한국음향학회지 제25권 제4호 pp. 151~158 (2006)

다중 레벨 양자화 기법을 적용한 오디오 핑거프린트 추출 방법

Audio Fingerprint Extraction Method Using Multi-Level Quantization Scheme

송 원 식*, 박 만 수*, 김 회 린*
(Wonsik Song*, Mansoo Park*, Hoirin Kim*)

*한국정보통신대학교

(접수일자: 2006년 2월 9일; 수정일자: 2006년 4월 20일; 채택일자: 2006년 5월 2일)

본 논문은 필립스의 음악 검색 기법을 기반으로 필터 뱅크 에너지 변화량과 음악의 통계적인 특성을 이용한 오디오 핑거프린트 추출 방법을 제안하였다. 기존의 필립스 방식은 제한된 주파수 영역을 너무 많은 필터 뱅크로 분할하여 분석함으로써 밴드들 사이에 연계성 및 왜곡에 대한 민감도가 증가하는 특징을 보일 수 있다. 제안된 방법은 필터 뱅크의 밴드 수를 줄여 왜곡에 대한 강인성을 증진시키고, 필터 뱅크 에너지의 변화량의 부호와 크기 정보를 통계적 특성을 고려한 양자화 기법을 이용해 2비트로 할당함으로써 오디오 핑거프린트의 고유성을 확보하였다. 추출된 2비트는 4개의 레벨로 정보를 표현함으로 각 레벨 사이에 연계성이 존재하게 된다. 이 같은 레벨 사이의 연계성은 유사도 측정 시 이용될 뿐만 아니라 오디오 핑거프린트를 기준으로 검색 영역을 확장하는 제안된 방식에서는 효율적인 검색 영역을 선택할 수 있는 정보로 활용 되었다. 제안된 방식은 다양한 주변 잡음 환경 (거리, 백화점, 자동차, 사무실, 식당)에서의 실험을 통하여 주변 잡음에 강인한 특성을 보일 뿐만 아니라 검색 속도 또한 향상되는 특징을 보였다.

핵심용어: 오디오 핑거프린트, 확률 분포, 양자화, 필터뱅크 에너지 변화량

투고분야: 음성처리 분야 (2)

In this paper, we proposed a new audio fingerprint extraction method, based on Philips' music retrieval algorithm, which uses the energy difference of neighboring filter-bank and probabilistic characteristics of music. Since Philips method uses too many filter-banks in limited frequency band, it may cause audio fingerprints to be highly sensitive to additive noises and to have too high correlation between neighboring bands. The proposed method improves robustness to noises by reducing the number of filter-banks while it maintains the discriminative power by representing the energy difference of bands with 2 bits where the quantization levels are determined by probabilistic characteristics. The correlation which exists among 4 different levels in 2 bits is not only utilized in similarity measurement, but also in efficient reduction of searching area. Experiments show that the proposed method is not only more robust to various environmental noises (street, department, car, office, and restaurant), but also takes less time for database search than Philips in the case where music is highly degraded.

Keywords: Audio Fingerprint, Probabilistic Characteristics, Quantization, Energy Difference of Neighboring Filter-Banks

ASK subject classification: Speech Signal Processing (2)

I. 서 론

전통적인 내용기반 음악 검색 시스템에 관한 연구는 피치나 스펙트럼 포락선의 히스토그램 기반의 확률적 패턴을 모델링 하는 형태로 이루어졌다. 이 같이 구축된 모델들은 벡터 양자화 기법[1, 4]을 이용하여 추출된 오디오 특징 벡터를 군집화한 것이다. 그러나 이 같은 형태는 음악 검색 시스템의 확장성과 성능을 보장하지 못하기 때문에 상업적 목적으로 적합하지 못하다. 이 같은

책임저자: 송 원 식 (songcode78@icu.ac.kr)
305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교 공학부 음성언식기술연구실
(전화: 042-866-6221; 팩스: 042-866-6245)

이유에서 최근 연구 방향은 확장성과 성능을 보장할 수 있는 대용량 데이터 베이스 기반의 음악 검색기 개발에 초점이 맞춰져 있다.

최근 유·무선 통신의 발달과 함께 이를 기반으로 한 음악 검색 기술은 음악 서비스 업체들에게 매력적인 어플리케이션으로 각광 받고 있다. 예로 소비자의 요청에 의한 음악 검색, 음악 방송 모니터링, peer to peer network[5, 7]상에서 인증되지 않은 음악 파일의 공유 차단 등의 서비스가 제공되고 있다.

필립스의 오디오 핑거프린트 기법은 최근에 발표된 내용 기반의 음악 검색 기술로써[8-9] 대용량 데이터 베이스 검색에 적합한 특성을 보이고 있다. 그러나 이 방법도 실제 환경에서 발생되는 주변 잡음이나 음악 재생 속도 변화 등의 변위에 음악 검색기의 성능이 현저히 저하되는 특징을 보인다. 특히 실제 녹음 환경에서 주변 잡음에 의해 음악에 왜곡이 발생할 경우 추출된 오디오 핑거프린트를 기준으로 검색 영역을 제한하는 시스템에서 검색 성능 저하뿐만 아니라 많은 검색 시간을 요구하는 특징을 보인다. 이 같이 주변 잡음이 쿼리 음악에 포함되는 경우는 일상 생활에서 빈번하게 발생 할 수 있기 때문에 왜곡에 강인하면서 빠른 검색 시간을 갖는 음악 검색기 설계는 실제 어플리케이션에서 매우 중요한 요소이다.

