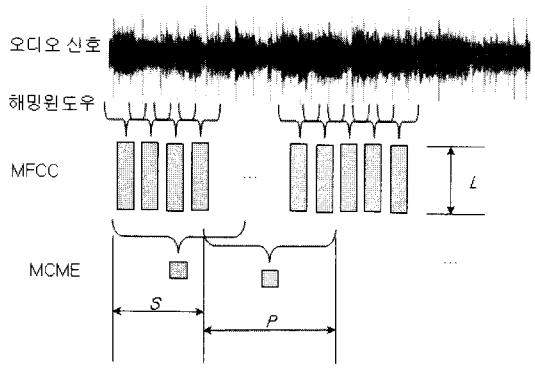
학습을 위한 데이터는 RDCD 세트에서 선정하였다. 음성 모델을 학습하기 위한 음성 데이터로는 한국어 문장 데이터베이스 중 남, 녀 5명분 총 1,060 문장을 사용하였으며, 음악 모델을 학습하기 위한 데이터로는 RWC - Music Database :

Music Genre 데이터베이스의 33개의 서브카테고리에서 무작위로 1곡씩을 선정하여 음악 모델의 학습을 위해 사용하였다. 선정된 음악 데이터 33곡 중 가수의 노래가 포함된 곡은 16곡이다. 학습 데이터를 제외한 전량이 테스트 데이터로 사용되었으며, 그 크기 비율은 236분(학습 데이터) 대 1,084분(평가 데이터)로 약 1:4.6이다.

음악 CD 데이터의 경우 16 kHz로 다운샘플링하고 16 비트로 양자화하여 음성 데이터의 포맷과 맞추었다. 방송 데이터의 경우 '콩3'을 통하여 방송을 수신하면서 실시간으로 16 kHz, 16 비트로 양자화 하여 저장하였다. 테스트를 위한 음악 데이터 및 방송 데이터의 경우 그 길이가 판별 성능에 영향을 미칠 수 있으므로 각 파일을 평균 15초 길이의 세그먼트들로 분할하여 저장하고 이를 테스트용으로 사용하였다.

4.2. 실험 환경

판별 실험을 위한 특징을 추출하기 위하여 25 ms의 해밍 윈도우를 10 ms 단위로 프레임을 이동하면서 24차의 필터 뱅크 결과와 에너지, 12차의 MFCC 결과와 에너지를 추출하였다. 필터 뱅크의 차수는 음성 인식을 위해 사용되는 12차 MFCC를 추출하기 위하여 24차의 필터 뱅크를 주로 사용하므로 12차의 MFCC 기반 MCME와 공정한 비교를 위해 24차를 선택하였다. 위의 과정을 거쳐 추출된 MFCC로부터 MCME를 추출하기 위한 후속 과정을 <그림 3>에 나타내었다.



<그림 3> MCME 추출 과정

차수가 L인 MFCC열로부터 MCME 추출을 위하여 P개의 MFCC를 이용하여 각 계수별로 대역 필터를 통과시킨 후 그 결과의 합을 해당 구간의 평균 로그 에너지로 나누어 준 후 최종적으로 1차의 MCME를 추출한다. 이후 S만큼 이동하여 MCME 추출과정을 반복한다. ME를 추출하기 위한 과정은 <그림 3>에서 MFCC열

을 필터 뱅크 분석 결과 열로, L을 필터 뱅크 결과의 차수 M으로 대치하면 그대로 동일하다.

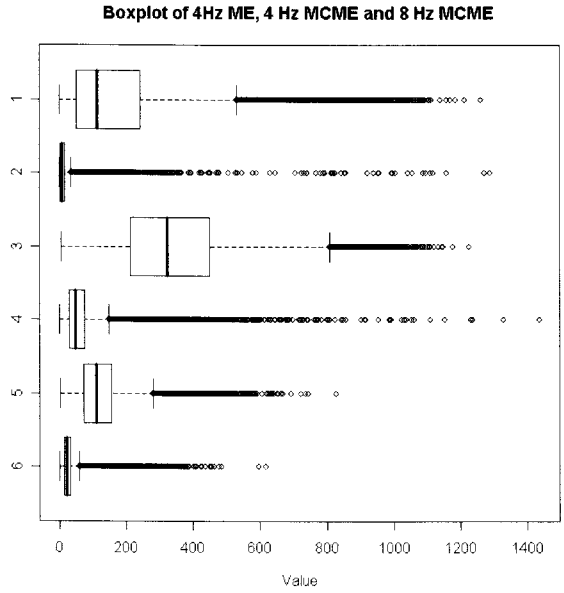
본 논문에서는 실험을 위하여 $P=25, S=10$의 값을 사용하였다. 따라서 최종적으로 1초당 10개의 차수가 1인 ME 및 MCME를 추출하였다. 판별 실험에 사용된 분류기는 단일 혼합(single mixture) GMM을 사용하였으며 HTK[20]를 이용하여 실험을 수행하였다.

