식별함수를 이용한 오디오신호의 내용기반 분류

김영섭*·이광석**·고시영***·허강인*

*동아대학교 · **진주산업대학교 · ***경일대학교

Content Based Classification of Audio Signal using Discriminant Function

Young-Sub Kim*, Kwang-Seok Lee**, Si-Young Koh***, Kang-In Hur*

*Dept. of Electronic Engineering, Dong-A University

**Dept. of Electronic Engineering, JinJu National University

***Dept. of Electronic & Information Engineering, Kyung-il University

E-mail : sad6513@nate.com

요 약

본 논문은 오디오 색인 · 검색 시스템을 구현하기 위하여 오디오 신호에 대한 특징 파라미터 풀(pool)을 구성하고, 구성되어진 특징 파라미터 풀을 이용한 오디오 데이터의 내용분석 및 분류에 관한 연구이다. 오디오 데이터는 기본적으로 다양한 형태의 오디오 신호로서 분류되어진다. 본 논문에서는 오디오 데이터의 분류에 이용 가능한 특징 파라미터를 분석하고 추출하는 방법에 대하여 논한다. 그리고 특징 파라미터 풀을 색인 그룹 단위로 구성하여 오디오 카테고리에 대한, 설정된 특징들의 포함 정도와 색인기준을 오디오 데이터의 내용을 중심으로 비교, 분석한다. 그리고 마지막으로 위의 결과를 바탕으로 분류카테고리 별로 오디오 데이터의 특징 벡터를 구성한 뒤 이를 이용하여 식별함수 분류기를 통한 분류를 실험한다.

ABSTRACT

In this paper, we research the content-based analysis and classification according to the composition of the feature parameters pool for the auditory signals to implement the auditory indexing and searching system. Auditory data is classified to the primitive various auditory types. we described the analysis and feature extraction method for the feature parameters available to the auditory data classification. And we compose the feature parameters pool in the indexing group unit, then compare and analysis the auditory data centering around the including level and indexing criterion into the audio categories. Based on this result, we composit feature vectors of audio data according to the classification categories, then experiment the classification using discrimination function.

키워드

Audio, Audio retrieval, Audio classification

Ⅰ. 서 론

최근 들어 대용량 디지털 오디오 데이터베이스가 다양해지고 있으며, 오디오의 내용 분석에 따른 오디오 데이터베이스에 대한 효과적인 관리의 중요성이 인식되어지고 있다. 그럼에도 불구하고 오디오신호가 가지는 동적인 특성과 다양성으로 인해 멀티미디어 스트림의 오디오 색인 · 검색에 관한 연구는 내용기반 이미지나 비디오 데이터베이스에 대해 행해지는 연구와 비교하여 매우 미흡한 실정이다. 그러나 다양한 멀티미디어 요소들의 통합적인 색인 · 검색에 대한 연구가 비디오를 완전하게 파싱(parsing)하기 위해서 해결해야 될 필수적인 과제이므로 내용기반 오디오 데이터에 대한 색인 · 검색 연구는 지속되어질 필요성이 있다. 본 논문에서는 이미 연구되어 왔던 음성이나 음악과 같은 오디오형태의 기본적인 분류카테고리를 확장하여, 이전의 연구에서는 소홀히 다루어져 왔던 노래를 포함시키고, 배경 오디오의 구별을 강조하여 음악과 음성만으로 구성된 카테고리

뿐만 아니라 배경음이 음악인 음성, 배경음이 음악인 노래 등과 같이 다양한 사운드를 분류대상으로 설정하였다. 설정된 대상에 따라 최적의 색인 · 검색을 위하여 일차적으로 오디오 신호분석을 통하여 각 카테고리의 특징을 비교 · 분석한 후, 이를 바탕으로 몇 가지 특징을 제안하고 제안된 오디오 신호의 특징을 기반으로 하여 특징파라미터 풀을 구성한 뒤, 분류실험을 행하였다. 본 논문은 다음과 같이 구성되어진다. 2장에서는 오디오특징 파라미터 풀에 포함 가능한 여러 가지 파라미터들을 오디오 신호의 내용에 기반하여 소개하고, 그 특징의 계산과 추출방법을 소개한다. 3장에서는 2장에서 언급한 특징들을 이용하여 특징 파라미터 풀을 구성하고 이를 바탕으로 베이즈정리(Bayes theorem)에 기반한 통계적 식별함수 분류기에 의한 분류에 대하여 설명한 뒤 4장에서 실험 및 결과, 그리고 5장에서 결론을 맺는다.