본 논문에서는 필립스의 오디오 핑거프린트 추출방법[8-9]을 기반으로 필터 뱅크 에너지 변화량을 확률 통계적 특성을 고려하여 양자화 시킴으로써 좀더 효율적으로 음악을 표현 할 수 있는 오디오 핑거프린트 추출 기법을 제안하였다. 논문은 다음과 같은 순서로 기술되었다. 2장에서는 기존의 오디오 핑거프린트 추출 방법에 대하여 살펴보고, 3장에서는 제안된 다중 레벨 양자화 기법을 이용한 오디오 핑거프린트 추출 기법 및 검색 영역 확장 방법에 대해 기술하였다. 4장과 5장에서는 실험 결과 분석 및 결론에 대하여 언급하였다.

II. 기존의 오디오 핑거프린트 추출 방법

필립스에서 제안한 오디오 핑거프린트[8-9] 추출 방식은 음악의 특성을 시간에 따른 주파수 밴드 에너지의 증감에 의해 표현한다. 그림 1은 전체 오디오 핑거프린트가 추출되는 과정을 보여준다. 입력 신호는 일정한 사이즈의 프레임으로 분리된 후 FFT (Fast Fourier Transform)을 이용해 주파수 영역에서 분석된다. 그리고 주파수 영역의 에너지는 33개의 필터 뱅크 신호로 묶이게 된다. 여기서 33개 밴드는 300~3000Hz 사이의 주파수 대역으로 mel/bark 스케일을 갖는다. 33개의 필터 뱅크에서 추출된 에너지 값은 시간과 주파수 영역에서 필터 링 시킨 후, 그 값의 부호를 기준으로 32비트의 오디오 핑거프린트로 변환된다.

$$ED(n,m) = E(n,m) - E(n,m+1) - (E(n-1,m) - E(n-1,m+1)) \quad (1)$$

그림 1의 전체 필터 링 과정은 식 (1)로 정의된다. 식 (1)에서 E (n,m)는 n차 프레임의 m차 필터 뱅크 밴드[8-9]의 에너지를 나타내고 ED (n,m)은 필터 뱅크의 에너지 변화량을 나타낸다. 추출된 필터 뱅크의 에너지 변화량은 식 (2)에 의해 부호에 따라 1개의 비트가 할당되고, 추출된 32개 비트의 조합을 통해 sub-fingerprint 혹은 해쉬 값으로 표현된다.

$$H(n,m) = \begin{cases} 1 & if \quad ED(n,m) \geq 0 \\ 0 & if \quad ED(n,m) < 0 \end{cases} \quad (2)$$

필립스 방식[8-9]은 프레임 마다 핑거프린트를 추출하기 때문에 brute-force 검색 방식은 비 효율적이다. 따라서 그림 2와 같이 쿼리 데이터에서 추출된 오디오 핑거프린트 값을 기반으로 검색 범위를 제한하여 검색을 수행한다. 제한된 도메인의 경우 메모리의 효율적인 활

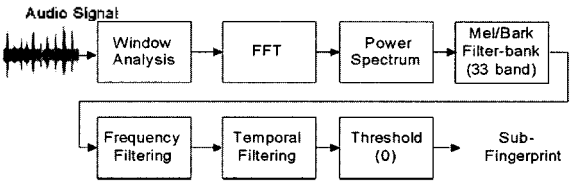

그림 1. 필립스의 오디오 핑거프린트 추출 과정
Figure 1. Overview of Philips' audio fingerprint extraction process

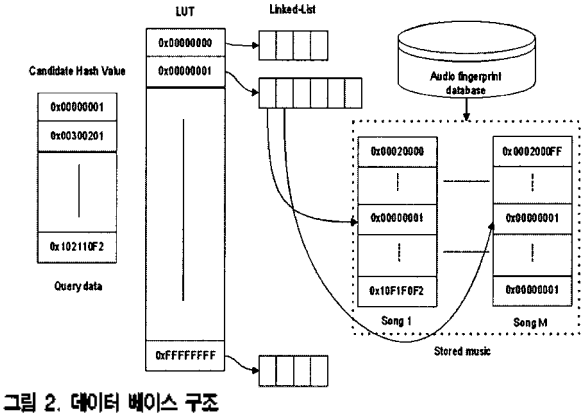

그림 2. 데이터 베이스 구조
Figure 2. Database Layout structure.

용을 위해 look-up table대신에 hash table을 이용하여 색인 목록을 생성하게 된다. 왜곡에 의해 해당 곡의 해당 구간의 데이터 베이스에 접근하지 못하는 경우가 발생 하기 때문에 일반적으로 쿼리 데이터의 오디오 핑거프린트를 기반으로 일정한 Hamming Distance의 값까지 검색 대상을 확장하게 된다. 또한 필립스 방식[8-9]은 유사도 측정 방법으로 해쉬 값들의 Hamming Distance를 사용한다. 즉 각 비트의 일치 성을 판단하는 것이다. 최종 검색 결과는 사전에 결정된 오디오 핑거프린트의 블럭 당 BER (Bit Error Rate)를 임계치로 이용해 검색 결과를 검증하게 된다.

