2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

음성/음악 분류를 위한 특징 비교

이경록, 서봉수, 김진영
전남대 전자공학과 멀티미디어 DSP 연구실
Tel)062-530-0472, Fax)062-530-0472

The Comparison of Features for Speech/Music Discrimination

Kyong Rok Lee, Bong Su Seo, Jin Young Kim
Multimedia DSP Lab Dept. of Electronic Engineering, Chonnam National University
E-Mail : krlee@dsp.chonnam.ac.kr
{bsseo, kimjin}@dsp.chonnam.ac.kr

요 약

본 논문에서는 멀티미디어 정보에서 원하는 정보를 추출하는 멀티미디어 인덱싱 중 오디오 인덱싱의 전처리 부격인 음성/음악 분류실험을 하였다.

오디오 인덱싱에 있어서 음성/음악 분류기는 원 오디오 신호에서 정보를 가진 음성 부분을 분리하는 역할을 한다.

실험에서는 음성/음악 분류에서 널리 쓰이는 멜캡스트럼(Mel Cepstrum), 정규화 로그 에너지(normalized log energy), 영교차(Zero-Crossings)를 특징 파라미터로 사용하였다[1, 2, 3].

특징공간은 GMM(Gaussian Mixture Model)에 의해 모델링되었고, 오디오 신호의 분류는 각각 3가지 분류항목(음성, 음악, 음성+음악)과 2가지 분류항목(음성, 음악)을 적용하였다.

실험결과 3가지 분류항목 적용시와 2가지 분류항목 적용시 모두 멜캡스트럼을 사용하였을 때 가장 좋은 결과를 보였다.

I. 서 론

정보화시대의 도래에 의한 정보의 디지털화는 대용량 멀티미디어 데이터베이스를 일반적인 추세로 만들었다. 이러한 대용량 멀티미디어 데이터베이스에서 사용자가 원하는 정보로의 접근성을 양호하게 하기 위해서 제시된 방법이 멀티미디어 인덱싱이다.

멀티미디어 인덱싱은 크게 비디오 인덱싱과 오디오 인덱싱으로 나눌 수 있다. 비디오 인덱싱은 영상을 커트로 구분하고 오브젝트를 추출함으로써 컨텐트를 구분하는 방식이다. 오디오 인덱싱은 오디오에서 정보를 가진 부분(음성)을 분리하고 오브젝트를 추출하여 컨텐트를 구분하는 방식이다. 오디오 인덱싱은 비디오 인덱싱의 보조수단으로서 사용되기도 하며, 요즘은 음성인식기술을 접목한 컨텐트 기반 오디오 정보탐색 기술이 널리 연구되고 있다.

오디오 인덱싱에서는 처리의 용이와 계산시간 단축을 위해 전처리부로 원신호로부터 정보를 가진 부분인 음성을 분리해내는 음성/음악 분류를 필요로 한다.

음성/음악 분류에서 사용된 특징 파라미터들은 멜캡스트럼, 에너지, 영교차이다.

데이터베이스는 1시간씩 총 3회 분량의 공중파방송 뉴스와 대중가요, 음악자료를 사용하여 수작업으로 음성, 음악, 음성+음악으로 분류하였다.

음성/음악 분류기에서 사용되는 분류 알고리즘으로는 GMM을 사용하였다[1].

II. 특징 파라미터

2.1 멜캡스트럼

오디오 신호는 8kHz로 샘플링 되었고, 26개의 필터로 구성된 필터뱅크를 이용하여 필터링을 실시하였다. 이때 필터뱅크는 멜 특성을 가진다. 한 프레임의 크기는

