다중 레벨 양자화 기법 기반의 음악 검색기 구현

송원식, 박만수, 김회린
한국정보통신대학교
{songcode78, mansoo, hrkim}@icu.ac.kr

Music retrieval system implementation based on multi-level quantization scheme

Wonsik Song, Mansoo Park, Hoirin Kim
Information and Communications University

요 약

본 논문은 필립스의 오디오 핑거프린트 추출 방식을 기반으로 기존의 방식이 주파수 영역을 너무 조밀하게 분석하는 특징을 지적하고 개선 방안으로 양자화를 통해 필터 뱅크의 에너지 변화율을 오디오 핑거프린트 추출시 반영하는 방법을 제안하였다. 또한 제안된 알고리즘을 사용하여 PDA 로 실제 어플리케이션을 구현하는 것을 목적으로 하고 있다. 제안된 방식은 필립스 방식과 동일한 메모리 크기를 유지하기 위하여 필터 뱅크의 개수를 33 개에서 17 개로 줄이고 필터 뱅크의 변화량을 2 비트로 할당하는 방식을 사용하였다. 변화량을 비트에 할당하기 위하여 음악 데이터 베이스로부터 추출된 각 밴드의 pmf 를 통해 음악의 고유성을 최대로 증진 시킬 수 있는 임계치를 찾아내고 이것을 바탕으로 필터 뱅크의 변화량을 2 비트로 할당하였다. 이 같이 추출된 오디오 핑거프린트를 기반으로 PDA 와 음악 검색기 서버와의 통신을 이용하여 사용자가 요청한 쿼리 음악에 관련된 정보를 제공하는 시스템을 구현했다. 제안된 방식은 다양한 주변 잡음 환경에서 평가되어 기존의 필립스 방식 보다 성능 향상 물론 검색 속도 또한 개선되는 특징을 확인할 수 있었다.
.

Keyword : 오디오 핑거프린트, 필터 뱅크 에너지 변화량, 양자화, PDA

1. 서론

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 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 기반한 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 필터 뱅크 에너지 변화량을 확률 통계적 특성을 고려하여 양자화 시킴으로써 좀 더 효율적으로 음악을 표현 할 수 있는 오디오 핑거프린트 추출 기법을 제안하고 제안된 알고리즘을 사용한 실제 어플리케이션을 구축하였다.

논문은 다음과 같은 순서로 기술되었다. 2 장에서는 기존의 오디오 핑거프린트 추출 방법에 대하여 살펴보고, 3 장에서는 제안된 다중 레벨 양자화 기법을 이용한 오디오 핑거프린트 추출 기법 및 PDA 를 이용한 음악 검색 시스템에 대하여 기술하였다. 4 장과 5 장에서는 실험 결과 분석 및 결론에 대하여 언급하였다.

2. 기존의 오디오 핑거프린트 추출 방법

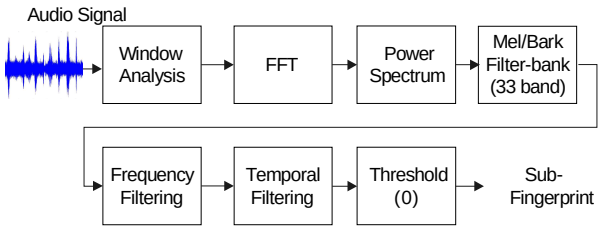

그림 1. 필립스의 오디오 핑거프린트 추출 과정

필립스의 오디오 핑거프린트 기법은 그림 1 처럼 33 개의 필터 뱅크 에너지의 필터 링에 의해 추출된 값의 부호를 기준으로 추출된다. 필터 링을 통해 추출된 값은 필터 뱅크의 에너지 변화량을 나타내며, 이러한 과정을 통해 추출된 32 비트 오디오 핑거프린트는 sub-fingerprint 또는 해쉬 값으로 표현된다.