III. 다중 레벨 양자화 기법을 적용한 오디오 핑거프린트 추출 방법

필립스 방식[8]에서 오디오 핑거프린트 추출을 위해 사용되는 필터뱅크 에너지 정보는 사람의 청각특성을 반영한 특징 벡터로써 일반적으로 300~3000Hz사이의 주파수 영역에서는 16~20사이의 필터뱅크 개수를 사용한다. 그러나 필립스 방식은 오디오 핑거프리트를 추출하기 위하여 33개의 필터뱅크 개수를 사용하였다. 이것은 주파수 영역을 너무 조밀이 분석하여 왜곡에 대한 민감도와 필터 뱅크 사이에 연계성을 증대시킬수 있다. 또한 필터뱅크 에너지의 변화량은 음악의 고유한 특징을 표현할 수 있는 정보임에도 불구하고 필립스 방식에서는 부호의 정보만을 취하고 크기 정보는 사용하지 않았다.

제안된 방법은 위의 특징들을 이용하여 필터뱅크의 밴드수를 17개로 줄임으로써 주변 잡음에 대한 민감도와 필터뱅크 밴드들 간의 연계성을 줄이고, 필터뱅크 에너지의 변화량의 부호와 크기 정보를 2비트로 매핑하여 오디오 핑거프린트의 고유성을 증대시켰다. 이 같이 16개의 필터뱅크 에너지 변화량을 각각 2 비트로 매핑함으로써 필립스 방식과 동일한 메모리 사이즈의 오디오 핑거프린트를 추출하였다. 제안된 방식은 필터뱅크 에너지 변화량의 부호정보 뿐만 아니라 크기 정보를 추가 함으로써 음악의 멜로디적 요소를 더 정확하게 표현하였다. 즉 추출된 오디오 핑거프린트에 필터뱅크의 에너지 변화량의 포락선 정보가 좀더 정확히 표현됨으로써 음악의 고유한 특징을 반영하게 되는 것이다.

제안된 방식은 데이터 베이스를 바탕으로 추출된 임계치 기준으로 오디오 핑거프린트를 추출하게 된다.

3.1. 오디오 핑거프린트 추출

임계치 추출 과정은 오디오 핑거프린트를 추출하기 위한 사전 과정으로 그림 3처럼 필터 뱅크의 밴드 수를 17개로 줄인 것을 제외하고는 필립스 방식과 동일하게 필터 뱅크의 에너지 변화량을 추출한다. 추출된 값은 상대적인 필터뱅크의 에너지 변화량 값으로 변환되기 위해 프레임 정규화 과정을 거치게 된다. 정규화 과정은 프레임의 절대 값의 평균으로 각 밴드의 에너지 변화량을 나누어 주는 것이다. 정규화된 각 밴드의 에너지는 프레임 안에서의 밴드들 간의 상대적인 에너지 크기를 나타내고 이는 프레임 안에서 상대적인 밴드 에너지 집중도를 나타내게 된다. 식 (3)은 프레임 정규화 과정을 식으로 표현하고, $ED_N(n,m)$는 정규화 되어진 값, BN (n) 은 한 프레임의 필터 뱅크 에너지 변화량의 절대 값의 평균을 의미한다.

$$ED_N(n,m)=\frac{ED(n,m)}{BN(n)},\ m=1...16$$
$$BN(n)=\frac{\sum_{m=1}^{16}|ED(n,m)|}{16} \quad (3)$$

데이터 베이스에서 추출된 각 필터 뱅크의 정규화된 에너지를 이용해서 각 밴드의 pmf (probability mass function)을 구하게 된다. pmf를 추출하는 이유는 비트 할당을 위한 임계치를 구하기 위해서이다. 즉 각 밴드의 에너지 분포도를 조사하여 가장 효율적으로 값을 매핑할 수 있는 임계치를 찾는 것이다. 식 (4)는 pmf 추출 방

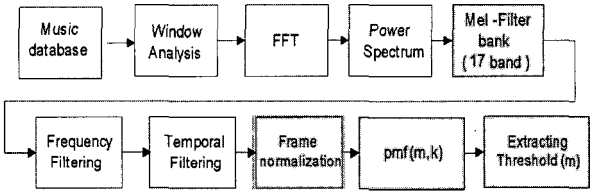

그림 3. 양자화를 위한 임계치 추출과정
Figure 3. Thresholds extraction process for quantization.

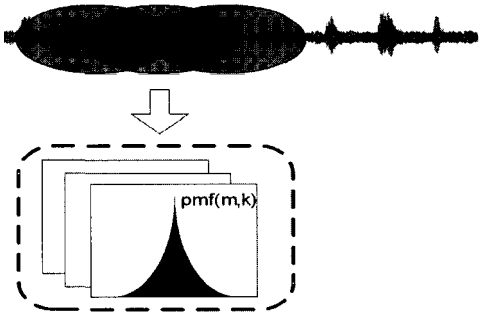

그림 4. 필터 뱅크 에너지 변화양의 pmf 추출 과정
Figure 4. Probability mass function (pmf) extraction process of the energy different of neighboring filter-banks.