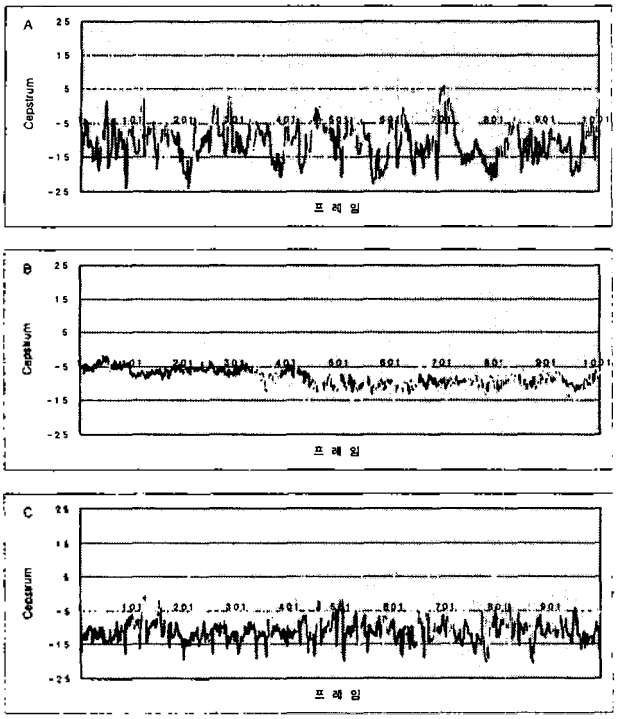

그림 1. 각 분류항목별 멜캡스트럼 변화 (1,2번째 계수 값). A : 음성, B : 음악, C : 음성+음악

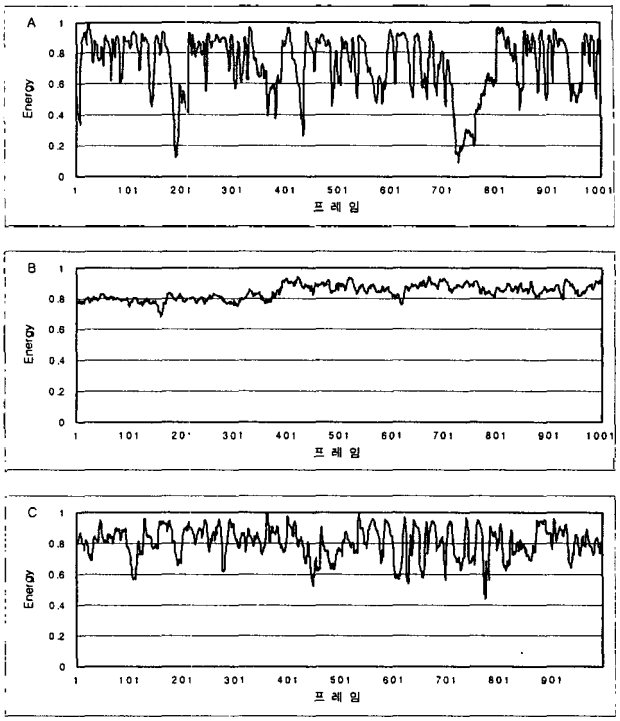

그림 2. 각 분류항목별 정규화 로그 에너지 변화. A : 음성, B : 음악, C : 음성+음악

25ms이며, 25ms의 윈도우를 10ms씩 이동시켜가면서 파라미터를 추출하였다. 연속적인 5개 프레임 구간에 대해서 델타 함수를 적용하였다. 특징 파라미터로는 12개의 멜캡스트럼 계수들과 12개의 델타 멜캡스트럼 계수를 사용하였다.

그림 1은 각 분류항목별 멜캡스트럼 변화를 나타낸 것으로 12개의 멜캡스트럼 계수값 중 각 분류항목의 특징을 가장 잘 나타낼 수 있는 2개의 계수값(1,2번째 계수)을 선별하여 나타내었다. 계수값들의 변화를 살펴보면 음성, 음성+음악, 음악 순으로 격렬한 변화를 관찰할 수 있었다. 전체 12개의 계수값에서는 음성과 음성+음악의 계수값의 변화가 비슷한 면을 보였으며 음악은 다른 분류항목에 비해 계수값의 변화폭이 적었다.