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

그림 1 의 전체 필터 링 과정은 식 (1)로 정의 된다. 식 (1)에서 EB(n,m)는 n 차 프레임의 m 차 필터 뱅크 밴드[8][9]의 에너지를 나타내고 E(n,m)는 필터 뱅크의 에너지 변화량을 나타내게 된다. 추출된 필터 뱅크의 에너지 변화량은 식 (2)에 의해 부호에 따라 1 개의 비트를 할당하고, 추출된 32 개 비트의 조합을 통해 한 개의 해쉬 값을 추출하게 된다.

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

이 같이 추출된 32 비트 해쉬 값은 고유한 특성 때문에 데이터 베이스 색인에 직접적인 접근 포인트로 사용된다. 이 같은 검색 방법은 효율적으로 검색 영역을 제한할 수 있는 기법으로 사전에 결정된 Hamming Distance 에 대한 색인 목록까지만 검색하게 된다.

필립스 방식에서는 유사도 측정 방법으로 해쉬 값들의 Hamming Distance 를 사용한다. 즉 연속되는 해쉬 값들의 Hamming Distance 스코어가 유사도를 판별하는 척도가 되는 것이다. 최종 검색 결과는 사전에 결정된 오디오 핑거프린트의 블럭 당 BER(Bit Error Rate)를 임계치로 사용해 검색 결과를 검증하게 된다.

그러나 32 개의 필터 뱅크의 에너지 변화율의 부호해 의해 추출된 해쉬 값은 변화량의 정보는 버리고 기울기 정보만 취하는 것이다. 또한 약 300~3000Hz 의 주파수 밴드를 33 개의 필터 뱅크로 나누는 것은 너무 세밀하게 주파수 특성을 보는 경향이 있다. 즉 너무 조밀한 분석은 왜곡에 대한 오디오 핑거프린트의 민감도 증가와 각 필터 뱅크들 사이의 연계성을 증가시킬 수 있는 특징을 가진다.

3. 제안된 알고리즘 및 PDA 를 이용한 어플리케이션 구현

3-1 다중 레벨 양자화 기법을 적용한 오디오 핑거프린트 추출 방법

이상적인 오디오 핑거프린트 기법은 추출된 정보의 메모리 사이즈, 강인성, 고유성 등의 척도에 의하여 그 효용성을 측정할 수 있다. 일반적으로 이 같은 요소들은 상이한 특징을 가지므로 여러 가지 요소들을 적절히 고려하여 오디오 핑거프린트를 추출하여야 한다. 본 논문에서는 왜곡에 강인한 오디오 핑거프린트 추출을 위해 필터 뱅크의 밴드 수를 줄이는 방법을 선택하였고, 고유성을 증가시키기 위해 필터 뱅크의 상대적인 에너지 변화량을 사용하였다. 즉 그림 2 처럼 오디오 핑거프린트 추출 과정에서 필립스 방식과 달리 필터 뱅크의 밴드 수를 33 에서 17 로 줄임으로써 잡음에 대한 민감도를 낮추고 2 비트를 이용해 필터 뱅크 에너지 변화량의 크기 정보와 부호 정보를 오디오 핑거프린트에 매핑함으로써 고유성을 증진시켰다. 오디오 핑거프린트 추출 방법은 2 비트 양자기의 임계치 추출 과정과 오디오 핑거프린트 추출 과정으로 이루어진다.

3-1-1 임계치 추출 과정

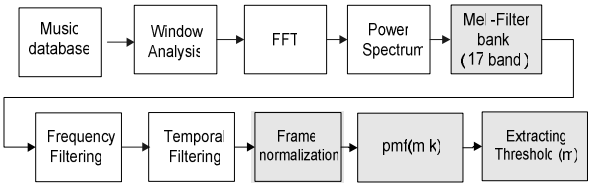

그림 2. 임계치 추출 과정

임계치 추출 방법은 그림 2 처럼 필터 뱅크의 밴드 수를 17 개로 줄여 필립스 방식과 동일하게 필터 뱅크의 에너지 변화량을 추출한다. 추출된 필터 뱅크의 에너지 변화량은 식 (3)를 이용하여 각 프레임의 각 밴드의 절대 값의 합으로 각 필터 뱅크 값을 정규화 시켜준다. 정규화된 필터 뱅크의 에너지 변화량을 이용해서 각 밴드의 pmf 를 구한다. 이 과정은 임계치를 구하기 위한 사전 작업으로써 평균 값은 0 를 갖고 laplacian 모양을 갖는 pmf 가 구해진다.

