한국음향학회지 제32권 제5호 pp. 415~422 (2013)
The Journal of the Acoustical Society of Korea Vol.32, No.5 (2013)
http://dx.doi.org/10.7776/ASK.2013.32.5.415

pISSN : 1225-4428
eISSN : 2287-3775

경첩 손실 함수 최소화를 통한 오디오 핑거프린트 이진화

Audio Fingerprint Binarization by Minimizing Hinge-Loss Function

서진수†

(Jin Soo Seo†)

강릉원주대학교 전자공학과

(접수일자: 2013년 5월 21일; 수정일자: 2013년 7월 22일; 채택일자: 2013년 8월 18일)

초 록: 본 논문에서는 경첩 손실 함수를 최소화를 통해서 강인한 이진 오디오 핑거프린팅 방법을 제안하였다. 특히 제안된 방법에서 오디오 핑거프린트는 이진값을 가지므로 핑거프린트 DB 크기를 줄여줄 수 있는 장점이 있다. 일반적으로 특징을 이진화하는 과정에서 핑거프린트의 강인성, 식별성 등 성능의 손실이 불가피하므로 손실을 최소화하는 것이 필요하다. 본 논문에서는 핑거프린팅에서 두 오디오 클립 간의 유사도가 경첩 함수 형태로 주어지는 것에 착안하여 경첩 손실을 최소화하는 방법으로 특징을 이진화하여 핑거프린트를 구하는 방법을 제안한다. 유도된 경첩 손실 함수는 최소 손실 해싱 기법을 통해서 최소화 하였다. 수 천곡 규모의 오디오에 대해서 다양한 변환들에 대한 인식 성능을 실험하였으며, 제안된 경첩 손실 함수 최소화를 통해서 핑거프린트의 식별성과 강인성이 개선됨을 확인하였다.
핵심용어: 오디오 핑거프린팅, 오디오 인식, 이진 해싱, 특징 학습

ABSTRACT: This paper proposes a robust binary audio fingerprinting method by minimizing hinge-loss function. In the proposed method, the type of fingerprints is binary, which is conducive in reducing the size of fingerprint DB. In general, the binarization of features for fingerprinting deteriorates the performance of fingerprinting system, such as robustness and discriminability. Thus it is necessary to minimize such performance loss. Since the similarity between two audio clips is represented by a hinge-like function, we propose a method to derive a binary fingerprinting by minimizing a hinge-loss function. The derived hinge-loss function is minimized by using the minimal loss hashing. Experiments over thousands of songs demonstrate that the identification performance of binary fingerprinting can be improved by minimizing the proposed hinge loss function.
Keywords: Audio fingerprinting, Audio identification, Binary hashing, Feature learning
PACS numbers: 43.60. Lq

I. 서 론

디지털 신호처리 기술이 발달함에 따라 문서, 음악, 영상, 영화 등 다양한 매체들이 전자기적 장치에 의하여 디지털화 되어 효율적으로 저장, 접근, 이용이 가능하게 되었다. 기존에는 디지털 콘텐츠의 생성, 저장 기술이 주를 이루었던 것에 반해서, 현재는 효율적인 관리 및 유통, 검색 기술의 필요성이 커지고 있다.[1-2] 이와 관련하여 워터마킹 기술이 관심을 받아왔지만, 워터마킹의 경우 콘텐츠에 워터마크를 삽입하므로 원본에 손상을 가해야 하는 단점이 있다. 이러한 문제를 해결하기 위해 콘텐츠 식별(content identification 또는 fingerprinting, media hashing등으로 불리기도 함)[3-6] 기술이 주목을 받고 있다. 콘텐츠 식별은 생체 식별에서 사람의 지문, 홍채 등을 이용하여 그 사람을 인식하는 것처럼 콘텐츠의 특징을 이용하여 해당 콘텐츠를 식별하는 기술을 말한다. 즉, 콘텐츠 식별은 특징 정보를 이용한 인식 시스템으로 콘텐츠들의 특징과 정보 데이터(metadata)를 미리 DB(database)에 저장시켜두고, 식별이 필요한 콘텐츠의 정보를 그 콘텐츠의 특징으로 DB를 검색하여

†**Corresponding author:** Jin Soo Seo (jsseo@gwnu.ac.kr)
Department of Electronic Engineering Gangneung-Wonju National University, 7 Jukhun-gil, Gangneung, Gangwon-Do 210-702, Repubic of Korea
(Tel: 82-33-640-2428, Fax: 82-33-646-0740)

찾아내는 것이다. 콘텐츠 식별 시스템은 P2P/UCC 등을 통한 불법 파일 공유를 막는 필터링, 방송 모니터링, 무선망을 통한 음악 찾기, 대용량 콘텐츠 라이브러리를 자동으로 태깅 또는 인덱싱 등 다양한 용도로 활용 가능하다.