법을 식으로 표현한 것이다. 식 (4)에서 pmf (m,k) 는 정규화된 m^{th} 필터 뱅크 에너지 변화량이 k구간 $(\varepsilon(k)_1 \sim \varepsilon(k)_4)$ 사이의 크기를 가질 확률을 나타낸다. 본 논문에서는 N = 500, $\varepsilon(k)_1 \sim \varepsilon(k)_4$ 구간 사이의 간격은 - 16~16을 1000등분하는 범위로 설정하였다. 각 밴드로부터 추출된 확률분포는 그림 4처럼 일반적으로 평균 값은 0를 갖고 라플라시안 모양을 갖는 pmf가 구해진다.

$$
\begin{aligned}
&pmf(m,k) = \frac{\Delta(m,k)}{nframe} \\
&\Delta(m,k) = \#\ of\ \varepsilon(k)_1 < ED_N(n,m) < \varepsilon(k)_4 \\
&nframe = \#\ of\ frame\ in\ music\ database \\
&n = 1,,,nframe,\ k = -N,-N+1,...,N
\end{aligned} \tag{4}
$$

오디오 핑거프린트 추출을 위한 임계치는 데이터 베이스의 확산성을 고려하여 각 pmf의 면적을 양자화 레벨의 개수에 따라 정확히 나누는 점으로 선택하였다. 식 (5)는 pmf를 이용하여 각 밴드의 임계치를 추출하는 과정을 나타낸다. 식 (5)에서 $threshold_1(m)$, $threshold_2(m)$, $threshold_3(m)$ 는 각각 m번째 필터 뱅크 에너지 변화량의 첫 번째, 두 번째, 세 번째 임계치를 나타낸다. 특히 모든 pmf는 평균 값이 0인 특징을 가지므로 $threshold_2(m)$의 값은 항상 0이 되는 특징을 갖게 된다. 이 같은 특성은 실제 추출된 오디오 핑거프린트가 부호 값을 반영하는 것을 의미하게 된다. 즉 오디오 핑거프린트가 필터 뱅크 에너지 변화량의 크기 정보와 부호 정보를 모두 포함하게 되는 것이다.

$$
\begin{aligned}
&threshold_1(m) = n_1,\ \sum_{k=-N}^{n_1} pmf(m,k) = \frac{1}{4}\ for\ k = -N,-N+1,...,N \\
&threshold_2(m) = n_2,\ \sum_{k=-N}^{n_2} pmf(m,k) = \frac{1}{2}\ for\ k = -N,-N+1,...,N \\
&threshold_3(m) = n_3,\ \sum_{k=-N}^{n_3} pmf(m,k) = \frac{3}{4}\ for\ k = -N,-N+1,...,N
\end{aligned} \tag{5}
$$

추출된 임계치를 기준으로 정규화된 필터뱅크 에너지의 크기에 따라 오디오 핑거프린트가 추출된다. 그림 5는 이 같은 과정을 보여준다.

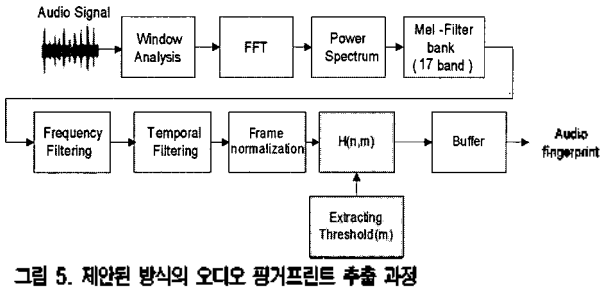

그림 5. 제안된 방식의 오디오 핑거프린트 추출 과정
Figure 5. Audio fingerprint extraction process of the proposed method.

정규화된 에너지가 2비트로 매핑 되는 과정은 식 (6)에 의해 표현된다. 식 (6)에서 볼 수 있듯이 각 밴드의 정규화된 에너지의 크기에 따라 2비트의 오디오 핑거프린트가 할당되는 것을 볼 수 있다.

$$
H(n,m) = \begin{cases} 11 & if \quad ED_n(n,m) \geq threshold_3(m) \\ 10 & if \quad threshold_2(m) \leq ED_n(n,m) < threshold_3(m) \\ 01 & if \quad threshold_1(m) \leq ED_n(n,m) < threshold_2(m) \\ 00 & if \quad ED_n(n,m) < threshold_1(m) \end{cases} \tag{6}
$$

3.2. 비트 사이의 레벨 차이를 이용한 유사도 측정 (Modified Hamming Distance)

