한국음향학회지 제27권 제5호 pp. 239~243 (2008)

부밴드 스펙트럼의 무게중심을 이용한 강인한 오디오 인식기

Robust Audio Identification Using Spectro-Temporal Subband Centroids

서 진 수*, 이 승 재**
(Jin Soo Seo*, Seungjae Lee**)

*강릉대학교 전자공학과, **한국전자통신연구원 SW 콘텐츠 연구부문
(접수일자: 2008년 6월 17일; 채택일자: 2008년 7월 25일)

본 논문에서는 스펙트럼의 주파수 및 시간 방향의 특성을 결합한 오디오 인식 방법을 제안하였다. 특히 스펙트럼의 형태를 모사하기 위해 부밴드로 나누고 주파수와 시간 방향의 무게중심을 구하고 정규화하여 인식기에 사용하였다. 무게중심 값은 스펙트럼의 형태적 특징을 잘 나타내면서도 간결하여 인식기에 사용되는 특징 DB의 크기를 줄여줄 수 있는 장점이 있다. 수 천곡 규모의 오디오에 대해서, 부밴드 스펙트럼의 주파수와 시간 방향 무게중심의 인식 성능을 비교하였다. 실험 결과 주파수와 시간 방향 특징을 결합하면 상보적으로 인식 성능을 높일 수 있음을 발견하고, 선형 변환을 이용하여 주파수와 시간 방향 특징을 하나로 결합하는 방법을 제안하였다.
핵심용어: 오디오 인식, 오디오 핑거프린팅, 부밴드 주파수-시간 무게중심, 지각적 해싱
투고분야: 음향 신호처리 분야 (1,2)

This paper proposes a new audio identification method based on a combination of the instantaneous and dynamic spectral features of the audio spectrum. Especially we propose the spectro-temporal subband centroids that are easy to compute and effective to summarize the instantaneous and dynamic spectral variations. Experimental results demonstrate that the identification performance can be greatly improved by combining both the spectral and the temporal subband centroids.
Keywords: Audio identification, Audio fingerprinting, Subband spectro-temporal centroid, Perceptual hashing
ASK subject classification: Acoustic Signal Processing (1,2)

I. 서 론

컴퓨터, 네트워킹, 저장장치 등의 발달로 대용량 음악 데이터를 빠르고 신뢰성 있게 보호, 관리 및 검색할 수 있는 오디오 인식의 필요성이 높아지고 있다. 지문을 이용해서 사람을 인식하듯이, 오디오 인식 시스템은 오디오의 고유한 특징을 이용하여 해당 오디오를 인식한다. 그림 1과 같이 인식하고자 하는 오디오 파일들에서 특징을 추출하여 특징 DB를 만들고 오디오 파일의 메타 정보와 연동시키게 된다. 인식하고자 하는 미지의 오디오에서 추출된 특징으로 미리 만들어둔 특징 DB를 검색하고 최종 검증 과정을 거쳐 인식하게 된다. 오디오 인식기는 P2P/UCC 등을 통한 불법 파일 공유를 막는 필터링, 방송 모니터링, 무선망을 통한 음악 찾기, 대용량 오디오 라이브러리를 자동으로 태깅 (tagging) 또는 인덱싱 (indexing) 하는 등 여러 가지 실제적인 용도를 가지고 있어 최근 많은 관심을 받고 있으며 다양한 기법들이 제안되고 있다 [1-3].

기존의 데이터 색인 기법인 해싱 (hashing)에 비유하여, 오디오 인식을 핑거프린팅 (fingerprinting) 또는 지각적 해싱 (perceptual hashing)이라고 부르기도 한다. 기존의 암호적인 해싱 기법의 경우 데이터가 조금만 변화해도 해쉬 (hash) 값이 크게 변화하므로, 오디오 데이터와 같이 압축, 잡음처리, AD/DA 변환 등의 다양한 신호처리 과정에 대한 강인성이 요구되는 경우에는 적합하

책임저자: 서 진 수 (jsseo@kangnung.ac.kr)
210-702 강원도 강릉시 강릉대학로 120 강릉대학교 전자공학과
(전화: 033-640-2428; 팩스: 033-646-0740)

지 않다. 일반적으로 오디오 인식기에 사용되는 특징은 간결하면서도 다양한 변환에 대한 강인성을 가지고, 서로 다른 오디오에 대해 차별성을 줄 수 있어야 한다 [1]. 따라서 오디오 인식기 설계에 있어서 간결성, 강인성, 차별성을 두루 갖추고 있는 특징을 찾는 것이 중요하다.