$$E_N(n,m) = \frac{E(n,m)}{BN(n)}\ ,\ m = 1...16$$
$$BN(n) = \frac{\sum_{m=1}^{16} |E(n,m)|}{16} \qquad (3)$$

$$pmf(m,k) = \frac{\Delta(m,k)}{nframe}$$
$$\Delta(m,k) = \#\ of\ \ \varepsilon(k)_- < E_N(n,m) < \varepsilon(k)_+ \qquad (4)$$
$$nframe = \#\ of\ \ frame\ in\ music\ database$$
$$n = 1,,,nframe,\ k = N, -N+1, ..., -N$$

추출된 각 밴드의 pmf 를 이용해서 양자화를 위한 임계치를 추출하게 된다. 임계치는 각각의 pmf 의 면적을 양자화 레벨 수로 나누는 포인트를 임계치로 추출하게 된다.

3-1-2 오디오 핑거프린트 추출

오디오 핑거프린트를 추출하는 방법은 그림 3 처럼 임계치를 추출하는 과정과 동일하게 필터 뱅크의 에너지 변화량을 추출하고, 사전에 구해진 임계치 값을 사용하여 비트를 할당하는 과정이다.

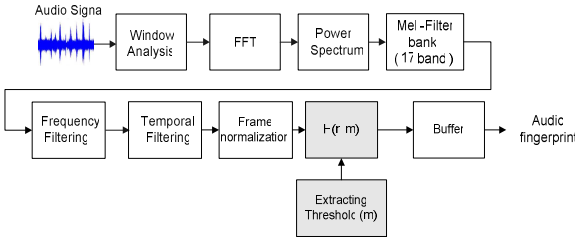

그림 3. 오디오 핑거프린트 추출 과정

음악 데이터 베이스를 통해 추출된 임계치를 통해 식 (5)처럼 2 비트를 할당하게 된다. 식 (5)에서 볼 수 있듯이 추출된 값의 크기에 따라 4 개의 다른 값을 갖게 된다. 제안된 방법은 유사도 측정을 위하여 ED(Euclidean Distance)와 HD(Hamming Distance)가 결합된 형태의 Modified HD 를 사용하였다.

$$H(n,m)=\begin{cases}11 & if \quad E_n(n,m)\geq threshold_3(m)\\ 10 & if \quad threshold_2(m)\leq E_n(n,m)<threshold_3(m)\\ 01 & if \quad threshold_1(m)\leq E_n(n,m)<threshold_2(m)\\ 00 & if \quad E_n(n,m)<threshold_1(m)\end{cases} \quad (5)$$

Modified HD 는 식 (6)처럼 2 비트 값들의 ED 합을 기준으로 오디오 핑거프린트의 유사도를 측정하게 된다. 제안된 방식은 오디오 핑거프린트들 사이의 일치성 보다는 유사성 측정에 초점을 두었기 때문에 BER(Bit Error Rate)대신에 BDM(Bit Dissimilar Measurement)으로 유사도를 표현된다.

$$bdm = \frac{\sum_{n=1}^{nframe}\sum_{m=1}^{16}\left|H_{query}(n,m)-H_{DB}(n,m)\right|}{48*nframe} \quad (6)$$
$$nframe = \#of\ query\ frames$$

제안된 방식의 검색 영역 확장 방법은 2 비트 단위로 추출된 오디오 핑거프린트 값을 기준으로 가장 작은 레벨 차이를 갖는 2 개의 오디오 핑거프린트 값을 찾아 확장하는 방식이다. 본 논문에서는 이 같은 확장 방법을 NSC(Number of Search Candidate)=2 확장으로 표기하였다.