제안된 방법에 의하여 추출된 32비트의 오디오 핑거프린트 값은 2비트가 크기 값을 의미하기 때문에 필립스 방식에서 사용된 Hamming Distance대신 ED (Euclidean Distance)와 HD (Hamming Distance)가 결합된 형태의 Modified HD를 사용하였다. Modified HD는 식 (7)처럼 2비트 값들의 ED합을 기준으로 오디오 핑거프린트의 유사도를 측정하게 된다. 오디오 핑거프린트들 사이의 유사도는 BER (Bit Error Rate)대신에 BDM (Bit Dissimilar Measurement)으로 표현된다. 제안된 유사도 측정 방법은 비트의 일치성 보다는 2비트 사이의 유사도 측정에 초점을 맞추었기 때문에 이와 같이 표현 하였다. BDM은 오디오 핑거프린트가 일치할 경우는 0, 완전히 상이할 경우는 1의 값을 갖게 된다.

$$
BDM = \frac{\sum_{n=1}^{nframe} \sum_{m=1}^{16} \left| H_{query}(n,m) - H_{DB}(n,m) \right|}{48 * nframe} \tag{7}
$$

실제 검색 과정은 그림 6처럼 BDM이 일정 임계치 값보다 작아질 때까지 검색하게 된다. 사용된 thr_{break}값은 실험적인 방법을 통하여 얻어졌다.

그림 7은 제안된 방식에 의해 실제 추출된 오디오 핑거프린트 블락과 블락들 사이의 유사도 측정결과의 예를 보여준다. 그림 7a 는 원곡 (가수: 휘성, 곡명: with me), 7b는 5dB 자동차 노이즈에 왜곡된 동일한 곡의 오디오 핑거프린트 블락을 나타낸다. 추출된 오디오 핑거프린트는 크기에 따라 2비트로 필터 뱅크의 에너지 변화량을 매핑함으로 그림 7에서 보듯이 한 개의 픽셀당 4개의 레벨을 가진다. 그림 7c는 7a와 7b의 유사도 측정 결과를 나타낸다. 측정된 결과는 유사도에 따라 4개의 값을 가진다. 그림 7c의 각 픽셀은 두 비트가 비슷할수록

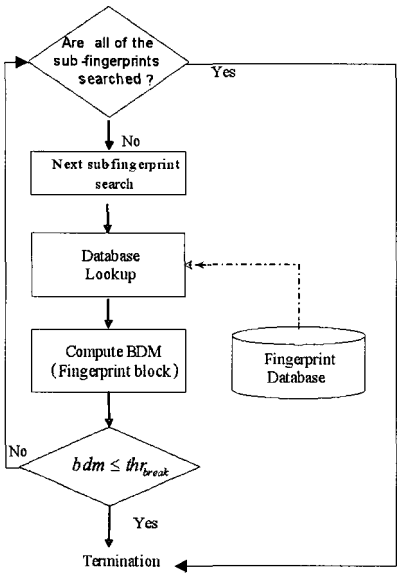

그림 6. 데이터 베이스 검색 과정
Figure 6. Database searching process.

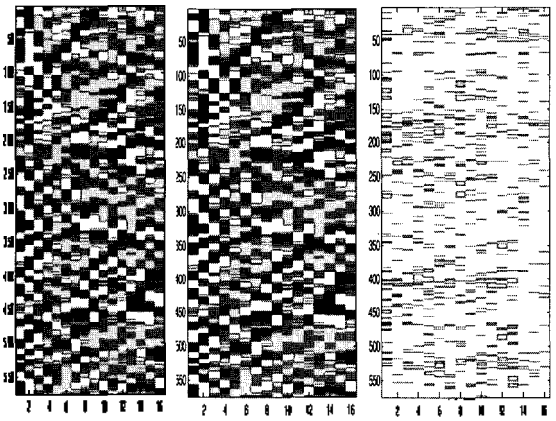

그림 7. 제안된 방식의 오디오 핑거프리트 블락의 추출된 예
(a): 원곡의 오디오 핑거프린트 블락, (b): 왜곡된 원곡의 오디오 핑거프린트 블락, (c): 유사도 측정 결과 (bdm = 0.07)
Figure 7. Example of audio fingerprint block extraction of the proposed method.
(a): Audio fingerprint block of original music. (b): Audio fingerprint block of distorted music. (c): Similarity measure between (a) and (c) (bdm =0.07).

흰색에 가깝게, 차이가 클수록 검은색에 가깝도록 표시하였다.

3.3. 데이터 베이스 검색 후보 확장

쿼리 데이터에서 추출된 오디오 핑거프린트를 데이터베이스 검색 포인트로 사용할 경우 왜곡으로 인한 인식 대상 곡의 해당 구간에 접근하지 못하는 경우가 빈번히 발생 할 수 있다. 그렇다고 모든 경우에 대해서 검색할 경우는 프레임 기반의 오디오 핑거프린트 방식에서는 많은 검색량을 요구하게 된다. 따라서 오디오 핑거프린트의 특성을 고려하여 오디오 핑거프린트 검색 영역을 확장할 경우 아와 같은 단점을 보완 할 수 있다. 필립스 방식의 경우는 HD (Hamming Distance)를 사용하는 반면, 본 연구에서는 2비트로 매핑된 필터 뱅크의 에너지 변화량 사이의 연계성을 이용하여 검색 영역을 확장하게 된다. 즉 오디오 핑거프린트의 레벨 차이를 이용해서 검색 영역을 확장하는 것이다. 제안된 확장 방법은 2비트 단위로 추출된 오디오 핑거프린트 값을 기준으로 가장 유사도가 큰 2개의 값을 찾아 확장하는 방식이다. 본 논문에서는 이 같은 확장 방법을 NSC (Number of Search Candidate)=2 확장으로 표기하였다. 표 (1)은 NSC=2 확장의 예를 보여준다. 제안된 방식은 필립스 HD (Hamming Distance) ≤ 1 확장과 동일하게 32배의 확장된 검색 영역을 갖게 된다.