인간의 음향 지각 및 기계적인 음성 인식 시스템 연구에서 현시점의 (instantaneous) 주파수 특성 뿐만 아니라 그 특성들의 시간적 (dynamic) 변화도 중요함이 널리 알려져 있지만 [4, 5], 이를 오디오 인식에 적용한 예는 거의 없었다. 따라서 본 연구에서는 현시점의 주파수 특성 뿐만 아니라 그 특성의 시간적 변화도 고려해 줄 수 있는 부밴드 주파수-시간 무게중심을 이용하여 오디오 인식기를 구성하고 성능을 평가하였다. 무게중심 값은 그 형태가 선형회귀 분석에서 기울기와 연관 되어있어서 [4], 부밴드 주파수-시간 무게중심은 스펙트럼의 주파수 및 시간 방향의 형태적 특성을 간결하게 요약해서 나타내면서도, 오디오가 변형을 겪더라도 그 값이 크게 변하지 않아서 강인성을 유지할 수 있다. 주파수 방향 부밴드 무게중심의 경우 이미 오디오 인식 [3]과 음성 인식 [6]에 적용되어 우수한 성능을 보였다. 본 논문에서는 기존의 주파수 방향 특성만을 고려하던 오디오 인식시스템 [3]에 시간 방향 특성을 추가하여 인식률을 높일 수 있음을 보였다. 또한 선형변환을 이용하여 시간 방향 특성을 추가하여 늘어난 특징의 차수를 줄이는 방안에 대해서도 연구를 수행하였다.

본 논문에서는 실험을 통해 오디오 부밴드 스펙트럼의 시간 방향 특성을 포함하는 것이 실제로 오디오 인식 성능을 크게 향상 시킬 수 있음을 확인했고, 선형 변환을 이용하여 성능을 크게 저하시키지 않고 인식에 사용되는 특징의 차수를 줄일 수 있음을 확인하였다. II장에서 제안된 부밴드 주파수-시간 무게중심을 이용한 오디오 인식 방법에 대해 살펴보고, III장에서 제안된 방법의 성능을 실험하고 결과를 분석한다.

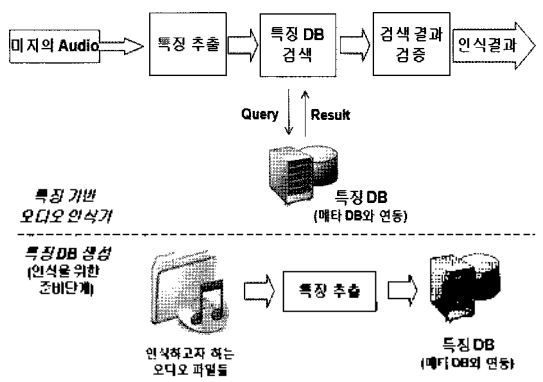

그림 1. 특징 기반 오디오 인식 과정 블록선도
Fig. 1. Block diagram of feature-based audio identification.

II. 주파수-시간 무게중심을 이용한 오디오 인식

2.1. 부밴드 주파수-시간 무게중심

오디오 인식을 위해 본 연구에서 고려한 특징은 부밴드 주파수-시간 무게중심 (spectro-temporal subband centroid; STSC)이며 그림 2와 같은 블록선도로 구성된다. 먼저 입력 오디오 신호를 모노로 바꾸고 11025 Hz로 샘플링 주파수를 맞춘 후, 4096 길이의 해닝 (Hanning) 윈도우를 50%씩 겹쳐 가면서 (overlap) 적용하고 Fourier 변환을 가한다. 이렇게 주파수 도메인으로 신호를 변환해서 얻은 각 오디오 프레임의 파워 스펙트럼 (power spectrum)을 M개의 부밴드로 나눈다. 특히 본 논문에서는 300 Hz에서 5300 Hz 사이의 16개의 인간 청각의 임계 대역 (critical band) [7]을 부밴드로 사용하였다. 입력 오디오 신호 n번째 프레임의 스펙트럼의 k번째 주파수 계수를 $S[n, k]$ 라고 하면, 주파수 방향의 무게중심 Cs와 시간 방향 무게중심 Ct는 다음과 같이 주어진다.