콘텐츠 식별시스템에서 사용되는 특징을 핑거프린트라고 부르며, 일반적으로 오디오 식별에 사용되는 핑거프린트는 다음 4가지 조건들을 만족시켜야 한다.[3,4,7]

- 식별성(collision-free): 서로 다른 음악 간에 오인식이 일어나지 않도록 지각적으로 서로 다른 음악에서 추출된 핑거프린트 값이 충분한 차별성을 가지고 있어야함
- 강인성(invariance to distortions): 오디오 신호가 압축, EQ, 잡음첨가, sampling rate 변화 등 다양한 변환을 겪어 음악 신호에 왜곡이 가해지더라도 핑거프린트 값이 일정한 범위 내에서 유지되어야함
- 간결성(compactness): 다수의 오디오에서 핑거프린트를 추출해서 DB에 저장하므로, 각 오디오 신호에 작은 크기의 표현이 필요함
- 계산용이성(computational efficiency): 핑거프린트 추출에 있어서 계산량과 걸리는 시간이 작아야함

특히 본 논문에서 주목하고자 하는 것은 위의 조건들이 상호 배치된다는 것이다. 콘텐츠의 식별성이 높다는 것은 콘텐츠로부터 얻어진 핑거프린트 값들 간에 차별성이 높다는 것이고, 반면에 강인하다는 것은 차이를 어느 정도 용인한다는 의미로써 둘 사이에 대치되는 관계가 있다. 이렇게 핑거프린트가 가져야할 조건들 간에 상호 배치가 발생하므로, 이를 고려하여 핑거프린트 추출 과정을 최적화할 필요가 있다. 기존의 연구 결과들은 대부분 강인성을 중점을 두고 설계된 특징을 사용하였으며, 핑거프린트가 갖춰야할 여러 조건들 간의 균형을 맞추기 위해서 학습을 적용한 경우는 많지 않았다. 최근 이러한 상황을 고려한 연구결과들[5]이 나오고 있으나, 본 논문에서는 기존의 연구 방법들과 다르게 핑거프린트를 이진화하는 과정에 학습을 직접 적용하였다. 기존의 방법[5]에서는 특징을 선택하는 과정에 학습을 적용하고, 이진화는 따로 수행하였다. 제안된 방법은 이진화를 동시에 고려할 수 있도록 경첩 손실(hinge-loss)을 목적함수로 하여 이진 핑거프린트 추출 함수를 최적화 하였다. 핑거프린트 값으로 실수와 이진수를 모두 사용할 수 있으나, 일반적으로 이진수를 사용하는 것이 핑거프린트의 조건들 중 간결성을 높여서 같은 크기의 DB에 수용할 수 있는 오디오 파일의 개수가 늘어나는 장점이 있다. 경첩 손실을 고려한 목적함수로부터 이진 핑거프린트 추출 함수를 학습하기 위해서 MLH(Minimal Loss Hashing) 학습[8]을 적용하였다. 핑거프린트의 조건들 중 가장 중요한 2가지 특징인 식별성과 강인성을 MLH 학습에 고려하였다. 부밴드 기반 여러 다른 오디오 특징들에 MLH 학습을 적용하여 이진 핑거프린트 함수를 구하였으며, 학습 전후와 비교하여 모든 특징들에서 성능의 개선을 확인하였다. 특히 MLH 학습 과정에서 고려되지 않은 오디오 변환들에 대해서도 성능의 개선이 있음을 확인하였다.

본 논문에서는 이진 핑거프린트 추출 함수를 설계하는 목적함수를 제안하고, 실제 학습을 통해 이진 핑거프린트의 식별성과 강인성을 향상 시킬 수 있음을 실험으로 확인하였다. 또한 제안된 방법은 이진 핑거프린트를 사용하므로 간결성이 우수하며, 핑거프린트 추출 시 계산복잡도가 크게 증가되지 않는다. II장에서는 경첩 손실에 기반한 이진 핑거프린트 추출 함수를 위한 목적함수를 제안하고, MLH 학습을 적용한다. III장에서 학습된 핑거프린트들의 오디오 인식 성능을 실험하고 결과를 비교 분석한다.