3-2 PDA 를 이용한 실제 어플레케이션 구현

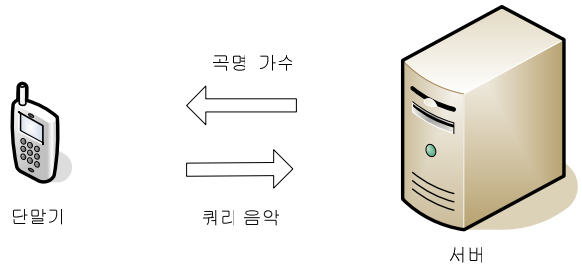

그림 4. 음악 검색 과정

그림 4 는 일반적인 음악 검색 과정을 보여준다. 그림은 실제 음악 검색 어플리케이션 구축을 위해서는 소비자 단말기와 음악 검색 서버와의 통신이 필수적인 요소임을 보여준다. 즉 현재 소비자들이 사용하는 단말기는 계산 속도나 저장 용량 등에 많은 제약을 받으므로 검색과 같은 복잡한 과정은 서버를 통해 이루어지고 검색 결과만을 단말기를 통해 확인하는 방식이 이용되는 것이다.

본 연구에서 사용자 단말기로 주변에서 쉽게 접할 수 있는 PDA 를 사용하였다. 실제 구축된 음악 검색 어플리케이션에서의 PDA 의 역할은 쿼리 음악의 일부분을 녹음하고, 녹음된 음악을 검색 서버로 전송는 부분을 담당한다. 서버는 전송된 음악을 토대로 일치하는 음악을 검색하고 결과를 PDA 로 넘겨주는 역할을 하게 된다.

3-2-1 PDA 를 이용한 음악 검색 과정

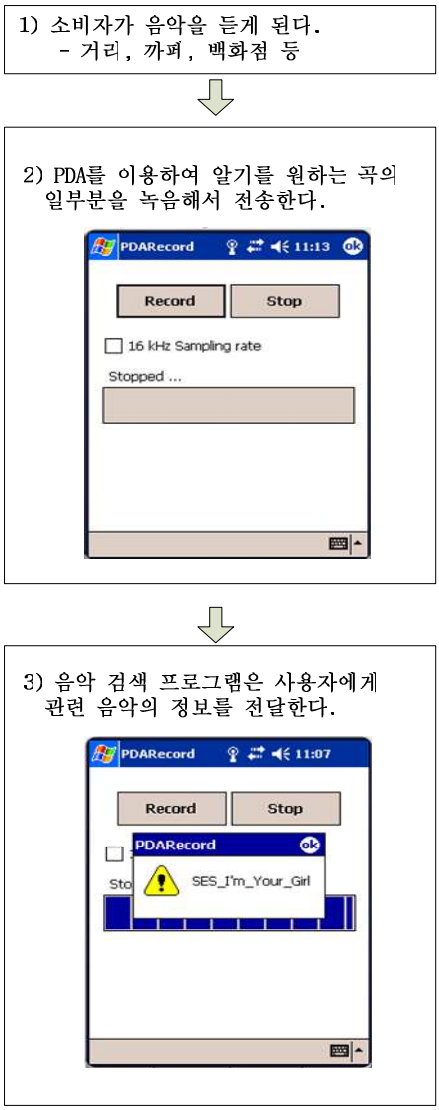

그림 5. 음악 검색 과정

그림 5 는 실제 구축된 음악 검색기의 음악 검색 과정을 보여준다. 구축된 시스템은 사용자의 직관성을 높이기 위하여 단순한 인터페이스를 제공하였다. 즉 단순히 녹음 버튼과 정지 버튼을 이용하여 음악을 검색하는 것이다. 사용자가 원하는

곡의 정보를 얻기 위해서는 쿼리 음악의 일부를 녹음 버튼과 정지 버튼을 사용하여 녹음 시키면 해당 곡을 바로 서버로 보내도록 설계하였다. 이 같이 보낸진 정보는 서버에서의 음악 검색 과정을 거친 후 그림 5 의 3)과 같이 음악의 정보를 얻게 된다. 즉 PDA 를 통하여 곡명(I' m your girl)과 가수(SES)의 정보를 얻게 되는 것이다.