$$Cs[n,m]=\frac{\sum_{k=B[m]+1}^{B[m+1]} k\,S[n,k]}{\sum_{k=B[m]+1}^{B[m+1]} S[n,k]}$$
$$Ct[n,m]=\frac{\sum_{t=n-W+1}^{n} t\,\bar{S}[t,m]}{\sum_{t=n-W+1}^{n} \bar{S}[t,m]} \quad (1)$$

여기서 $B[m]$은 m번째 부밴드의 주파수 경계값을 나타내고, W는 시간 방향 무게중심을 구하기 위해 사용되는 프레임의 수, $\bar{S}[t,m]$는 t번째 프레임의 m번째 부밴드내의 파워 스펙트럼 값들을 합한 것이다.

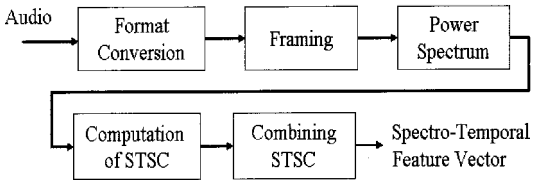

그림 2. 부밴드 주파수-시간 무게중심 (STSC)을 이용한 오디오 인식 특징 추출 블록선도
Fig. 2. Block diagram of the spectro-temporal feature extraction from STSC for audio identification.

얻어진 두 가지 종류의 무게중심은 각기 주파수와 시간 방향 오디오의 특성을 나타내는 것으로 서로 상호보완적으로 인식에 도움이 될 수 있다. 따라서 두 특징을 모두 사용하는 것이 가장 효율적이지만, 인식기에서 특징의 차수가 높아지면 특징 DB 검색 및 검색 결과 검증 과정에서 많은 부하가 걸리게 된다 [8]. 본 논문에서 부밴드의 개수가 M개이므로 주파수와 시간 방향 무게중심인 M차원의 Cs와 Ct를 모으면 $2M$ 차수의 벡터 C가 얻어진다. 이 경우 [3]에서처럼 한가지 형태의 무게중심만을 이용하는 것에 비해서 저장공간이 2배가 되고, 특징 검색에 드는 계산량은 비선형적으로 더 크게 증가하게 된다 [8]. 일반적으로 이러한 문제를 해결하기 위해서 특징의 차수를 줄이는 (dimensionality reduction) 다양한 기법들이 제안되어 왔다. 특히 선형변환을 이용하여 특징의 차수를 줄이는 방법들이 널리 쓰이고 있으며, Cosine, Hadamard, Haar 등 신호에 독립적인 변환과 KLT (Karhunen– Loève transform) 등 신호에 종속적인 변환으로 크게 나눌 수 있다 [9]. 오디오 인식에서는 다양한 신호적 특성을 가지는 오디오 데이터를 모두 다룰 수 있어야 하고, 실시간 동작을 위해 계산량이 중요하다. 따라서 본 논문에서는 신호 독립적인 Cosine, Hadamard, Haar 변환을 고려하였다. 또한 이들 선형 변환은 변환 전 신호에 대해서 하한성질 (lower bounding)을 가지고 있어서, 변환된 특징을 인식기에 사용하더라도 특징 DB 인덱싱에서의 오인식을 방지할 수 있는 장점이 있다 [10]. 본 논문에서 사용된 변환의 파라미터는 [9]에 나온 것과 같다. 주파수와 시간 방향 무게중심 특징을 모아서 얻은 $2M$ 차수의 무게중심 특징 C에 대해 선형변환 행렬 L을 가하고 강인성을 위해 M개의 저주파 (low-frequency) 계수를 취하면 다음 식과 같이 M차의 특징 벡터 F를 얻을 수 있다.