II. 제안된 이진 핑거프린트 추출 함수

본 논문은 핑거프린트 추출 함수에 학습을 적용하여 성능을 개선하고자 한다. 기존의 논문에서는 대부분 저자들의 직관에 의해서 강인할 것으로 예상되는 특징들을 조합하여 핑거프린트로 사용하였다. 이러한 과정을 최적화하기위해서 일부 기존 연구들[5]에서 핑거프린트를 위한 특징 추출 학습에 최적화를 적용하였다. 즉, 기존 연구들은 핑거프린트 이진화

과정이 아니라 특징추출과정에 최적화를 수행하였다. 본 논문에서는 특징이 선택된 후에 이진화하는 과정에 적용할 수 있는 최적화 기법을 연구하였다. 특히 본 논문은 이진 오디오 핑거프린트에 대해서 다루며, 널리 사용되고 있는 데이터 이진화 방법인 RP_LSH(Random Projection 기반 Locality-Sensitive Hashing) 방법을 사용하였다.[8-9] 이 방법은 데이터 벡터 x에 다음과 같이 임의로 선택된 projection matrix W를 곱하고 그 값의 부호에 따라 -1 또는 1의 이진값을 할당하여 식(1)과 같이 입력 벡터 x에 대한 이진 핑거프린트 벡터 h_x 를 구한다.

$$H(x) = sign(Wx). \tag{1}$$

식(1)에서 데이터 벡터 x의 평균값을 0으로 미리 조정한 것으로 가정한다. 본 논문에서 사용되는 오디오 특징 벡터 x는 평균을 미리 구하여 차감하여 평균을 0으로 조정하였다. RP_LSH 방법에서는 projection matrix W를 임의로 선택하여 이진 핑거프린트 벡터의 차수를 높일 경우 성능이 보장이 되나, 차수가 낮은 경우 W의 선택에 따라 성능에 차이가 있을 수 있다. 일반적으로 핑거프린트의 경우 이진 핑거프린트 벡터의 차수가 16 또는 32정도로 상당히 작으므로 RP_LSH 를 적용하는 것이 적당하지 않다. 따라서 본 논문에서는 식(1)과 같은 형태의 이진화를 사용하지만, projection matrix W를 핑거프린트의 조건들 중 식별성과 강인성에 기반한 목적함수를 통해 학습하여 이진 핑거프린트의 차수가 작을 경우에도 성능을 개선하고자 한다.

2.1. 오디오 핑거프린트를 위한 부밴드 특징 추출

본 논문에서는 오디오 핑거프린팅에서 널리 사용되고 있는 부밴드의 특징들을 이용하여 이진 핑거프린트를 구하였다. 먼저 핑거프린트를 추출하기 위한 전처리 과정으로 입력 오디오 신호를 모노로 바꾸고, 샘플링 주파수를 11025 Hz로 맞춘다. 일반적으로 음악 유통과정 중에 11025 Hz 이상의 샘플링 주파수가 사용되므로 더 낮은 값으로 전처리하면 핑거프린트 매칭 시 오류를 줄일 수 있다. 리샘플링된 오디오 신호에 4096 길이의 해닝(Hanning) 윈도우를 75 %씩 겹쳐 가면서 적용하고 Fourier 변환을 가한다. 이렇게 주파수 도메인으로 신호를 변환해서 얻은 각 오디오 프레임의 파워 스펙트럼을 Q개의 부밴드로 나눈다. 기존 논문들과 마찬가지로 300 Hz에서 5300 Hz 사이의 16개의 인간 청각의 임계 대역[10]을 부밴드로 사용하였다. 따라서 매 프레임당 16차수의 특징 벡터가 나온다. 사용된 부밴드 특징은 기존 논문들에서 사용되었던 부밴드 무게중심[11], 부밴드 분산[11], 부밴드 평탄도[12]의 세 가지를 고려하였다. 각각 식(2), (3), (4)와 같이 주어진다. 입력 오디오 신호 n번째 프레임의 스펙트럼의 k번째 주파수 계수를 $P[n, k]$ 라고 하면 q번째 부밴드의 무게중심 $C[n, q]$, 분산 $S[n, q]$, 평탄도 $F[n, q]$ 특징은 부밴드 주파수 경계 $B[q]$ 과 간격 $N_q = B[q+1]-B[q]$ 에 대해서 각각 식(2), (3), (4)와 같이 주어진다.

