실험에 의한 음성·음악 분류 특징의 비교 분석

Comparison & Analysis of Speech/Music Discrimination Features through Experiments

이경록, 류시우*, 곽재영*
남부대학교, (주)세오전자*

Lee Kyung-Rok, Ryu Shi-Woo*,
Gwark Jae-Young*
Nam Bu Univ, SEO electronics*

요약

본 논문에서는 각 특징 파라미터 조합의 음성/음악 분류 성능을 비교 분석하였다. 음향신호는 3가지(음성, 음악, 음성+음악)로 분류하였다. 본 실험에서는 분류 특징으로 멜캡스트럼, 에너지, 영교차 3가지 형태가 사용되었다. 음성/음악 분류 성능이 가장 좋은 특징간의 상호 조합을 비교 분석하였다. 실험결과 멜캡스트럼, 영교차 조합이 가장 좋은 결과(음성: 95.1%, 음악: 61.9%, 음성+음악: 55.5%)를 보인다는 것을 확인할 수 있었다.

Abstract

In this paper, we compared and analyzed the discrimination performance of speech/music about combinations of each features parameter. Audio signals are classified into 3 classes (speech, music, speech and music). On three types of features, Mel-cepstrum, energy, zero-crossings used to the experiments. Then compared and analyzed the best of the combinations between features to speech/music discrimination performance. The best result is achieved using Mel-cepstrum, energy and zero-crossings in a single feature vector (speech: 95.1%, music: 61.9%, speech & music: 55.5%).

1. 서 론

최근 급속한 정보 기술의 발달로 대용량 멀티미디어 데이터 베이스 구축이 일반화 되고 있다. 이러한 데이터베이스에서의 멀티미디어 인덱싱 기반 정보 검색 기술이 활발히 연구되고 있다.

오디오 인덱싱은 음향 신호에서 정보를 가진 부분(음성, 폭발음, 함성 등)을 분리하고 내용를 구분하는 방식이다.

본 논문에서는 차후 음성인식을 통한 내용 분석에 필요한 음성부분만을 추출하는 오디오 인덱싱의 전처리부격인 음성/음악 분류기에 사용되는 특징들의 상호조합에 대하여 실험하였다.

음성/음악 분류기에 사용되는 특징으로서는 멜캡스트럼, 피치, 에너지, 영교차가 가장 널리 사용되고 있다[1]. 실험에서는 이들 중 3가지 특징(멜캡스트럼, 에너지, 영교차)들을 선정하고, 특징들간의 상호조합을 통해 가장 양호한 결과를 생성하는 조합을 찾아내는데 중점을 두었다.

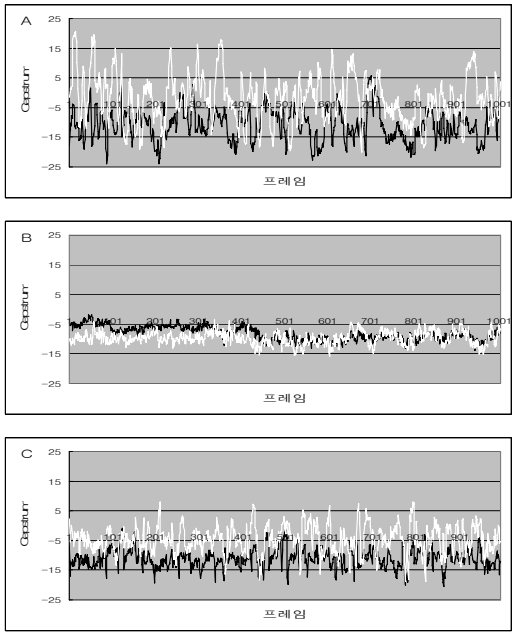
음성/음악 분류기에서 사용되는 분류 알고리즘으로는 Gaussian mixture model, k-nearest-neighbors classification, 벡터 양자화, 결정 트리 등이 있으나 본 논문에서는 GMM을 사용하였다[1-3].