2.2 에너지

에너지는 계산비용이 저렴한 특징 파라미터이다. 오디오 신호는 8kHz로 샘플링 되었다, 프레임은 25ms 의 길이를 가지며, 25ms의 윈도우를 10ms씩 이동시켜가면서 파라미터를 추출하였다. 연속적인 5개 프레임 구간에 대해서 델타를 적용하였다. 에너지의 특징 파라미터로는 로그 에너지와 델타 로그 에너지를 사용하였다. 에너지는 음성/음악 분류기가 레벨 정보에 의해 분류하는 것을 방지하기 위해서 정규화를 적용하였다.

그림 2는 각 분류항목별 정규화 로그 에너지의 변화를 나타내고 있다. 음성의 정규화 로그 에너지의 변화는 음악과 음성+음악의 정규화 로그 에너지의 변화에 비해 격렬한 것을 볼 수 있다. 특히 발성간의 무음구간에서는 정규화 로그 에너지의 변화가 크게 일어났다.

2.3 영교차

영교차는 계산비용 대 효율비가 좋은 특징 파라미터이다. 본 실험에서는 영교차를 기반으로 하는 4가지 특징 파라미터들(영교차율, 영교차율 변화량의 표준편차, 영교차 평균 상하값들간의 차, 영교차율의 삼차 모멘트)을 적용하였다 [2, 4].

오디오 신호는 8kHz로 샘플링 되었다. 한 프레임은 25ms의 길이를 가지며, 25ms의 윈도우를 10ms씩 이동시켜가면서 파라미터를 추출하였다.

각각의 영교차 특징들의 연속적인 5개 프레임 구간에 대해서 델타를 적용하였다. 영교차율을 제외한 나머지 특징들은 음성/음악 분류기가 레벨 정보에 의존하여 분류하는 것을 방지하기 위해서 정규화를 적용하였다.

그림 3은 영교차 특징 파라미터 중 영교차율의 각 분류항목별 변화를 나타낸 것이다. 그림에서 보면 음성과 음성+음악의 영교차율이 음악의 영교차율보다 격렬하게 변화하였다.

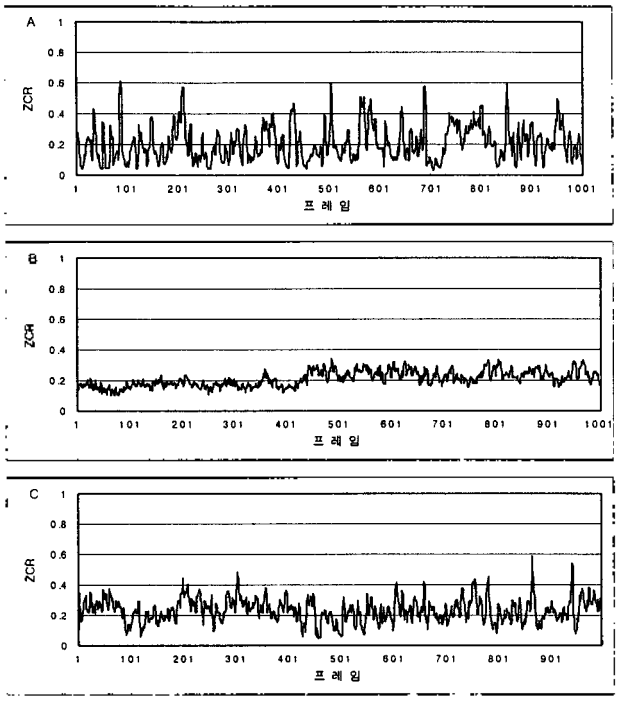

그림 3. 각 분류항목별 영교차율 변화.
A : 음성, B : 음악, C : 음성+음악

2.4 특징에 대한 고찰

음성의 특징 파라미터 수치변화는 타 분류항목에 비해 격렬하게 변화함으로써 뚜렷하게 구별되었다. 음악은 다른 특징 파라미터에 비해서 비교적 변화가 급격하지 않은 모습을 볼 수 있었다. 각 특징 파라미터들의 분류항목별 특징변화를 살펴보면 특징값들의 평균은 비슷한 반면, 그 값들의 변화는 분류항목별로 현격한 차이를 보이는 것을 알 수 있다. 이는 특징 파라미터들의 수치보다는 그 수치들의 변화가 분류에 큰 영향을 끼친다는 것을 의미한다.