$$C[n,q] = \frac{\sum_{k=B[q]+1}^{B[q+1]} kP[n,k]}{\sum_{k=B[q]+1}^{B[q+1]} P[n,k]}, \tag{2}$$

$$S[n,q] = \left[\frac{\sum_{k=B[q]+1}^{B[q+1]} (k - C[n,q])^2 P[n,k]}{\sum_{k=B[q]+1}^{B[q+1]} P[n,k]} \right]^{1/2}, \tag{3}$$

$$F[n,q] = \frac{\left[\prod_{k=B[q]+1}^{B[q+1]} P[n,k] \right]^{1/N_q}}{\sum_{k=B[q]+1}^{B[q+1]} P[n,k]/N_q}. \tag{4}$$

Fig. 1에 나온 바와 같이 얻어진 부밴드 특징들은 식(1)의 이진화 과정을 거쳐서 이진 핑거프린트를 얻는다. 본 논문에서는 학습 오디오 파일들로부터 각 특징별로 평균을 구하고 이를 차감하여 zero mean을 만들어 사용하였다.

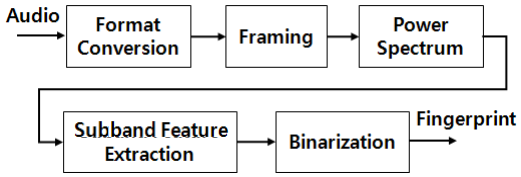

Fig. 1. Block diagram of the proposed fingerprint extraction.

2.2. 경첩 손실 함수 기반 이진 핑거프린트 목적함수

핑거프린트의 성능을 최적화하기 위해서 특징을 이진화하여 핑거프린트를 만드는 식(1)에서 projection matrix W를 핑거프린트의 식별성과 강인성을 고려하여 학습해야 한다. 이를 위해서 다양한 종류의 음악 파일들로부터 N개의 오디오 프레임들을 임의로 선택하고 부밴드 특징을 추출하여 원본 오디오의 특징벡터들의 집합 $O=\{o_1, o_2, ..., o_N\}$을 구성한다. 또한 선택된 오디오 프레임들을 왜곡한 후 부밴드 특징을 추출한 변형 오디오의 특징벡터들의 집합 $M=\{m_1, m_2, ..., m_N\}$을 구성한다. Fig. 2에서처럼 원본 오디오와 변형되어 왜곡된 오디오에서 각각 얻은 특징들을 비교 분석하여 최적의 이진 핑거프린팅 파라미터를 구해야 한다. 이상적인 이진 핑거프린트 함수 H는 $H(o_i)$ 들 간의 거리는 최대화하고, $H(o_i)$와 $H(m_i)$ 간의 거리는 최소화해야 한다. 이진 핑거프린트의 경우 식(1)에서처럼 부호를 취하는 과정 등의 비선형 연산이 수반될 수 밖에 없으므로 직접적으로 핑거프린트 거리값들에 대해서 최적화를 적용하는 것이 어렵다. 또한 일반적으로 핑거프린트 DB 검색 또는 매칭 시에 완벽히 모든 핑거프린트 값이 일치하지 않더라도 일정한 문턱값 이상만 동일할 경우 같은 오디오라고 가정하게 된다. 이러한 사안들을 고려하여 이진 핑거프린트 추출 함수 설계를 위한 목적함수 유도를 위해서 경첩 손실을 사용하였다. 즉, 원본과 해당 원본을 변형시킨 오디오 프레임에서 각각 얻은 이진 핑거프린트 $H(o_i)$와 $H(m_i)$ 간의 해밍 거리 $D_H()$가 정해진 값 λ 이하일 경우 손실이 0 이고, λ 이상이면 해밍 거리와 λ의 차이값을 손실로 정하였다. 반대로 서로 다른 오디오 프레임에서 각각 얻은 이진 핑거프린트 $H(o_i)$와 $H(o_j)$ 간의 해밍 거리가 정해진 값 λ 이상일 경우 손실이 0 이고, λ 보다 작을 경우 손실을 해밍 거리와 λ의 차이값으로 정하였다. 이와 같은 경첩 손실은 식(5)과 (6)에 수식으로 나타내었다.