II. 오디오 신호의 특징 파라미터들

본 논문에서의 특징 파라미터는 오디오 신호의 속성에 기반한 파라미터 풀을 구성하여 이용한다. 본 장에서는 신호처리에서 일반적으로 사용되는 특징 파라미터뿐만 아니라 "배음도(HD)", "주파수 지속도(FDD)", "주파수 집중도(FCD)"라는 특징을 제안하고 오디오 데이터 분류 파라미터로 적용가능한가를 분석한다.

2.1 단구간 평균 에너지

오디오 신호의 에너지는 일반적으로 식 (1)과 같이 자승평균 에너지로 정의하는데, 신호 파형의 변화정도를 크게 하여 임계값에 여유를 줄 경우에는 식 (2)을 사용한다.

$$E_n = \frac{1}{N}\sum_{m=0}^{N-1} X^2(m) \tag{1}$$

$$E_n(dB) = 10\log E_n \tag{2}$$

에너지는 일반적으로 유/무성음에서 보여주는 차이로 인해, 에너지에 임계값을 주어 묵음을 분리할 수 있는 유용한 정보가 된다. 또한 음악을 포함한 오디오 신호에서의 에너지는 순수한 음성신호보다 높은 값을 가지기 때문에 오디오신호의 분류 특징으로 유용하게 사용될 수 있다.

2.2. 단구간 평균 영교차율

영교차는 이산신호의 인접 샘플이 다른 부호를 가질 경우에 나타나며 식 (3)과 같이 정의한다.

$$L_n = \frac{1}{2}\sum_{m=0}^{N-1} |sgn[x(n-m)] - sgn[x(n-m-1)]|$$

$$where,\quad sgn[s(n)] = 1, \quad s(n) \geq 0 \tag{3}$$

$$= -1, \quad s(n) < 0$$

오디오 신호는 ZCR커브 특성의 규칙성, 주기성, 안정성 그리고 진폭의 범위와 분산 값은 매우 다른 특성들을 보여준다. 특히 순수한 음성은 ZCR의 4가지의 조건으로 분류 가능하다. 첫 번째 조건은 Fig. 1과 같이 ZCR과 에너지의 순간적 커브간의 상호 배타적인 관계이다. 이것을 이용하여 다음과 같이 음성신호를 구별한다. 즉, 최대진폭의 2/3 지점에서 ZCR과 에너지 커브를 클리핑하여 적은 부분을 제거하고 두 커브의 단일 피크부분만을 남도록 한 뒤 두 신호 잔차커브의 내적을 계산한다. 음성의 경우, 이 내적은 대개 에너지와 ZCR의 피크가 다른 시간에 존재하므로 "0" 근방의 값을 가지는 반면에, 다른 오디오 형태에 대하여는 큰 값을 가진다.

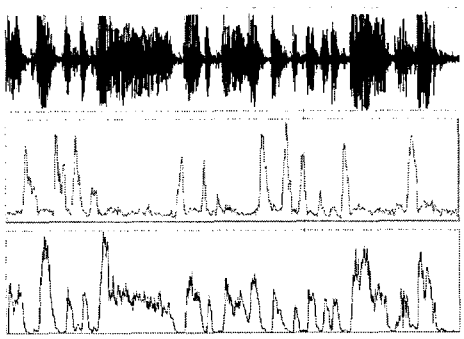

Fig. 1. Exclusive relation between energy and ZCR for the speech

두 번째는 ZCR커브의 모양이다. 음성은 ZCR커브의 범위가 크며 낮은 하한선을 가진다. 세 번째와 네 번째는 각각 ZCR커브 진폭의 범위와 분산이다. 음악 세그먼트의 경우에는 평균 ZCR이 음성에 비해 Fig. 2와 같이 특정 범위내의 단구간에서 훨씬 안정된 형태이고 파형의 변화는 불규칙적이며 진폭이 음성에 비해 적은 범위를 가지기 때문에 진폭의 범위와 분산이 적은 값을 가지게 되는 반면에, 음성의 경우에는 큰 값을 가지게 된다.