2. 특징 파라미터

각 특징 파라미터의 분석을 위해 음향 신호는 8kHz로 표본을 추출하고, 한 프레임의 크기는 25ms이며 10ms씩 이동시켜가면서 특징 파라미터를 추출하였다. 델타 파라미터는 연속적인 5개 이상의 프레임 구간에서 추출하였다.

2.1 멜캡스트럼

특징 파라미터로는 12개의 멜캡스트럼 계수들과 12개의 델타 멜캡스트럼 계수들을 사용하였다.



▶▶ 그림 1. 각 분류항목별 멜캡스트럼 변화 (1,2번째 계수값). a : 음성, b : 음악, c : 음성+음악

분석을 위해 사용된 캡스트럼 수식은 다음과 같다.

$$C_x(m) = \frac{1}{N}\sum_{k=0}^{N-1}\log|X(k)|\, e^{j2\Pi\frac{mk}{N}} \qquad (1)$$

식 (1)의 대수진폭은 아래와 같다.

$$\log|X(k)| = \sum_{m=0}^{N-1} C_x(m)\, e^{-j2\pi\frac{mk}{N}} \qquad (2)$$

캡스트럼 파라미터는 멜 스케일 특성을 가진 26개의 필터로 구성된 필터뱅크로 필터링을 실시하였다.

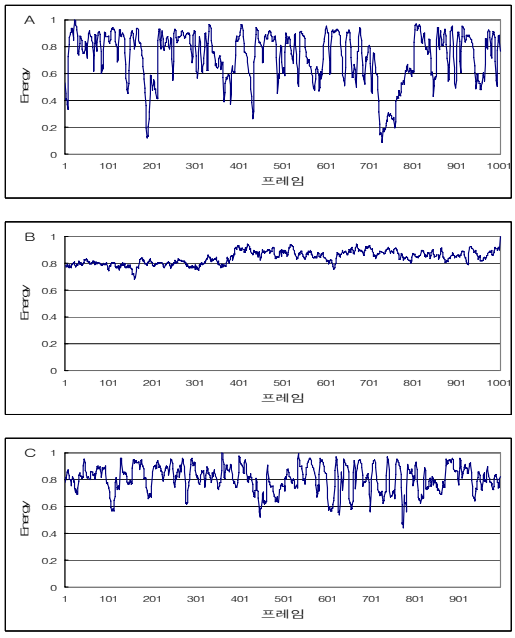
멜캡스트럼은 차후 화자변화탐색 분야에서도 활용가능하며, 효율면에서나 분류능력면에서도 가장 양호한 것으로 알려져 있다[1].

그림 1은 멜캡스트럼 계수값의 변화를 나타낸 것으로 12개의 계수값 중 각 분류항목의 특징을 가장 잘 나타낼 수 있는 2개의 계수값만을 선별하여 나타내었다(1, 2번째 계수값).

선별된 멜캡스트럼 계수값의 변화를 살펴보면 음성, 음성+음악, 음악 순으로 격렬한 변화를 보인다.

2.2 에너지

에너지의 특징 파라미터로는 정규화 로그 에너지와 델타 정규화 로그 에너지를 사용하였다.



▶▶ 그림 2. 각 분류항목별 정규화 로그 에너지 변화. a : 음성, b : 음악, c : 음성+음악

사용된 에너지 파라미터의 수식은 다음과 같다.

$$E = \frac{1}{E_{\max}} \left(\log \left(\frac{1}{N} \sum_{k=0}^{N-1} x(k)^2 \cdot h(m-k) \right) \right) \quad (3)$$

로그 에너지는 음성/음악 분류기가 크기 정보에 의존하여 분류하지 않도록 정규화를 실시하였다.

그림 2는 각 분류항목별 프레임 변화에 따른 정규화 로그 에너지 변화를 나타내고 있다. 음성의 변화가 음악과 음성+음악에 비해 훨씬 격렬하다.