표 1. NSC = 2 검색 영역 확장 방법
Table 1. NSC = 2 search area expansion method.

H(n,m)	확장 된 비트 쌍	
(00)	(01)	(10)
(01)	(00)	(10)
(10)	(01)	(11)
(11)	(10)	(01)

IV. 실험 결과

4.1. 실험 데이터

실제 녹음 환경을 반영하기 위하여 테스트 데이터는 저렴한 스탠드 마이크를 이용하여 약 10~20cm의 거리로부터 2.1채널 스피커로 출력되는 mp3 파일의 음악으로부터 추출하였다. 모든 테스트 데이터의 길이는 7초로 추출하였고 16비트 양자화 레벨에 샘플링 주파수는 11,025kHz의 standard PCM 포맷으로 변환하여 사용하였다. 데이터 베이스는 대량의 음악 CD를 구입하는 대신에 1,000곡의 mp3 파일을 16bit 양자화 레벨에 샘플링 주파수 11,025kHz의 standard PCM 포맷으로 변환하여 데이터 베이스 구축 시 사용하였다. 구축된 데이터 베이스는 락, 팝, 댄스, 발라드 랩 등의 다양한 장르의 음악으로 구성되었다.

제안된 방식의 성능을 평가하기 위하여 장치 및 주변 잡음에 따라 총 4개의 테스트 데이터 셋을 구축하였다.

- Set I: mp3 파일로부터 직접 추출
  - 100곡으로부터 1,056개의 테스트 데이터 추출

- Set II: 조용한 환경에서 스텐드 마이크와 2.1채널 스피커를 통해 데이터 수집
  - 50곡으로부터 237개의 테스트 데이터 추출
- Set III: 다양한 주변 환경 노이즈와 Set I을 SNR (10, 5, 0dB)에 따라 합성
  - SNR 및 주변 잡음당 1,056개의 테스트 데이터 (총 15,840(5×3×1,056))
- Set IV: 다양한 주변 환경 노이즈와 Set II를 SNR (10, 5, 0dB)에 따라 합성
  - SNR 및 주변 잡음당 237개의 테스트 데이터 (총 3,555(5×3×237))
- 노이즈 데이터: 거리, 백화점, 자동차, 사무실, 식당 등의 5가지 경우의 실제 환경에서 빈번히 발생할 수 있는 상황을 선택하여 MD (Sharp: IM-DR 580H)를 이용하여 녹음

오디오 추출 과정에서 프레임 사이즈는 0.37초, shift 사이즈는 11.6ms로 설정하여 오디오 핑거프린트 값을 추출하였다. 또한 인간의 청각 특성을 반영하기 위해 주파수 밴드 추출 영역을 300~3,000Hz로 제한 하였다.

4.2. 실험 결과

실험은 제안된 방식의 오디오 핑거프린트 추출 방법의 왜곡에 대한 강인성과 검색 속도를 평가하기 위하여 실시되었다. 일반적으로 검색 시간은 CPU 속도나 컴퓨터 내의 현재 작업 중인 프로세스에 따라 많은 영향을 받으므로 절대적인 검색 속도를 평가하는 지표로써 적합하지 않기 때문에 본 논문에서는 평균 색인 목록 (Look Up Table) 검색 횟수를 검색 속도 측정을 위하여 사용하였다. 색인 목록 (LUT) 검색 횟수는 일치하는 음악을 찾기 위해 데이터 베이스에 접근한 횟수를 나타내며, 그림 2의 데이터 베이스 구조에서 색인 목록에 연결된 linked list와 밀접한 관계를 가진다. 즉 특정한 색인 목록에 많은 linked list가 연결될 경우 검색 시 쿼리 데이터와 일치되는 모든 색인 목록에 연결된 linked list를 검색해야 함으로 검색 시간이 오래 걸리는 특징을 갖게 된다.

모든 실험은 필립스 방식의 경우 Hamming Distance (HD ≤ 1)를 이용하여 검색 범위를 제한하였고, 제안된 방식은 오디오 핑거프린트들의 레벨 차이 (NSC = 2)를 이용하여 확장하였다.