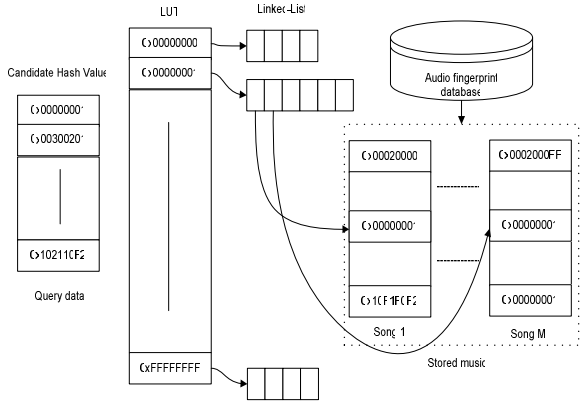

그림 6. 데이터 베이스 구조

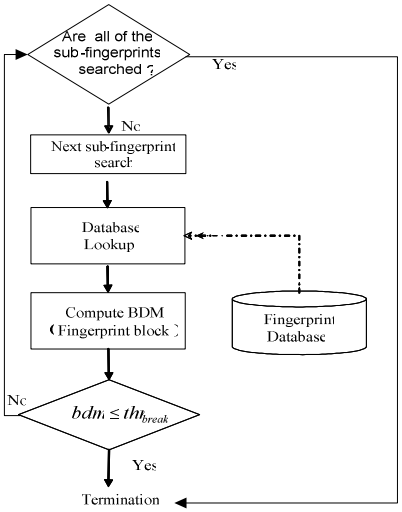

그림 7. 검색 과정

서버로 전송된 쿼리 음악은 그림 6 과 같이 데이터 베이스의 접근 포인트로 사용된다. 즉 전손된 쿼리 음악은 오디오 핑거프린트 블락으로 추출되고 이를 바탕으로 NSC = 2 방식으로 확장하여 그림 6 의 candidate hash value 들을 검출하는 것이다. 추출된 값을 기반으로 look-up table 에 접근하여 후보 곡의 후보 구간에 접근하는 것이다. 그림 7 는 쿼리 음악의 검색 과정을 보여주고 있다. 일정 임계치 보다 작은 값을 갖는 음악을 찾을 때까지 검색하고 일치하는 음악을 찾은 경우 해당 정보를 사용자에게 제공하는 것이다.

4 실험 결과

4-1 실험 데이터

여러 가지 주변 잡음에서의 성능을 테스트하기 위하여 몇 가지 잡음이 발생할 수 있는 상황을 설정하여 노이즈 데이터를 수집하였다. 추출된 노이즈 데이터는 거리, 백화점, 자동차, 사무실, 식당 등의 실제 환경에서 빈번히 발생할 수 있는 상황을 선택하였고, 데이터 수집 시 MD(Sharp: IM-DR 580H)을 사용하여 노이즈 데이터를 추출 하였다. 이처럼 추출된 노이즈 데이터는 왜곡 정도에 따른 제안된 알고리즘의 성능을 측정하기 위하여 SNR(signal to noise ratio)에 따라 10, 5, 0dB로 나누어 데이터를 추출하였다. 각 쿼리 데이터의 길이는 7초로 하였고, 각 실험마다 1,056개의 데이터를 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.

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

4-2 실험 결과 및 검토

실험 결과는 성능을 평가하기 위하여 필립스의 오디오 핑거프린트 방식과 제안된 방식을 비교하여 성능을 표시 하였다. 모든 실험은 검색 영역을 확장한 경우만을 고려하여 실험 하였다. 실험은 5가지(자동차, 거리, 사무실, 백화점, 식당) 왜곡 환경을 SNR에 따라 나누어 실험하였다.