4.3. 실험 결과 및 검토

음성의 경우 음절 발음의 영향으로 인해 약 4 Hz에서 ME의 피크가 발생한다는 연구 결과[4]를 기반으로 대부분의 ME 기반 음성/음악 판별 시스템은 4 Hz의 ME를 특징으로 사용한다. 그러나 3장에서 살펴본 바와 같이 MCME의 경우 음성 및 음악 판별을 위해 적합한 모듈레이션 주파수도 변화할 수 있으므로 실험을 통하여 이를 검증하기 위해 4 ~ 20 Hz의 범위에 걸쳐 실험을 진행하였다. 그 결과는 <표 1>과 같다.

<표 1> 음성/음악 분류 성능(%); ME와 MCME의 비교

Feature	Test set		Modulation Frequency (Hz)					Average
			4	8	12	16	20	
ME	RDCD	Speech	100.00	99.90	99.90	100.00	99.70	99.90
		Music	98.20	97.50	94.80	93.80	93.70	95.60
		Average	99.10	98.70	97.35	96.90	96.70	97.75
	BRDC	Speech	82.60	89.10	86.40	72.10	57.70	77.58
		Music	96.80	96.20	95.90	97.10	97.30	96.66
		Average	89.70	92.65	91.15	84.60	77.50	87.12
	Total	Speech	98.90	99.20	99.10	98.10	96.90	98.44
		Music	98.00	97.30	94.90	94.20	94.10	95.70
		Average	98.45	98.25	97.00	96.15	95.50	97.07
MCME	RDCD	Speech	100.00	100.00	100.00	100.00	99.90	99.98
		Music	98.70	99.50	98.80	98.60	98.60	98.84
		Average	99.35	99.75	99.40	99.30	99.25	99.41
	BRDC	Speech	99.60	99.20	99.60	98.10	94.30	98.16
		Music	95.30	96.20	95.90	95.90	95.00	95.66
		Average	97.45	97.70	97.75	97.00	94.65	96.91
	Total	Speech	100.00	100.00	100.00	99.90	99.50	99.88
		Music	98.30	99.10	98.40	98.30	98.20	98.46
		Average	99.15	99.55	99.20	99.10	98.85	99.17



<그림 4> 4 Hz ME, 4 Hz MCME 및 8 Hz MCME의 상자 그림 (1: 4 Hz ME (음성), 2: 4 Hz ME (음악), 3: 4 Hz MCME (음성), 4: 4 Hz MCME (음악), 5: 8 Hz MCME (음성), 6: 8 Hz MCME (음악))

RDCD 세트 음성의 경우, 학습 데이터로는 한국어 문장 음성을 사용하고 평가 데이터로는 한국어/영어/중국어의 문장 음성을 사용했음에도 불구하고 골고루 좋은 판별 성능을 보이는 것으로 보아 모듈레이션 에너지 영역에서 언어간 차이가 미치는 영향은 별로 크지 않음을 알 수 있다.

ME의 경우 대체로 4 Hz에서 좋은 성능을 보임을 알 수 있다. 또한 모듈레이션 주파수의 변동에 따라 8 Hz를 넘어가면서 급격하게 성능이 저하되는 것을 볼 수 있다. MCME의 경우 전반적으로 ME에 비해 좋은 성능을 보이고 있고 모듈레이션 주파수의 변동에도 성능 저하의 폭이 크지 않음을 볼 수 있으며, 8 Hz에서 가장 좋은 성능을 보임을 알 수 있다.

MCME가 ME에 비해 전체적으로 평균 약 72%의 오류 감소율을 볼 수 있으며, 특히 8 Hz의 경우 4 Hz ME에 비해 71%, 8 Hz HE에 비해 74%의 오류 감소율을 보이고 있다. RDCD 세트 뿐만 아니라 BRDC 세트에서도 역시 비슷하게 성능이 향상되고 있음을 볼 수 있어 제안된 MCME가 유효함을 알 수 있다.