표 2는 필립스의 오디오 핑거프린트 추출 방식의 밴드 개수 변화에 따른 주변 잡음에 대한 검색 성능과 평균 색인 목록 검색량을 표시한 것이다. 표2에서 볼 수 있듯이 17개의 필터 뱅크를 사용했을 했을 경우가 월등히 뛰어난 성능을 보이는 것을 볼 수 있다. 이 같은 결과는 필터 뱅크의 밴드 개수를 줄임으로 인해 왜곡에 대한 민감도가 현저히 감소한 것을 보여준다. 그러나 16비트로 오디오 핑거프린트 추출함으로써 오디오 핑거프린트의 고유성이 현저히 저하되는 특징 때문에 색인 목록 검색 양이 급격히 증가하는 특징을 보인다. 이 같은 특징 때문에 오디오 핑거프린트의 효용성을 측정하기 위해서는 검색 성능은 물론 색인 목록 검색량 또한 고려해야 하는 것이 필수적임을 알 수 있다. 따라서 본 연구에서는 제안된 방식의 성능을 측정하기 위하여 검색 성능과 평균 색인 목록 검색량을 함께 비교하였다.

표 2. 필립스 방식의 필터 뱅크 개수에 변화에 따른 성능 평가
Table 2. Performance evaluation with the different number of the filter-banks in Philips method.

성능 평가 지표	필립스 33밴드			필립스 17밴드		
	HD = 0					
	10dB	5dB	0dB	10dB	5dB	0dB
검색 성능 (%)	83.70	63.20	33.30	99.00	92.03	89.05
평균 색인 목록 검색 횟수	14.95	23.49	32.92	2.4×103	1.26×105	2.15×105

표 3. 채널 노이즈에 대한 검색 성능 평가
Table 3. Performance evaluation according to channel noise.

성능 평가 지표	필립스 방식		제안된 방식	
	HD = 0	HD ≤ 1	NSC = 0	NSC = 2
검색 성능(%)	96.20	100	94.09	100
평균 색인 목록 검색 횟수	10.99	20.88	12.97	16.29

표 3은 채널 노이즈에 대한 실험 결과를 보여준다. 즉 필터 뱅크의 에너지 변화량과 채널 왜곡과의 상관관계를 보여준다. 표에서 HD = 0, NSC = 0 는 필립스 방식과 제안된 방식이 검색 영역을 확장하지 않았음을 나타내고, HD ≤ 1, NSC = 2 는 두 방식 모두 검색 영역을 확장 했음을 표시한다. 검색 영역을 확장하지 않은 경우 필립스 방식이 더 좋은 검색 성능을 나타낸다. 이 같은 특징은 필터 뱅크의 밴드의 개수를 줄임으로써 밴드 사이의 채널 특성 차이의 증가로 기인한 것으로 볼 수 있다. 비록 밴드 별 왜곡의 차이가 크지는 않지만 어느 정도 성능에 영향을 미치는 것이다. 그러나 채널 왜곡에 대한 효과는 대부분 제거되기 때문에 검색 영역을 확장함으로써 이 같은 문제점을 해결할 수 있음을 표 3은 보여준다. 또한 평균 색인 목록 검색 횟수도 검색 영역을 확장한 경우 제안된 방식 더 작은 값을 갖는 것을 볼 수 있다.

표 4, 5는 주변 잡음에 대한 검색 성능과 평균 색인 목

표 4. 주변 잡음에 대한 검색 성능 평가
Table 4. Retrieval accuracy according to environmental noises.

잡음 환경	SET III ->검색 성능(%)					
	필립스 방식			제안된 방식		
자동차	HD ≤ 1			NSC = 2		
거리	10dB	5dB	0dB	10dB	5dB	0dB
사무실	100	99.60	98.50	100	99.81	98.67
	98.76	96.78	91.76	99.05	97.34	94.03
백화점	99.43	97.53	92.80	99.43	97.72	96.21
	92.40	82.30	59.20	93.75	84.28	62.12
식당	92.61	80.58	56.72	94.03	83.90	57.67

표 5. 주변 잡음에 대한 평균 색인 목록 검색 횟수
Table 5. Database access trials according to environmental noises.

성능 평가 지표	필립스 방식			제안된 방식		
	HD ≤ 1			NSC = 2		
	10dB	5dB	0dB	10dB	5dB	0dB
평균 색인 목록 검색 횟수	32.22	74.106	157.94	27.55	67.43	151.59

록 검색 횟수를 보여준다. 대부분의 잡음 환경에서 제안된 방식이 검색 성능과 검색 속도 측면에서 뛰어난 것을 볼 수 있다. 특히 주변 잡음 비가 증가할수록 제안된 방식의 성능이 향상되는 것을 볼 수 있다. 검색 성능의 향상은 필터 뱅크의 밴드 수를 줄임으로써 smoothing 효과에 의하여 주파수 영역에서의 왜곡에 대한 민감도 감소와 밴드들 사이에 연계성을 줄임으로써 인접 밴드로 왜곡의 확산을 낮춘 것으로 볼 수 있다. 또한 검색 횟수의 감소는 오디오 핑거프린트 추출 시 최대한 확산성을 고려하여 양자기의 임계치를 추출했기 때문이다. 이 같은 특징은 데이터 베이스에서 오디오 핑거프린트가 일정 영역으로 집중되지 않고 확산되어 분포함을 보여준다.