2.3 영교차

영교차는 계산비용 대 성능비가 우수한 것으로 알려져 있다. 영교차를 기반으로 하는 4가지 특징은 영교차율, 영교차율 변화량의 표준편차, 영교차 평균보다 높은 값과 낮은 값들간의 차, 영교차율의 제삼 모멘트이다[2, 4].

실험에 사용된 영교차의 수식은 다음과 같다.

$$Z_m = \sum_m |S[x(n)] - S[x(n-1)]| \cdot h(n-m) \quad (4)$$

이때, $S[x(n)]$은 $x(n)$이 0 이상일 때만 1이 된다.

영교차 특징들은 음성/음악 분류기가 크기 정보에 의존하여 분류하는 것을 방지하기 위해서 영교차율을 제외한 나머지 3가지 특징들에 대해서 정규화를 실시하였다.

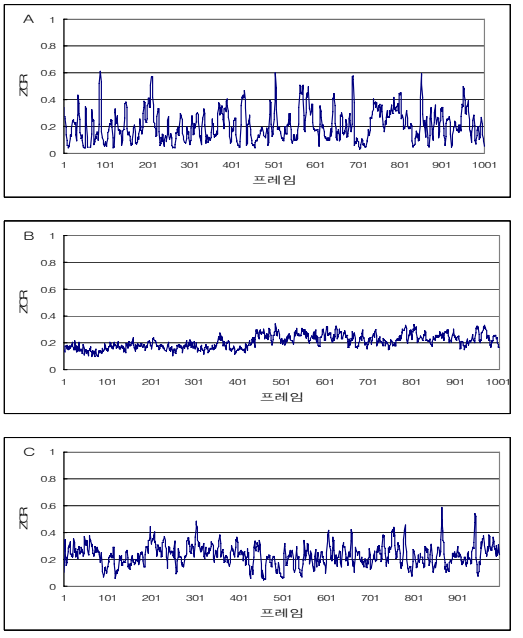
[2]에서는 분석 윈도우를 이동함에 있어 중첩을 허용하지 않았지만 각 특징들간의 상호조합을 지원하기 위해서 동일하게 중첩을 적용하였다.

중첩 적용 전과 후의 분류능력의 변화를 살펴보면 표 1과 같다. 표 1에서 알 수 있듯이 음악에서의 분류능력은 향상된 반면, 음성과 음성+음악의 분류 결과가 나빠졌음을 확인할 수 있다.

[표 1] 영교차의 중첩 적용여하에 따른 분류결과(백분율). S:음성, M:음악, S+M:음성+음악

구 분		영교차(중첩적용전)			영교차(중첩적용후)		
		16Mix	32Mix	64Mix	16Mix	32Mix	64Mix
분류항목	S	87.8	79.2	82.9	85.3	80.4	81.7
	M	95.2	71.4	61.9	95.2	95.2	95.2
	S+M	11.1	11.1	16.6	5.5	5.5	5.5

그림 3은 중첩이 적용된 상태의 영교차 특징 파라미터 중 영교차율의 각 분류항목별 변화를 나타낸 것이다. 음성과 음성+음악의 영교차율이 음악의 영교차율보다 격렬하게 변화하는 것을 알 수 있으며, 특히 음성의 영교차율은 음성+음악의 영교차율보다 큰 차이로 변화하는 것을 볼 수 있다.



▶▶ 그림 3. 각 분류항목별 영교차율 변화.
a : 음성, b : 음악, c : 음성+음악

3. 실험 구성

3.1 데이터베이스 구축

공중파 방송의 50분 분량의 뉴스 3회분, 대중가요, 기타 음악파일들로 데이터베이스를 구성하고 음성, 음악, 음성+음악의 3가지로 분류하였다.

각각의 데이터베이스는 적용한 분류기준별로 균등한 비율로 구성되어 있다. 음성 데이터베이스는 방송뉴스을 대상으로 3가지 분류기준(무소음시, 배경소음 존재시, 다른 화자의 음성 혼재시)을 적용하였다. 음악 데이터베이스는 방송뉴스와 대중가요를 대상으로 3가지 분류기준(발라드, 락, 클래식)을 적용하였다. 음성+음악 데이터베이스는 방송뉴스와 대중가요를 대상으로 3가지 분류 기준(발라드+음성, 락+음성, 클래식+음성)을 적용하였다.