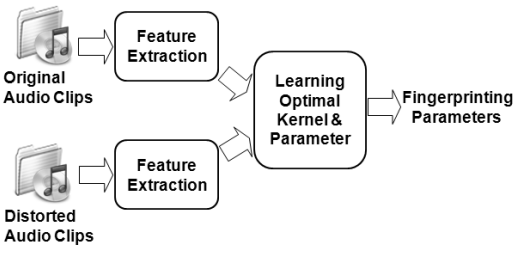

Fig. 2. Learning optimal fingerprinting parameters.

$$l_w(i) = \max(D_H(H(o_i), H(m_i)) - \lambda + 1, 0), \quad (5)$$

$$l_b(i,j) = \max(\lambda - D_H(H(o_i), H(o_j)) + 1, 0). \quad (6)$$

따라서 학습에 사용된 전체 N개 오디오 프레임에 적용하면 식(7)과 같이 핑거프린트 추출 함수 H()에 대한 목적함수가 유도된다.

$$L_H = \sum_{i=1}^{N} l_w(i) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} l_b(i,j). \quad (7)$$

식(7)을 보면 l_w에 비해 l_b 계산 시에 더 많은 경우가 있으므로, L_H에 l_b가 더 큰 영향을 미치게 된다. 핑거프린트의 조건에서 l_w는 강인성에 해당하고, l_b는 식별성에 해당하므로, 두 조건에 같은 가중치를 두기 위해서 실제 실험에서는 여러 종류의 변환을 가하여 변형 오디오 특징벡터들의 집합 M을 여러 개 만들었고, 식(7)에서 l_b 계산 시 에는 전체 쌍을 모두 고려하였으나 실제로는 (i,j) 쌍 중 임의로 선택하여 L_H 계산에서 l_w와 l_b가 같은 개수만큼 사용되었다. 자세한 설명은 III 장에 기술되어 있다.

2.3. MLH를 통한 목적함수 최적화

이진화 과정인 식(1)은 실제로 식(8)에서 이진 벡터인 h를 찾는 것과 같다.

$$H(x) = b(x; W) = \arg\max_{h} [h^T W x]. \quad (8)$$

이진화 과정인 식(8)에서 projection matrix W를 식(7)의 목적함수를 최소화하도록 정하는 것이 주어진 최적화의 목표이다. 경첩 함수를 이용해 얻어진 목적함수는 직접 최적화하하는 것이 어려우므로 식(7)의 상위 경계(upper bound)를 최소화하는 것이 MLH 학습 방법[8]이다. 상위 경계를 최소화하면 식(9)를 만족

하도록 각 특징벡터에 이진 핑거프린트를 할당해야 한다.

$$(\tilde{b}(x_i;W),\tilde{b}(x_j;W)) = \underset{(g_i,g_j)}{\arg\max}[l(x_i,x_j)+g_i^T W x_i + g_j^T W x_j]. \quad (9)$$

식(9)의 $l(x_i, x_j)$는 x_i와 x_j가 지각적으로 동일한 오디오 신호에서 나왔으면 l_w를 사용하고, 지각적으로 다른 오디오에서 나왔으면 l_b를 사용한다. 즉, 현재의 projection matrix W를 이용해서 핑거프린트를 구하는 것은 식(8)의 $b(x_i)$ 로 주어지고, 식(7)의 상위 경계를 더 작게 하기 위해서는 식(9)로부터 구해진 $\tilde{b}(x_i)$가 핑거프린트로 할당되도록 W를 변형시켜야 한다. 이를 위해서[8]에서 stochastic gradient를 이용한 방법을 통해서 식(10)과 같이 iterative하게 W를 업데이트 하게 된다.

$$Wnew = W+\eta[b(x_i;W)x_i^T+b(x_j;W)x_j^T - \tilde{b}(x_i;W)x_i^T-\tilde{b}(x_j;W)x_j^T]. \quad (10)$$

식(10)에서 η는 상수로써 학습률이다. 식(10)의 과정을 2.2절에서 준비한 학습 오디오 특징벡터 집합인 *O*와 *M*에 대해서 적용하여 충분한 학습 이후에 W의 값의 변화가 없으면 학습이 종료된 것으로 한다. 위 최적화에 대한 상세한 유도과정은[8]에 나와 있다.