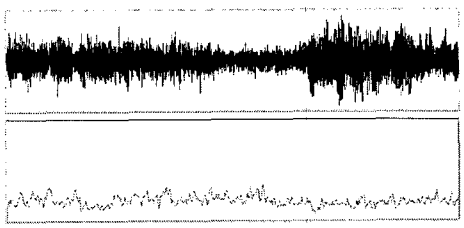

Fig. 2. The stable ZCR of Music signal

2.3. 배음도 (Harmonic Degree)

사운드는 기본 주파수와 그것의 정수배의 주파수인 배음을 가진 사운드 Fig. 3과 그렇지 않는 사운드 Fig. 4로 나눌 수 있다. 음성의 경우, 구성 요소들이 고조파적인 유성음과 비고조파적인 무성음의 혼합된 형태의 사운드이다. 반면에, 악기에 의한 음악적 요소를 가진 사운드는 대부분 배음을 가진 사운드에 해당한다.

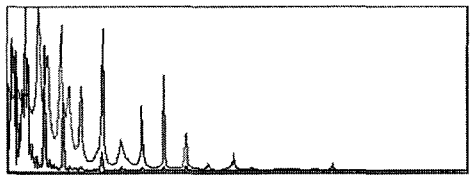

Fig. 3. Music signal with harmonics

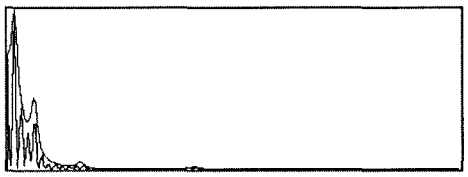

Fig. 4. Sound with non-harmonics

오디오 데이터에 대한 고조파적인 배음의 포함 유무검출은 AR모델로 신호의 스펙트럼을 추정함으로 가능하다. Fig. 3, 4은 Durbin법으로 추정한 주파수 스펙트럼을 보여준다. 단구간에서의 고조파적인 특성을 찾기 위해서는 AR모델로 스펙트럼의 포락선을 추정하여 피크부분을 검출하는 방법이 직접 신호의 스펙트럼을 계산하는 경우보다 피크부분의 추출이 용이하다. 실제 피크의 형태적인 특성은 그림에서 알 수 있듯이 비고조파보다 고조파의 경우에 두드러진 피크를 가진다. 만약 AR모델로 추정한 오디오 신호의 세그먼트 내에 날카롭고 주기적인 피크를 가지면 이 세그먼트는 배음을 가진 사운드로 음악적 요소를 가진 것으로 분류 가능하다. 사운드를 단구간 분석하여 해당 프레임이 배음을 가지는 경우에 음악적 요소를 가지는 것으로 인덱스 값을 1, 그렇지 않으면 0으로 설정한다. 인덱스 열에서 0의 수와 인덱스 총수의 비율을 "배음도 (이하 "HD")"라고 정의한다. 이 파라미터는 사운드에 포함되는 음악 성분이 적을수록 이 비율이 낮아지며 음악적 성분 포함 유무를 측정할 수 있는 중요한 파라미터가 된다. 배음도의 계산은 Fig. 5에서와 같은 알고리즘으로 계산하며 0-1사이의 값을 가진다. 본 논문에서 오디오 데이터의 배경음은 대부분 음악적 요소를 가진 것으로서, 배경음이 음악인 경우로 제한한다.