표 1은 제안된 방식과 필립스 방식 모두 검색 영역을 32배 확장한 경우의 음악 검색 성능을 표시한 것이다. 제안된 방식이 일반적으로 뛰어난 성능을 보이는 것을 볼 수 있다. 또한 왜곡이 심

해 질수록 제안된 방식의 성능 향상이 증가되는 것을 볼 수 있다. 이 같이 특징은 제안된 방식이 필립스 방식에 비해 왜곡에 강인한 특징을 갖는 것을 보여준다.

표 1. 주변 잡음에 대한 검색 성능

잡음 환경	필립스 방식 (%)			제안된 방식 (%)		
	10dB	5dB	0dB	10dB	5dB	0dB
자동차	100	99.60	98.50	100	99.81	98.67
거리	98.76	96.78	91.76	99.05	97.34	94.03
사무실	99.43	97.53	92.80	99.43	97.72	96.21
백화점	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

아래의 표 2 는 쿼리 음악당 평균 색인 목록 검색 횟수를 나타낸다. 색인 목록 검색 횟수는 음악 검색 속도를 측정하는 척도로써 표 2 에서 볼 수 있듯이 왜곡이 심해질수록 평균 색인 목록 검색 횟수가 현저히 증가하는 것을 볼 수 있다. 이 같은 특징은 쿼리 데이터를 기준으로 데이터 베이스에 접근하기 때문에 쿼리 데이터에 왜곡이 심할 경우 검색 성능 저하뿐만 아니라 검색 속도 또한 증가 하는 특징을 갖는 것이다. 이 같은 특징을 기반으로 필립스 방식과 제안된 방식을 비교했을 경우 제안된 방식이 더 적은 평균 색인 목록 검색 횟수를 나타냈다. 즉 제안된 방식이 필립스 방식에 비하여 성능뿐만 아니라 검색 시간에서도 더 뛰어난 성능을 보였다.

표 2. 평균 색인 목록 검색 횟수

필립스 방식			제안된 방식		
10dB	5dB	0dB	10dB	5dB	0dB
32.22	74.106	157.94	27.55	67.43	151.59

5. 결론

본 논문에서 필터 뱅크 에너지 변화량를 이용한 오디오 핑거프린트 검출 기법을 제안하였다. 제안된 방식은 필터 뱅크 밴드 수를 17 개로 줄임으로써 왜곡에 대한 강인성을 증대 시키고, 필터 뱅크 에너지의 변화량 정보를 확률적 분포를 고려하여 추출된 오디오 핑거프린트에 추가함으로써 오디오 핑거프린트의 고유성을 증대시키는 방법을 제안하였다. 또한 PDA 를 이용하여 사용자가 손쉽게 이용할 수 있는 음악 검색 프로그램을 구축하였다. 제안된 방식은 실험을 통하여 주변 잡음에 강인한 특징을 보일 뿐만 아니라 검색 속도 또한 빠른 특징을 보였다.

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.

<Acknowledgement>

본 연구는 정보통신부 및 정보통신연구진흥원의 디지털미디어연구소 지원사업의 연구결과로 수행되었음.

참고문헌

[1] Mansoo Park et al., “Content-based Music information Retrieval using Pitch Histogram of Band Pass Filter Signal,”*Proc. of AIRS2004*, pp.245-248, Oct. 2004.

[2] J. Herre, E. Allamanche, and O. Hellmuth,”Robbust matching of audio signals using spectral flatness features,” *Proc. of Workshop on Applications of Signal Processing to Audio and Acoustics2001*, *IEEE*, pp. 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*, pp. 197-204, 2001.

[4] Jonathan T. Foote, “Content-Based Retrieval of Music and Audio,” *Proc. of SPIE, Multimedia Storage and Archiving Systems II*, Vol. 3229, pp. 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*, Sept. 2001.

[9] J.A. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System,” *Proc. ISMIR2002*, pp. 144-148, Oct. 2002.