III. 실험 구성

3.1 데이터베이스 구축

실험을 위한 데이터베이스는 음성, 음악, 음성+음악으로 구성되었다. 전체적으로 8kHz의 샘플링 주파수를 적용하였다.

음성 데이터베이스는 3가지 분류기준(무소음시, 배경소음 존재시, 타 화자의 음성 혼재시)을 적용하여 구성하였다. 음악 데이터베이스는 뉴스와 대중가요에서 3가지 분류기준(발라드, 락, 클래식)을 적용하여 구성하였다. 음성+음악 데이터베이스는 뉴스와 대중가요에서 3가지 분류기준(발라드+음성, 락+음성, 클래식+음성)을 적용하여 구성하였다.

음성 트레인 데이터베이스는 3개 뉴스에서 각각 평균 1분의 30개 신호씩 총 90개 신호, 음악 트레인 데이터베이스는 평균 45초의 36개 신호, 음성+음악 트레인 데이터베이스는 평균 30초의 37개 신호로 구성되어 있다.

테스트 데이터베이스는 3개 뉴스에서 평균 1분의 음성신호 82개, 평균 45초의 음악신호 21개, 평균 45초의 음성+음악신호 18개의 신호를 사용하였다. 이 중 음성+음악 데이터베이스 중 9개는 대중가요에서 나머지 9개는 실제 뉴스에서 추출하였다. 차후 2가지 분류항목 적용시 뉴스에서 추출한 신호 9개는 음성으로 대중가요에서 추출한 신호 9개는 음악으로 분류하였다.

신호의 시작과 끝 부분의 무음구간은 제거하였으나 신호의 중간에 존재하는 대화간의 무음구간에 대해서는 3초 이내의 것에 한하여 허용하였다.

3.2 실험 시스템

음성/음악 분류는 GMM을 이용하였다. 음성/음악 분류기에서는 오디오 신호를 3가지(음성, 음악, 음성+음악)와 2가지(음성, 음악)로 분류하였다.

오디오 신호의 분류항목 결정은 각 분류항목별 트레인 DB에 대해 Expectation Maximum(EM)알고리즘을 적용하여 트레인을 한 GMM의 결과값을 입력 오디오 신호에 대응시켜 그 중 최대의 확률을 가지는 것을 선택하였다. 테스트는 각각의 특징에 대해 동일한 3가지 Mixture(16, 32, 64)를 적용하였다.

IV. 실험 결과

4.1 멜캡스트럼

실험결과 3가지 분류항목 적용시에는 16 Mixture에서, 2가지 분류항목 적용시에는 모든 Mixture에서의 결과가

표 2. 멜캡스트럼 특징을 적용한 GMM 분류 결과(백분율) S : 음성, M : 음악, S+M : 음성+음악.

구 분	16 mixture			32 mixture			64 mixture		
	S	M	S+M	S	M	S+M	S	M	S+M
S	92.6	0	7.4	93.9	0	6.1	93.9	0	6.1
M	0	76.1	23.9	0	71.4	28.6	0	61.9	38.1
S+M	33.3	11.2	55.5	33.3	11.2	55.5	33.3	5.6	61.1

구 분	16 mixture		32 mixture		64 mixture	
	S	M	S	M	S	M
S	100	0	100	0	100	0
M	3.4	96.6	3.4	96.6	3.4	96.6

표 3. 에너지 특징을 적용한 GMM 분류 결과(백분율).
S : 음성, M : 음악, S+M : 음성+음악

구 분	16 mixture			32 mixture			64 mixture		
	S	M	S+M	S	M	S+M	S	M	S+M
S	90.2	0	9.8	87.8	0	12.2	89.0	0	11
M	38.2	14.2	47.6	38.2	14.2	47.6	38.2	14.2	47.6
S+M	19.5	25	55.5	11.2	22.2	66.6	11.2	22.2	66.6

구 분	16 mixture		32 mixture		64 mixture	
	S	M	S	M	S	M
S	92.3	7.7	93.4	6.6	92.3	7.7
M	40	60	40	60	40	60

양호하였다. 멜캡스트럼은 에너지 특징 파라미터에 비해 음악 분류에 있어서 3가지 분류항목 적용시의 결과가 좋지 않았다.