2.4. 단구간 기본 주파수 지속도 (FuF. Duration Degree)

단구간 기본 주파수 지속도 (이하 "FDD")는 일정 대역의 단구간 기본 주파수 값이 임계 프레임 이상 지속되는 구간의 총수로 정의한다. 단구간 기본 주파수의 지속도는 일정 피치가 지속되는 정도를 나타내므로 음악적 요소가 포함된 범주를 분류하는데 효과적으로 이용 가능하다.

2.5. 주파수 집중도 (Freq. Convergence Degree)

주파수 집중도 (이하 "FCD")는 색인 그룹 내의 주파수의 누적치 평균값이 임계치 이상되는 주파수의 총수로 정의한다. 주파수 대역의 분포는 SPT(Spectral Peak Track)를 추출한 분포에서 취하는 방법, 그리고 직접 FFT하여 구하는 방법이 있다. 본 논문에서는 직접 FFT에 의한 방법을 사용하였다. 오디오 신호의 내용에 기반한 분석 측면에서 보면 주파수 집중도는 배경음으로 음악 신호가 포함되지 않은 순수한 노래 신호에서 높은 값을 나타냄을 알 수 있다.

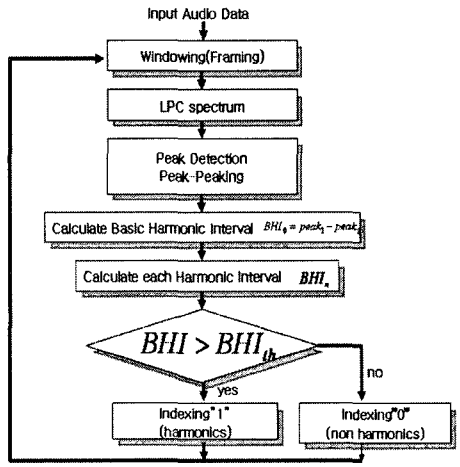

Fig. 5. Harmonic degree extraction algorithm

Ⅲ. 특징 파라미터 풀을 이용한 분류

3.1 분류 카테고리

멀티미디어 내에서 오디오 신호의 내용에 기반한 분류 카테고리는 묵음(Silence), 음악(Music), 음성(Speech), 노래(Song), 배경음이 음악인 음성(Back+Speech), 배경음이 음악인 노래(Back+Song), 효과음으로 나눌 수 있다. 그러나 효과음은 그 형태가 다양하고 분류의 기준이 되는 정량적인 요소가 부재하므로 분류에서 제외하였다.

3.2 파라미터 풀의 구성

입력된 오디오는 2장에서 언급한 특징파라미터 중에서 Table. 1과 같은 파라미터들로 풀을 구성한다. 본 연구에서는 단구간 기본 주파수의 지속도는 실시간 처리에 어려움이 있고, 다른 파라미터로 이 특징의 분류 기준은 대체 가능하므로 제외하였다.

Table 1. Analysis result of parameter pool.

구분	Speech	Music	Song	Back+ Speech	Back+ Song
Energy	161.8	704.8	435.3	295.9	1320.7
ZCR	20.4	25.4	27.5	21.9	27.3
ZCR Range	93	40.2	87.7	60.625	82.4
ZCR Var./10000	178	2	330	44	152
Inner Product	248.7	12866.6	2111.3	1424.45	8653.1
FCD	90.9	122	12301	90.95	60.1
HD	0.389	0.874	0.678	0.591	0.721

3.3 식별함수 분류기에 의한 분류

구성되어진 오디오 분류 카테고리에 대한 분류는 특징파라미터 풀로 구성되어진 7가지의 특징을 이용하여 각각의 클래스를 7차의 특징 파라미터로 구성하고, 이 값의 스케일을 일정하게 하기 위하여 로그값을 취한 뒤 행해진다. 식(4)로 표현되는 판별식은 주어진 데이터 집합을 이루는 확률밀도함수가 가우시안 분포와 같은 특정한 형태를 이루고 있다고 가정하고 확률밀도함수의 평균, 공분산을 파라미터로 추정하는 모수적 방법 중의 하나로, 베이즈의 이차판별 함수식을 사용한다.