$$F_{M\times 1} = L_{M\times 2M}\, C_{2M\times 1} \tag{2}$$

최종적으로 다양한 변환을 이용하여 얻어진 특징 F에 대해 III장에서 인식 성능을 비교하였다. 선형변환 외 다른 방법으로 단순히 Cs와 Ct를 합하거나 큰 값 또는 작은 값 등을 취하는 방법들이 있으나, 선형변환이 더 우수한 인식 성능을 보였다.

2.2. 특징 비교

본 연구의 주목적은 부밴드 주파수-시간 무게중심 특징의 오디오 인식 성능을 검증하는 것이다. 특징 비교 방법은 성능 비교를 용이하게 하기위해서 기존의 논문 [3]과 동일하게 특징을 정규화하고 Euclidean 거리를 사용하였다. 오디오 인식 문제는 오디오 특징 추출 함수인 H와 특징간 거리 비교 함수인 D를 이용하여 아래와 같은 가설검증 (hypothesis testing)으로 주어진다.

가설 L_0: 만약 D(H(A), H(B)) 이 문턱값 T보다 작다면, 두 오디오 클립 A와 B는 같은 오디오이다.

가설 L_1: 만약 D(H(A), H(B)) 이 문턱값 T보다 크다면, 두 오디오 클립 A와 B는 다른 오디오이다.

각각의 오디오 프레임에서는 2.1절에서 기술된 바와 같이 M차원의 특징이 추출된다. 실제로 한 프레임에서 추출된 M차원의 특징만으로는 다양한 오디오 변형들에 대해서 강인성을 유지하면서 위 가설검증을 엄밀하게 적용하는 것이 불가능하다. 따라서 인접한 N개의 프레임에서 나온 특징들을 모아서 MxN 형태의 행렬인 특징 블록 c [n, m]을 만들고 이를 오디오 인식에 이용하게 된다 [2, 3] ($1 \le n \le N$, $1 \le m \le M$). 먼저 특징 블록을 아래와 같이 그 블록내 m번째 부밴드의 특징들의 평균 $\mu_c[m]$과 표준편차 $\sigma_c[m]$값으로 정규화 한다.

$$p[n,m] = \frac{c[n,m] - \mu_c[m]}{\sigma_c[m]} \tag{3}$$

위 정규화 과정 후 특징 비교를 위해 Euclidean 거리를 아래와 같이 두 오디오 특징 블록 p와 q의 거리 비교에 사용하였다.

$$D = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}(p[n,m]-q[n,m])^2 \tag{4}$$

본 논문에서는 특징 추출 방법으로 부밴드 주파수-시간 무게중심을 사용하고, 식 (3)의 정규화와 식 (4)의 방법으로 특징 비교를 수행하여 얻은 두 오디오 간의 거리 D를 미리 정해둔 문턱값인 T와 비교하여 오디오 인식을 수행하게 된다.

III. 실험 결과

본 장에서는 부밴드 주파수 및 시간 방향 무게중심의 오디오 인식 성능을 비교하고, 선형변환을 이용하여 특

징의 차수를 줄이는 방법의 인식성능에 대해 실험하였다. 실험을 위해서 8000곡 분량의 다양한 장르의 음악 파일을 수집하였다. 수집된 음악 파일들에서 부밴드 스펙트럼의 주파수-시간 무게중심 특징을 추출하여 특징 DB를 만들었다. 시간 방향 무게중심 값은 3개의 인접한 프레임에서 구했고, 특징 비교에는 5초 길이의 음악에서 얻어진 특징들을 사용했다 (M = 16, W = 3, N = 27). 일반적으로 특징 간의 인식 성능 비교에는 ROC (receiver operating characteristic) 곡선이 이용된다 [11]. ROC 곡선은 인식 시스템에 존재하는 두 가지 형태의 오인식율인 FAR (false alarm rate)과 FRR (false rejection rate)을 가로와 세로축으로 하여 그래프를 그린 것이다. 오디오 인식 시스템에서 FAR은 서로 다른 오디오를 같다고 판정할 확률이며, FRR은 같은 오디오를 다르다고 판정할 확률이다. 본 장에서는 실험을 위해서 수집된 8000곡의 음악으로부터 얻은 각 특징 또는 특징의 선형 변환 값들로 특징 DB를 만들고, 다양한 변형들에 대한 인식 실험을 통해 ROC 곡선을 구하여 특징들의 성능을 비교하였다.