III. 실험 결과

본 장에서는 학습을 통해서 얻은 이진 핑거프린트의 성능을 실험하였다. 특히 학습 이전과 이후를 비교하여 성능향상의 정도를 확인하였다. 부밴드 특징으로는 2.1절에 언급한 부밴드 무게중심[11], 부밴드 분산 11], 부밴드 평탄도[12] 의 세가지를 고려하였다. 이진 핑거프린트를 구하기 위해 필요한 projection matrix W의 학습을 위해서 다양한 장르의 156곡을 선정하였고, 각 곡에서 100 프레임씩 선택하여 원본 오디오의 특징벡터들의 집합 O를 만들었다(*N*=15600). 또한 왜곡된 오디오의 특징벡터들의 집합 *M*을 만들기 위해서 MP3(48 kbps), auditorium, expander, nbass, notch filter, phone-like filter, small room, white noise 의 8가지 변형을 고려하였다. white noise를 제외한 다른 변형들을 가하는데는 모두 Cool Edit Pro 2.1 소프트웨어를 사용하였다. 실제로 선택된 15600 개의 프레임에 8가지 변형을 모두 가하였으므로 왜곡 오디오 특징벡터 집합인 *M*을 8개 얻게 된다. 따라서 2.2절 언급한 바와 같이 식(7)에서 l_w 계산은 실제로는 얻는 8개의 왜곡 집합을 대상으로 수행하므로 8*N* 번 수행하게 되고, 최적화 과정에서 식별성과 강인성의 균형을 맞추기 위해서 l_b 계산도(i, j) 쌍 중 임의로 선택하여 8*N* 번만 수행하였다. 2.3절에 주어진 방법으로 MLH 학습을 적용하였으며, 이 때 λ=4, $\eta=10^{-6}$ 을 사용하였고, 각 프레임당 이진 핑거프린트는 16차수로 정하였다. 또한 고려한 3가지 부밴드 특징은 모두 16차원이므로, 식(1)의 핑거프린트 함수에서 W는 16행-16열 행렬이다. 기존의 RP_LSH 에서는 W를 평균이 0이고 표준편차가 1인 정규분포로부터 random하게 발생된 행렬을 사용한다. 본 논문에서는 MLH 학습을 통해서 W를 학습데이터로부터 식별성과 강인성이 최적으로 조화가 되도록 구하였다.

일반적으로 콘텐츠 식별 시스템의 성능 비교에는 ROC(Receiver Operating Characteristic) 곡선이 이용된다. ROC 곡선은 인식 시스템에 존재하는 두 가지 형태의 오인식율인 FAR(False Alarm Rate) 과 FRR(False

Table 1. Specification of the composite audio distortion sets used for performance evaluation.

Composite distortion sets	Specification of composite distortions
A1	auditorium, notch filter, small room
A2	expander, phone-like filter, MP3 (48 kbps)
A3	old time radio, pitch increase by 1 %, 30-band classic & pop equalization, MP3 compression (128 kbps)
A4	ambient metal room, pitch decrease by 1 %, 30-band classic & pop equalization, MP3 compression (128 kbps)
A5	super loud, linear speed change by 1 %, 30-band classic & pop equalization, MP3 compression (128 kbps)
A6	rich chamber, time-scale modification by 4 %, 30-band classic & pop equalization, MP3 compression (128 kbps)