$$g_i = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) \quad (4)$$
$$-\frac{1}{2}\log(|\Sigma_i|)+\log(P(w_i))$$

여기서 x는 입력 데이터로, 분류 대상의 특징 파라미터가 되고 μ_i는 클래스 각각의 학습 데이터의 평균, Σ_i^{-1} 는 학습 데이터의 클래스 각각의 공분산의 역행렬이 된다. 그리고 $|\Sigma_i|$는 학습 데이터의 공분산의 행렬식이 된다. 마지막으로 식(4)의 마지막 항 $\log(P(w_i))$는 w_i의 사전확률로 여기서는 동일하다고 가정하고 사용하지 않았다.

IV. 실험 및 결과

4.1 분석환경

본 논문에서 오디오 신호의 분석을 위하여 사용한 DB는 "친구" , "Sound of Music" 등의 영화 오리지널 사운드 트랙에서 16KHz, 16bit로 샘플링하여 5초간의 오디오 클립 204개를 만들어 내용에 따라 7가지로 분류, 구축하고 분석하였다.

모의 분류 실험에 사용한 데이터는 이 중에서 카테고리마다 각각 40개의 색인 단위를 취하여 실험하였으며, 색인 단위는 2초간의 데이터(530 frame)이 기본 색인 단위가 되도록 분석하였다. 분석 도구는 윈도우 API함수를 이용하여 자체적으로 제작된 환경에서 행하였다.

4.2 분류결과

분류 카테고리 중에서 묵음은 에너지와 ZCR에 임계값을 설정하여 별도로 분류하였으며 임계값보다 에너지가 낮고 ZCR은 높게 될 때를 묵음으로 판별하여 100%의 분류율(%)을 보였으며 나머지 카테고리에 대해서는 Table. 2와 같은 분류결과를 보였다.

Table 2. Classification result of Audio Data

구분	음성	음악	노래	배경+음악	배경+노래	분류율(%)
음성	37	0	1	1	1	92.5
음악	1	32	2	1	4	80
노래	6	0	31	1	2	77.5
배경+음악	0	4	3	29	4	72.5
배경+노래	3	0	4	2	31	77.5

V. 결론

본 논문에서는 실시간 오디오 색인 · 검색 시스템을 구현하기 위하여 필요한 기초 연구로 오디오 신호에 유용한 특징 파라미터들을 분석한 후, 이를 바탕으로 몇 가지 새로운 파라미터를 제안하였다. 그리고 이를 바탕으로 파라미터 풀을 구성한 후, 베이즈의 이차판별식에 의해 분류 실험을 행하였다. 결론적으로, 오디오 데이터와 같이 다양하고 동적인 패턴에 대한 분류를 위해서는 단일 파라미터를 이용하거나 혹은 다양한 특징파라미터를 유기적으로 조합하여 이용하는 것보다도 분류 단계에서 이러한 파라미터 풀 내의 각 파라미터들로 차원을 형성할 경우 높은 분류율(%)을 나타냄을 알 수 있었다.

참고문헌

[1] 한학용, 김수훈, 허강인, "오디오 데이터의 특징 파라미터 구성에 따른 내용기반 분석," 한국음향학회지. 제21권 제2호, pp. 182-189, 2002.

[2] M. J. Carey, "A Comparison of feature for Speech, Music Discrimination," Proc. ICASSP, vol. 1, pp. 145-152, 2. 1999.

[3] J. Saunders, "Real Time Discrimination of Broadcast Speech/Music," proc. ICASSP, vol. 2, pp. 141-144, 3. 1996.

[4] 이경록, 서봉수, 김진영, "오디오 인덱싱을 위한 음성/음악 분류 특징 비교," 한국음향학회지. 제20권 제2호, pp. 10-15, 4. 2001.

[5] Y. Medan, E. Vair and D. Chazan, "Super Resolution Pitch Determination of Speech Signals," IEEE Trans. On Signal Processing, vol. 39, no. 1, pp. 40-48, 5. 1991.