MCME의 경우 4 Hz 보다 8 Hz 모듈레이션 에너지에서 좋은 성능을 내고 있는 결과를 검토하기 위해, 학습 데이터에서 추출한 특징 값들의 산포의 정도를 상자 그림(boxplot)을 <그림 4>에 나타내었다. 상자 그림에서 상자와 점선으로 연결되어 있는 왼쪽 직선은 최소값(min)을 나타내며 상자는 제1사분위수(Q1)와 제3사분위수

(Q3)를 나타낸다. 그리고 상자 안의 직선은 중앙값(median)을, 상자와 점선으로 연결되어 있는 오른쪽 직선은 Q3에 1.5배의 사분위수 범위(IQR)를 곱한 값이다. 이 선을 넘어서 원으로 표시된 데이터 값들은 이상치(outlier)들을 나타낸다.

<그림 4>에서 볼 수 있는 것과 같이, 4 Hz의 경우 MCME가 ME에 비해 음성과 음악이 좀 더 잘 분리되고 있는 것을 볼 수 있으므로 ME와 MCME의 성능 차이를 설명할 수 있다. MCME 8 Hz의 경우 MCME 4 Hz와 비교하여 그 평균값은 낮지만 학습 자료의 변동폭이 좁다는 것을 알 수 있으며, 또한 이상치들의 범위가 현저히 줄어 있다는 것을 볼 수 있다. 특히 음악의 경우 이상치의 범위가 음성에 비해 현저히 줄어 있는 것은 특기할 만하다. 이러한 현상을 감안할 때, MCME의 경우 8 Hz역시 음성 및 음악 판별을 위한 좋은 주파수로 판단된다.

4 Hz 모듈레이션 에너지 기반의 특징이외의 다른 특징을 사용하는 시스템과의 성능 비교를 위해 음성/음악 판별을 위한 특징 파라미터로 빈번하게 사용되는 CF와 성능 비교실험을 수행하였다. 비교를 위해 사용된 CF의 정의는 다음과 같다.

$$CF[n] = \frac{1}{P-1}\sum_{p=1}^{P-1}\sqrt{\sum_{l=0}^{L-1}(C[n+p,l]-C[n,l])^2} \tag{8}$$

여기에서 P는 켑스트럼을 비교할 프레임의 크기, L은 켑스트럼의 차수이다. MCME와 공정한 비교를 위하여 P는 이전의 실험과 동일하게 25 (MCME를 추출하기 위해 사용한 포인트 수), L은 12 (MFCC의 차수)를 사용하였다. [7]에서는 LPC 켑스트럼을 이용하였으나, 본 논문에서는 MCME와의 비교를 위해 멜 켑스트럼을 이용하여 실험하였다. 실험 결과는 <표 2>와 같다.

<표 2> 음성/음악 분류 성능(%); cepstral flux(CF)와 MCME의 비교

Test set		CF	8 Hz MCME
RDCD	Speech	99.90	100.00
	Music	98.50	99.50
	Average	99.20	99.75
BRDC	Speech	99.60	99.20
	Music	95.90	96.20
	Average	97.75	97.70
Total	Speech	99.90	100.00
	Music	98.20	99.10
	Average	99.05	99.55

CF의 경우 <표 1>의 ME에 비해 전반적으로 우수한 성능을 보이고 있으나, 본 논문에서 제안된 MCME에 비해서는 낮은 성능을 보임을 알 수 있다. 8 Hz MCME의 경우 CF에 비해 53%의 오류 감소율을 보이고 있어 제안된 MCME가 유효함을 알 수 있다. 또한 제안된 MCME는 음성 인식기에서 주로 사용되는 MFCC를 이용하므로, 동일한 특징 영역에서 음성/음악 판별 및 음성 인식을 수행할 수 있다는 장점이 있다. 8 Hz MCME가 판별 오류를 보인 파일들에 대하여 다음 <표 3>에 정리하였다.