FAR을 구하기 위해서 구축된 특징 DB에서 임의로 선택된 특징 쌍들 간의 거리를 구하고, 인식기의 문턱값 T를 변화시켜가면서 문턱값보다 작은 거리를 가지는 특징 쌍의 비율을 구하였다. FRR을 구하기 위해서, 각 음악 파일에 다양한 종류의 변형을 가하였다 (Cool Edit Pro 2.1 소프트웨어 사용). 실험1은 각 음악 파일에 filter emulating old time radio, 1% 피치 증가, 92.9 ms shift, 30-band pop EQ 의 변형을 가하고 MP3 (128 kbps) 압축을 가한 후 특징을 추출하였다. 변형된 음악에서 추출된 특징과 변형전 음악의 특징 사이의 거리를 인식기의 문턱값 T와 비교하여 FRR을 구하였다. 실험1에 대한 무게중심 및 선형변환 특징의 ROC 곡선은 그림 3(a)와 같다. 실험2는 각 음악 파일에 filter emulating ambient metal room, 1% 피치 감소, 92.9 ms shift, 30-band pop EQ 의 변형을 가하고 MP3 (128 kbps) 압축을 가한 후 실험1과 같은 과정으로 특징 비교를 수행하였고, 그 결과 ROC 곡선은 그림 3(b)와 같다. ROC 곡선 결과를 보면 주파수 방향 부밴드 무게중심 특징이 시간 방향 부밴드 무게중심보다 성능이 우수함을 알 수 있으며, 두 특징을 모두 사용할 경우 차수가 2배가 되지만 오류율이 크게 줄어듦을 알 수 있다. 두 특징을 모아서 하나의 특징 벡터로 만들고 선형변환을 적용하여 16차로 줄인 경우를 보면 Haar 변환이 가장 좋은 성능을 보였고, 그 외 Cosine과 Hadamard 변환을 사용해도 시간 또는 주파수 방향 특징만을 사용하는 것에 비해서 인식 성능을 높일 수 있음을 알 수 있다.

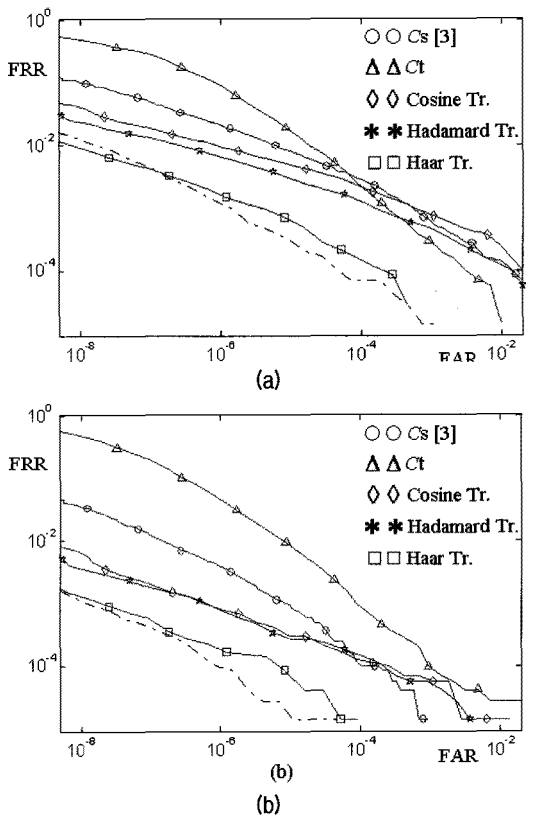

그림 3. 부밴드 스펙트럼 주파수-시간 무게중심 특징에 대한 ROC 곡선: M = 16, W = 3, N = 27, — 프레임당 16차 특징, - - 프레임당 32차 특징 (Cs와 Ct 모두 사용).
(a) 오디오 왜곡 실험 1 결과
(b) 오디오 왜곡 실험 2 결과