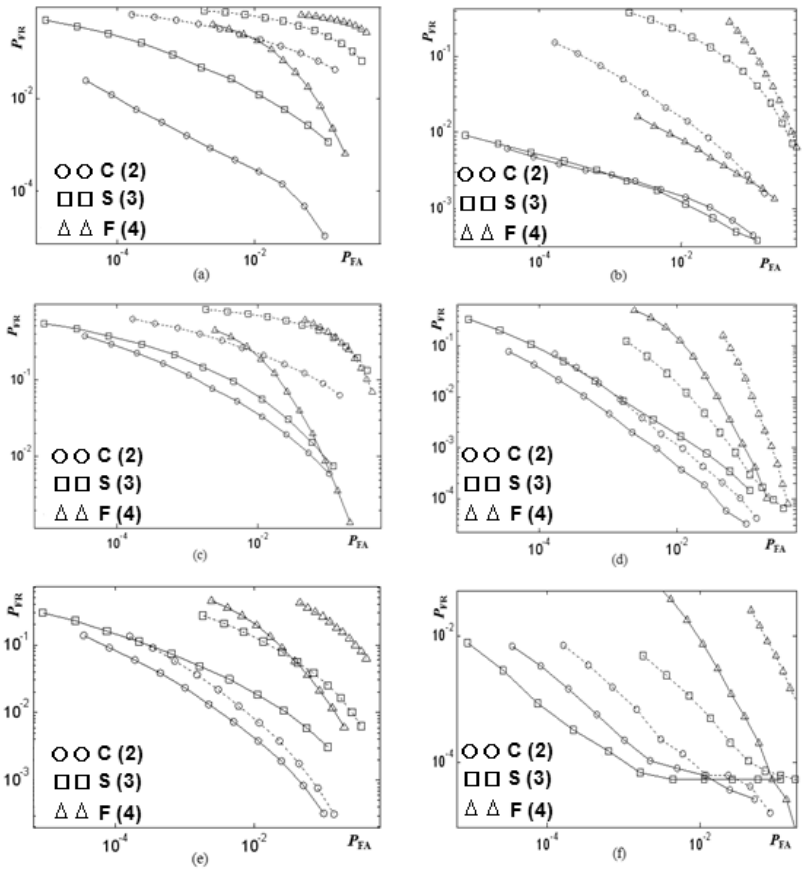

Fig. 3. ROC curves of the binary fingerprinting for — proposed binary fingerprinting, — RP_LSH [9]; ○○ Eqn. (2), □□ Eq. (3), △△ Eq. (4). (a) Test A1 (b) Test A2 (c) Test A3 (d) Test A4 (e) Test A5 (f) Test A6.

Rejection Rate)을 가로와 세로축으로 하여 그래프를 그린 것이다. 오디오 핑거프린팅 시스템에서 FAR은 서로 다른 오디오를 같다고 판정할 확률이며, FRR은 같은 오디오를 다르다고 판정할 확률이다. 실험을 위해서 수집된 8000곡의 음악으로부터 얻은 부밴드 무게중심, 부밴드 분산, 부밴드 평탄도 특징으로부터 학습과정을 통해서 얻은 W를 이용하여 식(1)과 같은 방법으로 핑거프린트 DB를 만들고, 다양한 왜곡들에 대한 식별 실험을 통해 ROC 곡선을 구하여 성능을 비교하였다. FAR을 구하기 위해서는 구축된 핑거프린트 DB에서 임의로 선택된 핑거프린트 쌍들 간의 해밍 거리를 구하고, 오디오 식별기의 문턱값을 변화시켜가면서 문턱값보다 작은 거리를 가지는 핑거프린트 쌍의 비율을 구하였다. FRR을 구하기 위해서 각 음악 파일에 다양한 종류의 변형을 가하였다. 본 논문에서는 6종류의 복합 왜곡에 대해서 실험하였다. 가해진 왜곡의 상세내용은 Table 1에 정리되어 있다. 복합 왜곡 A_1과 A_2는 학습에 사용된 왜곡들과 겹치며, 복합 왜곡 A_3, A_4, A_5, A_6은 학습과 겹치지 않는 왜곡들로 구성되어 있다. Fig. 3 의 ROC 곡선들을 얻기 위해서 프레임 단위의 매칭은 충분한 식별성이 없으므로, 기존의 다른 방법들[3,11] 과 마찬가지로 5초 구간의 오디오 신호의 핑거프린트(5초는 2.1절의 프레임 길이를 고려하면 54프레임이며, 각 프레임당 핑거프린트는 16 비트이므로 864 비트가 사용됨)를 이용해서 핑거프린트 매칭을 수행하여 FAR과 FRR을 구하였다. Fig. 3의 ROC 곡선을 구하기 위해서 RP_LSH에서 random projection W를 10번 발생시켜 FAR, FRR을 실험으로 구하고 그 평균값을 취하였다. 또한 제안된 방법에서도 10번 학습을 수행하여 매 학습된 W에 대해서 FAR, FRR을 실험으로 구하고 그 평균값을 취하였다. ROC 곡선 결과를 보면 학

습하지 않고 임의로 선택된 W를 사용하는 RP_LSH에 비하여 학습된 W를 사용함으로써 핑거프린팅 시스템 성능이 크게 개선됨을 확인할 수 있다. Fig. 3에 주어진 ROC 곡선들을 간편히 비교할 수 있도록 Table 2에 ROC 곡선들의 EER(Equal Error Rate)을 구하였다. EER은 인식 시스템에 존재하는 두 가지 형태의 오인식율인 FAR과 FRR이 같은 값을 가질 때의 오인식율이다. 즉, ROC 곡선이 원점을 지나고 기울기 1인 직선과 만나는 접점에서의 값이 된다. 일반적으로 EER이 작을수록 실제 적용에서 오인식 가능성이 낮다. ROC 곡선과 EER 값을 비교하면 고려한 세 가지 특징 모두에서 학습을 통해서 성능 개선을 확인할 수 있으므로 제안된 방법은 일반적으로 다른 종류의 특징들에도 적용되어 성능을 향상시킬 수 있을 것으로 생각된다. 고려한 세 가지 특징들 중에서는 부밴드 무게중심이 가장 좋은 성능을 보였으며, 부밴드 분산이 조금 낮은 성능을 보였고, 부밴드 평탄도는 좋지않은 성능을 보였다. 기존 RP_LSH와 비교해서 학습을 통한 상대적인 성능 향상도는 왜곡별로 차이가 있었다. 학습에 사용된 왜곡을 그대로 사용한 복합 왜곡 A_1과 A_2 에서 상대적으로 성능 개선이 크다는 것을 확인할 수 있다. 학습에 사용되지 않은 왜곡들만을 사용한 A_3, A_4, A_5, A_6 의 경우에도 어느 정도 성능이 개선되는 것을 확인하였다. 제안된 학습 방법을 통해서 임의의 특징으로부터 이진 핑거프린트를 얻는 과정을 최적화할 수 있음을 보였고, ROC 곡선을 구하여 실제로 식별성과 강인성의 성능이 개선됨을 확인하였다. 사용된 이진 핑거프린트는 핑거프린트 DB의 크기를 줄이고, 해밍 거리를 사용할 수 있으므로 계산용이성면에서도 장점이 있다.