<표 3> 8 Hz MCME의 판별 오류 유형

구분	오류 파일 수 (전체 파일 수)	오류 유형별 파일 수
음성	2 (4,005)	▪ 배경 음악이 우세한 방송 음성, 시간상으로도 음악 부분이 우세한 음성 데이터 : 1 ▪ 코믹하고 리듬이 있는 발성, 박수 소리 등 청중 소음이 포함된 방송 음성 데이터 : 1
음악	26 (2,988)	▪ 일반적인 음악 데이터 : 3 ▪ 랩 또는 힙합처럼 가수의 노래가 주도하는 음악 데이터 : 8 ▪ 음악 방송 프로그램의 인트로 음악 데이터 : 4 ▪ 주기적인 강한 비트 또는 리듬이 있는 음악 데이터 : 11

5. 결 론

본 논문에서는 음성/음악 판별을 위한 특징으로서 모듈레이션 에너지(ME)의 성능을 개선하기 위한 멜 켑스트럼 모듈레이션 에너지(MCME)를 제안하였다. MCME은 멜 켑스트럼을 이용하여 모듈레이션 에너지를 측정하는 방법이다. ME가 전체 스펙트럼의 시간에 따른 변화를 모델링하는데 비하여 스펙트럼 포락선에 영향을 주는 켑스트럼 저차 성분의 시간에 따른 변화를 모델링함으로써 ME에 비해 좋은 성능을 보여 주었다. 8 Hz MCME의 경우 4 Hz의 ME와 비교하였을 때 71%의 오류 감소율을 보였으며, 캡스트럴 플럭스(CF)와 비교해서도 53%의 오류 감소율을 볼 수 있었다.

MCME가 음성의 발음 특성을 감안한 특징으로 음성의 판별 성능은 높지만, 그에 반해 음악의 판별 성능은 음성에 비해 다소 낮음을 볼 수 있었다. 향후 음악의 특성을 보다 잘 표현하는 특징을 추가하여 판별 성능을 높이기 위한 연구를 진행하고자 한다.

참 고 문 헌

[1] 김수미, 김형순, "음성/음악 판별을 위한 특징 파라미터와 분류기의 성능 비교", *말소리*, 제46호, pp. 37-50, 2003.

[2] 박슬한, 최무열, 김형순, "켑스트럼 거리 기반의 음성/음악 판별 성능 향상", *말소리*, 제56호, pp. 195-206, 2005.

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

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

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

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

[7] T. Asano, M. Sugiyama, "Segmentation and classification of auditory scenes in time domain", *Proc. IWHIT*, pp. 13-18, 1998.

[8] J. Ajmera, I. McCowan, H. Bourlard, "Speech/music discrimination using entropy and dynamism features in a HMM classification framework", *Speech Communication*, Vol. 40, No. 3, pp. 351-363, 2003.

[9] 최무열, 송화전, 박슬한, 김형순, "다차원 MMCD를 이용한 음성/음악 판별", *말소리*, 제60호, pp. 191-201, 2006.

[10] J. Pinquier, J.-L. Rouas, R. Andre-Obrecht, "A fusion study in speech/music classification", *Proc. ICME*, Vol. 1, pp. 409-412, 2003.

[11] S. Karnebäck, "Discrimination between speech and music based on a low frequency modulation feature", *Proc. Eurospeech*, pp. 1891-1894, 2001.

[12] A. Eronen, A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features", *Proc. ICASSP*, Vol. 2, pp. 753-756, 2000.

[13] B. E. D. Kingsbury, N. Morgan, S. Greenberg, "Robust speech recognition using the modulation spectrogram", *Speech Communication*, Vol. 25, Nos. 1-3, 1998.

[14] V. Tyagi, I. McCowan, H. Bourlard, H. Misra, "On factorizing spectral dynamics for robust speech recognition", *Proc. Eurospeech*, pp. 981-984, 2003.

[15] V. Tyagi, I. McCowan, H. Misra, H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR", *Proc. ASRU*, pp. 399-404, 2003.

[16] SiTEC, http://www.sitec.or.kr.

[17] LDC, http://www.ldc.upenn.edu/.

[18] M. Goto, "Development of the RWC music database", *Proc. ICA*, Vol. 1, pp. 553-556, 2004.

[19] KBS Internet Radio KONG, http://www.kbs.co.kr/radio/kong/.

[20] HTK, http://htk.eng.cam.ac.uk/.

접수일자: 2007년 11월 10일
게재결정: 2007년 12월 16일

▶ 김봉완(Bong-Wan Kim) : 교신저자
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: bwkim@sitec.or.kr

▶ 최대림(Dea-Lim Choi)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: dlchoi@sitec.or.kr

▶ 이용주(Yong-Ju Lee)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 전기 전자 및 정보공학부, 음성정보기술산업지원센터
전화: 063) 850-7451
E-mail: yjlee@wonkwang.ac.kr