4.2 에너지

에너지는 3가지 분류항목을 적용하였을 경우 64 mixture에서, 2 가지 분류항목을 적용하였을 경우 32 mixture에서 가장 좋은 결과를 나타내었다. 분류 결과에 의하면 에너지는 타 특징에 비해 음악에 대한 분류능력이 현저히 낮았다.

3가지 분류항목 적용시에는 음성+음악의 분류능력은 타 특징들에 비해 양호한 것으로 나타났으나 음악의 분류능력은 가장 낮은 것으로 나타났다. 2가지 분류항목 적용시에는 음성의 분류능력에서는 양호한 결과를 나타내고 있으나, 음악의 분류능력은 타 특징들에 비해 현저히 낮았다.

4.3 영교차

영교차는 2가지 분류방법 모두 16 Mixture에서 가장 좋은 결과를 나타내었다.

결과를 살펴보면 2가지 분류항목 적용시 음성과 음악 모두에서 양호한 분류결과를 보였다. 3가지 분류항목 적용시에는 타 특징에 비해 음악의 분류능력이 가장 우수한 것으로 나타났다. 이에 반해 음성+음악의 분류능력은 타 특징에 비해 가장 떨어지는 것으로 나타났다.

Ⅳ. 결 론

본 논문에서는 오디오 신호에서 정보를 가지고 있는 음성을 추출하는 음성/음악 분류기에 대해 다루었다. 음성/음악 분류기에서는 멜캡스트럼, 에너지, 영교차를 특징 파라미터로 사용하였다.

표 4. 영교차 특징을 적용한 GMM 분류 결과(백분율).
S : 음성, M : 음악, S+M : 음성+음악

구 분	16 mixture			32 mixture			64 mixture		
	S	M	S+M	S	M	S+M	S	M	S+M
S	85.3	7.3	7.4	80.4	7.3	12.3	81.7	6	12.3
M	0	95.2	4.8	0	95.2	4.8	0	95.2	4.8
S+M	44.5	50	5.5	44.5	50	5.5	44.5	50	5.5

구 분	16 mixture		32 mixture		64 mixture	
	S	M	S	M	S	M
S	90.1	9.9	90.1	9.9	90.1	9.9
M	10	90	10	90	10	90

분류 알고리즘은 GMM을 사용하였으며, GMM의 mixture는 16, 32, 64를 일률적으로 적용하였다. 음성/음악 분류기는 오디오 신호를 3가지(음성, 음악, 음성+음악)와 2가지(음성, 음악)로 분류하였다. 실험결과 특징 파라미터 중 멜캡스트럼이 가장 좋은 결과를 보였다.

실험결과를 살펴보면 특징 파라미터의 장단점이 상호 보완적인 관계에 있는 경우가 있기 때문에 특징간의 상호조합을 통한 음성/음악 분류기의 성능향상이 기대된다.

참고문헌

[1] Michael J. Carey, 'A Comparison of Features for Speech, Music Discrimination', Proc. ICASSP 1999, Vol 1.

[2] John. Saunders, 'Real-Time Discrimination of Broadcast Speech / Music', Proc. ICASSP 1996, pp993-996.

[3] E. Scheier and M. Slaney, 'Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator', Proc. ICASSP 1997, pp1331-1334.

[4] B. Kedam, 'Spectral Analysis and Discrimination by Zero-Crossing', Proc. IEEE Vol. 74 No. 11 Nov 1986, pp 1477-1493.

[5] John D. Hoyt, 'Detection of Human Speech in Structured Noise', Proc. IEEE 1994, pp 237-240.

[6] T. Hain 'Segment Generation and Clustering in the HTK Broadcast News Transcription System'. In Proceedings of the Broadcast News Transcription and Understanding Workshop, 1998.