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

테스트 데이터베이스는 3개 뉴스에서 평균 1분 정도의 음성신호 82개, 평균 45초 정도의 음악신호 21개, 평균 45초 정도의 음성+음악신호 18개로 구성되었다. 이 중 음성+음악신호 데이터베이스는 총 18개 중 9개는 대중가요에서 나머지 9개는 실제 뉴스에서 추출하였다.

음향 신호의 시작과 끝의 무음구간은 제거하였으나 음향 신호의 중간에 존재하는 발성간의 무음구간에 대해서는 3초 이내의 것에 한하여 허용하였다[6].

3.2 실험 시스템

음성/음악을 분류하기 위한 알고리즘으로는 GMM을 이용하였고, 음성/음악 분류기에서는 입력 음향 신호를 3가지(음성, 음악, 음성+음악)로 분류하였다. 이는 오디오 인덱싱에서 배경음악 속에서의 음성신호 인식의 필요성(광고방송, 방송뉴스의 시작부분 등)이 제기됨에 따른 것이다.

음향 신호의 분류 결정은 각 트레인 데이터베이스에 대해 학습시킨 분류항목별 GMM의 모델을 입력 음향 신호에 대응시켜 그 중 최대의 확률을 가지는 것을 선택하였다. 테스트는 각각의 특징에 대해 동일한 3가지 mixture(16, 32, 64)를 적용, 상호비교 하였다.

4. 실험 결과

실험은 우선 각 특징들을 독립적으로 실험한 결과를 분석하여 상호조합에 반영함으로써 분류능력향상을 도모하였다.

4.1 멜캡스트럼

실험결과 16 mixture에서 결과가 양호하였으나, 음악 구분에 있어서 영교차 특징에 비해 좋지 않았다.

4.2 에너지

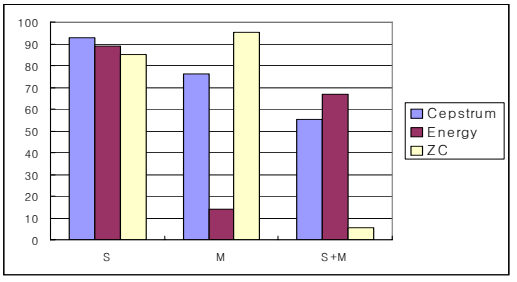
에너지는 64 mixture에서 가장 좋은 결과를 나타내었다. 분류 결과에 의하면 음성+음악 분류능력은 타 특징들에 비해 양호하였으나 음악의 분류능력은 가장 낮았으며, 음악 분류에서 다른 특징들과 달리 음성으로의 오분류 현상이 나타났다. 음악이 오분류 되는 양상을 살펴보면 평균적으로 오분류된 것 중 38.2%가 음성으로 47.6%가 음성+음악으로 분류되었다.

4.3 영교차

GMM의 mixture 수는 16, 32, 64 mixture를 적용하여 실시하였다.

표 4를 보면 타 특징에 비해 음악의 분류능력이 가장 우수한 것으로 나타났다. 이에 반해 음성+음악 분류능력은 타 특징에 비해 가장 떨어지는 것으로 나타났다. 또한 음성 분류에서 타 특징과 달리 음악으로의 오분류 현상이 나타났다. 영교차 특징 적용시 음성+음악이 오분류되는 양상을 살펴보면 평균적으로

오분류된 것 중 44.5%가 음성으로 분류되었고, 50%가 음악으로 분류되었다.



▶▶ 그림 4. 각 특징별 음성/음악 분류결과
(음성, 음악, 음성+음악)

4.4 특징간의 상호조합

특징들간의 상호조합은 다음와 같이 실시하였다.