표 6, 7 은 복합 잡음 즉 채널 노이즈와 주변 잡음을 모두 고려한 경우의 음악 검색 성능과 평균 색인 목록 검색 횟수를 보여준다. 채널 왜곡에 의한 효과 때문에 주변 잡음 레벨이 낮은 10dB의 경우 제안된 방식이 필립스 방식에 비하여 검색 성능 및 검색 속도 측면에서도 성능이 저하되는 특징을 보인다. 그러나 이 같이 잡음 비가 낮은 상황에서 두 방식 모두 높은 인식률을 보이기 때문에 약간의 검색 성능 저하는 큰 문제가 되지 않는다. 이와는 반대로 잡음 비가 증가할수록 제안된 방식의 검색 성능 및 검색 속도 또한 향상되는 것을 볼 수 있다. 이 같은 특징은 제안된 방식이 필립스 방식에 비하여 왜곡에 대한 강인한 특징을 보이는 것을 보여준다.

표 6. 복합 잡음에 대한 검색 성능 평가
Table 6. Retrieval accuracy according to combined noises.

잡음 환경	SET_IV->검색 성능(%)					
	필립스 방식			제안된 방식		
	HD ≤ 1			NSC = 2		
	10dB	5dB	0dB	10dB	5dB	0dB
자동차	99.57	99.57	98.73	99.57	99.57	99.57
거리	99.15	98.73	97.46	98.73	98.31	97.46
사무실	99.15	99.15	96.62	99.15	98.73	97.89
백화점	96.62	83.54	54.85	93.67	85.23	57.8
식당	96.2	79.74	55.69	95.78	81.43	60.33

표 7. 복합 잡음에 대한 평균 색인 목록 검색 횟수
Table 7. Database access trials according to combined noises.

성능 평가 지표	필립스 방식			제안된 방식		
	HD ≤ 1			NSC = 2		
	10dB	5dB	0dB	10dB	5dB	0dB
평균 색인 목록 검색 횟수	39.96	91.69	173.50	43.91	89.83	157.21

V. 결 론

본 논문에서 필터 뱅크 에너지 변화량를 이용한 오디오 핑거프린트 검출 기법을 제안하였다. 제안된 방식은 필터 뱅크 밴드 수를 17개로 줄임으로써 왜곡에 대한 강인성을 증대 시키고, 확률적 분포를 고려하여 추출된 오디오 핑거프린트에 추가함으로써 오디오 핑거프린트의 고유성을 증대시키는 방법을 제안하였다. 또한 2비트로 각 밴드의 에너지 변화량의 정보를 표현함으로써 크기 정보 사이의 연계성을 이용하여 검색 범위를 확장하였다. 이 같은 특징들은 실험을 통하여 주변 잡음에 강인한 특징을 보일 뿐만 아니라 검색 속도의 개선으로 나타났다.

향후 계획으로 음성 코덱 및 다양한 경우의 채널 왜곡에 대한 성능 평가와 필터 뱅크 에너지 변화량 외에 다른 특징 벡터를 이용한 오디오 핑거프린트 추출 방법에 대하여 연구할 계획이다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 디지털 미디어연구소 지원사업의 연구결과로 수행되었음.

참 고 문 헌

1. Mansoo Park et al., "Content-based Music information Retrieval using Pitch Histogram of Band Pass Filter Signal," Proc. of AIRS2004, 245-248, 2004.
2. J. Herre, E. Allamanche, and O. Hellmuth,"Robust matching of audio signals using spectral flatness features," Proc. of Workshop on Applications of Signal Processing to Audio and Acoustics2001, IEEE, 127-130, 2001.
3. E. Allamanche, J. Herre, and O. Hellmuth,"Content-based Identification of Audio Material Using MPEG-7 Low Level Description,"Proc. of ISMIR2001, 197-204, 2001.
4. Jonathan T. Foote, "Content-Based Retrieval of Music and Audio," Proc. of SPIE, Multimedia Storage and Archiving Systems II, 3229, 138-147, 1997.
5. AudibleMagic, http://audiblemagic.com.
6. ShazamEntertainment, http://www.shazam.com.
7. Gracenote, http://www.gracenote.com.
8. Haitsma J., Kalker T. and Oostveen J., "Robust Audio Hashing for Content Identification," Proc. the Content Based Multimedia Indexing2001, 2001.
9. J.A. Haitsma and T. Kalker, "A Highly Robust Audio Fingerprinting System," Proc. ISMIR2002, 144-148, 2002.
10. P.J.O. Doets and R.L. Lagendijk, "Theoretical Modeling Of A Robust Audio Fingerprinting System," Fourth IEEE Benelux Signal Processing Symposium, 2004

저자 약력

• 송 원 식 (Wonsik Song)

2004년 2월: 광운대학교 정보제어학과 (공학사)
2006년 2월: 한국정보통신대학교 (공학석사)

• 박 만 수 (Mansoo Park)

2000년 2월: 인하대학교 (공학사)
2002년 2월: 한국정보통신대학교 (공학석사)
2002년 3월~현재: 한국정보통신대학교 박사과정

• 김 회 린 (Hoirin Kim)

1984년 2월: 한양대학교 전자공학과 (공학사)
1987년 2월: 한국과학기술원 전기공학과 (공학석사)
1992년 2월: 한국과학기술원 전기공학과 (공학박사)
1987~1999년: ETRI 선임 연구원
2000년~현재: 한국정보통신대학교 공학부 교수