Fig. 3. ROC curves for the STSC feature vectors: M = 16, W = 3, N = 27, — 16 dimensions per frame, - - 32 dimensions per frame (Concatenation of Cs and Ct).
(a) Result of Distortion Test 1
(b) Result of Distortion Test 2

작은 차수의 특징은 특징 DB 저장 공간을 줄일 수 있고 검색 속도도 높일 수 있는 장점이 있다 [8, 11].

IV. 결 론

본 논문에서는 오디오의 순간 주파수 특성에 시변 주파수 특성을 같이 사용하는 것이 오디오 인식기의 성능을 높일 수 있음을 보였다. 또한 선형변환의 저주파 성분만을 취하여 늘어난 특징의 차수를 줄일 수 있는 방법을 제안하고 다양한 선형변환 방법들에 대해서 인식 성능을 비교하였다. 본 연구 결과는 부밴드 무게중심을 이용한 오디오 인식에 대해 다루었으나, 오디오를 다루는 다른 문제들인 내용기반검색 (retrieval)과 장르 분류 등의 다양한 문제에도 확장 적용하면 순간 주파수 특성에 시변

주파수 특성을 더하여 성능을 높이는 것이 가능할 것으로 기대된다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 IT 신성장동력 핵심기술개발사업의 일환으로 수행하였음. [2007-S017-01, 사용자 중심의 콘텐츠 보호 · 유통 기술]

참 고 문 헌

1. P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting", in Proc. IEEE Workshop on Multimedia Signal Processing, 169-173, Dec. 2002.
2. J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in Proc. International Conf. on Music Information Retrieval, 2002.
3. Jin S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C.D. Yoo, "Audio fingerprinting based on normalized spectral subband moments," IEEE Signal Processing Letters, 13(4), 209-212, 2006.
4. S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Transactions on Acoustics, Speech, and Signal Processing, 34(1), 52-59, 1986.
5. 김기석, 임은진, 황희용, "음성 인식 신경망을 위한 음성 파라미터들의 성능 비교," 한국음향학회지, 11(3), 61-66, 1992.
6. K. Paliwal, "Spectral subband centroid features for speech recognition," in Proc. IEEE ICASSP, 1998, 617-620.
7. E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, (Springer-Verlag, 1999).
8. C. Bohm, S. Berchtold, and D. Keim, "Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases," ACM Computing Surveys, 33(3), 322-373, 2001.
9. A.K. Jain, *Fundamentals of Digital Image Processing*, (Prentice-Hall, Upper Saddle River, NJ, 1989).
10. J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A Novel Symbolic Representation of Time Series", Data Mining and Knowledge Discovery, 15(2), 107-144, 2007.
11. A.K. Jain, Robert P.W. Duin, and J. Mao, "Statistical pattern recognition: A review", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 4-37, 2000.

저자 약력

•서 진 수 (Jin Soo Seo)

1976년 4월 12일생
1998년 2월: 한국과학기술원 전기 및 전자공학과 (공학사)
2000년 2월: 한국과학기술원 전자전산학과 (공학석사)
2005년 2월: 한국과학기술원 전자전산학과 (공학박사)
2005년 3월~2006년 2월: 한국과학기술원 정보전자연구소 BK21 연구원
2006년 3월~2008년 2월: 한국전자통신연구원 디지털콘텐츠 연구단 선임연구원
2008년 3월~현재: 강릉대학교 전자공학과 조교수

•이 승 재 (Seungjae Lee)

1977년 10월 28일생
2003년 2월: 연세대학교 전자공학과 (공학사)
2005년 2월: 한국과학기술원 전자전산학과 (공학석사)
2005년 2월~현재: 한국전자통신연구원 SW 콘텐츠 연구부문 선임연구원