Table 2. EER for the RP_LSH [9] and the Proposed MLH fingerprinting for distortion sets (DT SET) and features.

DT SET	C		S		F	
	RP	MLH	RP	MLH	RP	MLH
A1	7.4E-2	1.3E-3	1.7E-1	1.2E-2	3.4E-1	3.9E-2
A2	1.3E-2	2.3E-3	6.8E-2	2.0E-3	1.1E-1	8.2E-3
A3	8.6E-2	2.2E-2	2.3E-1	2.9E-2	2.2E-1	4.0E-2
A4	3.5E-3	2.1E-3	1.3E-2	4.1E-3	6.3E-2	2.6E-2
A5	9.6E-3	6.2E-3	4.9E-2	1.5E-2	1.7E-1	4.9E-2
A6	1.1E-3	5.1E-4	3.0E-3	2.7E-4	3.6E-2	9.3E-3

IV. 결 론

본 논문에서는 핑거프린트 추출 함수의 최적 선택을 위해서 경첩 손실 함수를 유도하고 학습을 통한 강인한 이진 오디오 핑거프린팅 방법을 제안하였다. 특히 핑거프린팅 시스템의 성능에 가장 큰 영향을 주는 식별성과 강인성 간의 상호 배치되는 관계를 MLH 학습을 통해서 최적화하였다. 수 천곡 규모의 오디오에 대해서 다양한 변환들에 대한 인식 성능을 실험하였으며, 제안된 경첩 손실 함수 최소화에 기반한 학습을 통해서 이진 핑거프린트의 성능이 개선됨을 확인하였다. 제안된 이진 핑거프린트 학습 방법은 여타 다른 종류의 특징에도 적용되어 핑거프린트 추출 함수 최적화에 사용가능할 것으로 기대된다.

감사의 글

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2012012876).

References

1. W. J. Yee, K. K Lee and K. S. Park, “A study on the Efficient feature vector extraction for music information retrieval system” (in Korean), J. Acoust. Soc. Kr. **23**, 532-539 (2004).
2. M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” Proceedings of the IEEE **96**, 668-696 (2008).
3. J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system,” in Proc. International Conf. on Music Information Retrieval, (2002).
4. P. Cano, E. Battle, T. Kalker, and J. Haitsma, “A review of audio fingerprinting,” Journal of VLSI Signal Processing **41**, 271-84 (2005).
5. D. Jang, C.D. Yoo, S. Lee, S. Kim, and T. Kalker, “Pairwise Boosted Audio Fingerprint,” IEEE Tr. Information Forensics and Security **4**, 995-1004 (2009).
6. M. Mohri, P. Moreno, and E. Weinstein, “Efficient and robust music identification with weighted finite-state

transducers," IEEE Tr. Audio, Speech, and Language Processing **18**, 197-207 (2010).
7. J. S. Seo and S. J. Lee, "Robust audio fingerprinting using compressed-domain features" (in Korean), J. Acoust. Soc. Kr. **28**, 375-382 (2009).
8. M. Norouzi and D.J. Fleet, "Minimal loss hashing for compact binary codes," in Proc.International Conference on Machine Learning, (2011).
9. P. Indyk and R. Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality." in Proc. ACM symposium on Theory of computing. (1998).
10. E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models* (Springer-Verlag, 1999).
11. Jin S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. Yoo, "Audio fingerprinting based on normalized spectral subband moments," IEEE Signal Processing Letters **13**, 209-212 (2006).
12. J. Herre, E. Allamanche, and O. Hellumth, "Robust matching of audio signals using spectral latness features," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 127-30 (2001).

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: 강릉원주대학교 전자공학과 조교수, 부교수