에너지+영교차 조합은 표 2와 같이 32mixture에서 좋은 결과를 나타내었다. 또한, 에너지 단독적용시에 비해서 음성과 음성+음악 분류에서는 성능저하가 관찰되었고, 음악 분류에서는 큰 성능향상이 관찰되었다. 영교차 단독 적용시에 비해서는 음성과 음성+음악 분류 결과는 향상되었지만, 음악 분류 결과가 크게 저하되었다.

멜캡스트럼+영교차 조합에서는 표 3과 같이 64 mixture에서 가장 좋은 결과를 나타내었다. 또한 멜캡스트럼 단독 적용시에 비해 음성, 음악 분류능력의 향상이 관찰되었고, 영교차 단독 적용시에 비해서 음성, 음성+음악 분류능력이 향상되었다.

멜캡스트럼+에너지 조합에서는 표4와 같이 32 mixture에서 좋은 결과를 나타내었다. 또한, 맬캡스트럼 단독 적용시에 비해 음악, 음성+음악 분류능력이 저하되었으며, 에너지 단독 적용시에 비해서는 음성, 음악 분류능력이 향상되었다.

멜캡스트럼+에너지+영교차의 조합에서는 표 5와 같이 32 mixture에서 가장 좋은 결과를 나타내었다.

실험에 의하면 특징 상호간 조합은 특징 단독으로 사용하였을 때보다 더 좋은 결과를 나타내고 있다.

결과에서 음악이 음성으로 오분류되는 것보다 음성이 음악으로 오분류되는 것이 더욱 치명적이다. 음악으로 분류되면 차후 실험에서 제외되기 때문이다.

연산시간면에서는 에너지(2초), 영교차(3.4초), 멜캡스트럼(7.4초)순으로 연산시간이 소요되었고, GMM의 mixture수의 증가에 비례하여 연산시간도 비례적으로 증가하였다. 이는 실험이 실시간성을 요구하지 않으므로 문제되지 않는다.

[표 2] GMM 분류결과 (에너지+영교차, 백분율).
S:음성, M:음악, S+M:음성+음악.

구 분		16mixture	32mixture	64mixture
분류항목	S	89.0	85.3	86.5
	M	76.1	85.7	80.9
	S&M	22.2	22.2	22.2

[표 3] GMM 분류결과(멜캡스트럼+영교차, 백분율).
S:음성, M:음악, S+M:음성+음악.

구 분		16mixture	32mixture	64mixture
분류항목	S	96.3	93.9	95.1
	M	80.9	76.1	85.7
	S&M	44.4	55.5	44.4

[표 4] GMM 분류 결과(멜캡스트럼+에너지, 백분율).
S:음성, M:음악, S+M:음성+음악.

구 분		16mixture	32mixture	64mixture
분류항목	S	96.3	93.9	95.1
	M	52.3	57.1	47.6
	S&M	61.1	61.1	66.6

[표 5] GMM 분류결과(멜캡스트럼+에너지+영교차, 백분율). S:음성, M:음악, S+M:음성+음악.

구 분		16mixture	32mixture	64mixture
분류항목	S	96.3	95.1	95.1
	M	71.4	71.4	61.9
	S&M	44.4	55.5	55.5

5. 결 론

본 논문은 음성/음악 분류 특징에 대해 다루었다.

음성/음악 분류에 사용된 특징들은 멜캡스트럼, 에너지, 영교차간의 상호조합이다. 분류 알고리즘은 GMM을 사용하였으며, 3가지의 mixture(16, 32, 64)를 일률적으로 적용하였다.

실험결과 각 특징들을 단독으로 적용한 분류결과에 비해 특징간 상호조합을 적용한 분류결과가 보다 양호하였으며, 멜캡스트럼+영교차 조합이 가장 양호한 결과를 보인다는 것을 확인할 수 있었다.

▌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.

[7] Jean-Luc Gauvain, Lori Lamel, "Partitioning and Transcription of Broadcast News Data", Proc. ICSLP 1998 Sydney.