특집논문-04-09-1-01

Pitch 히스토그램을 이용한 내용기반 음악 정보 검색

박 만 수*, 박 철 의*, 김 회 린*, 강 경 옥**

Content-based Music Information Retrieval using Pitch Histogram

Mansoo Park*, Chuleui Park*, Hoi-Rin Kim* and Kyeongok Kang**

요 약

본 논문에서는 내용 기반 음악 정보 검색에 MPEG-7에 정의된 오디오 서술자를 적용하는 방법을 제안한다. 특히 pitch 정보와 timbral 특징들은 음색 구분을 용이하게 할 수 있어 음악 검색뿐만 아니라 음악 장르 분류 또는 QBH(Query By Humming)에 이용될 수 있다. 이러한 방법을 통하여 오디오 신호의 대표적인 특성을 표현 할 수 있는 특징벡터를 구성 할 수 있다면 추후에 멀티모달 시스템을 이용한 검색 알고리즘에도 오디오 특징으로 이용 될 수 있을 것이다. 본 논문에서는 방송 시스템에 적용 하기 위해 영화나 드라마의 배경음악에 해당하는 O.S.T 앨범으로 검색 범위를 제한하였다. 즉, 사용자가 임의로 검색을 요청한 시점에서 비디오 컨텐츠로부터 추출한 임의의 오디오 클립만을 이용하여 그 컨텐츠 전체의 O.S.T 앨범 내에서 음악을 검색할 수 있도록 하였다. 오디오 특징벡터를 구성하기 위해 필요한 MPEG-7 오디오 서술자의 조합 방법을 제안하고 distance 또는 ratio 계산 방식을 통해 성능 향상을 추구하였다. 또한 reference 음악의 템플릿 구성 방식의 변화를 통해 성능 향상을 추구하였다. Classifier로 k-NN 방식을 사용하여 성능 평가를 수행한 결과 timbral spectral feature 보다는 pitch 정보를 이용한 특징이 우수한 성능을 보였고 vector distance 방식으로는 특징들의 비율을 이용한 IFCR(Intra-Feature Component Ratio) 방식이 ED(Euclidean Distance) 방식보다 우수한 성능을 보였다.

Abstract

In this paper, we proposed the content-based music information retrieval technique using some MPEG-7 low-level descriptors. Especially, pitch information and timbral features can be applied in music genre classification, music retrieval, or QBH(Query By Humming) because these can be modeling the stochasticpattern or timbral information of music signal. In this work, we restricted the music domain as O.S.T of movie or soap opera to apply broadcasting system. That is, the user can retrievalthe information of the unknown music using only an audio clip with a few seconds extracted from video content when background music sound greeted user's ear. We proposed the audio feature set organized by MPEG-7 descriptors and distance function by vector distance or ratio computation. Thus, we observed that the feature set organized by pitch information is superior to timbral spectral feature set, and IFCR(Intra-Feature Component Ratio) is better than ED(Euclidean Distance) as a vector distance function. To evaluate music recognition, k-NN is used as a classifier

Keywords : MPEG-7 Audio Descriptor, Intra-Feature Component Ratio, Pitch, Timbral Spectral, Music Information Retrieval

* 한국정보통신대학교 공학부
School of Engineering, ICU

** 한국전자통신연구원 방송미디어연구부
Electronics and Telecommunications Research Institute

※ 본 논문은 한국전자통신연구원 "지능형 방송서비스 핵심기술 개발"에 관한 공동연구과제 수행의 일환으로 얻어진 연구결과 입니다.

I. 서 론

디지털 방송 서비스가 활성화 되면서 시청자에게 다양한 기능을 제공할 필요성이 제시 되고 있고 현재 지능형 TV 서비스를 위한 SmarTV와 같은 미래 지향적인 디지털 TV

시스템에 관한 연구가 활발히 진행되고 있다. 그 중에서 시청자가 원하는 정보를 효율적으로 검색할 수 있는 기능이 중요시 되고 있다. 컨텐츠의 모든 정보가 메타데이터에 서술되어 있다면 시청자에게 정보제공을 손쉽게 할 수 있다. 하지만 모든 정보를 메타데이터에 서술하기 위해서는 대부분을 수작업으로 작성해야 한다. 이러한 단점을 보완하기 위해 컨텐츠 내용 기반의 특징을 이용하여 검색을 수행하게 된다. 이 경우 검색에 필요한 템플릿 구성 및 메타데이터를 사람의 수작업 대신 자동으로 생성하게 된다. 예를 들어, 오디오 분야에서는 오디오 신호의 특징만을 이용한 음악 장르 분류[1][2], QBH(Query By Humming)[3], 그리고 음표 검출[4]과 같은 연구가 진행되고 있고, 비디오 특징과 더불어 오디오 특징의 조합을 이용한 멀티모달 시스템에 관한 연구도 활발히 진행되고 있다. 또한 이러한 검색 기능이 가능하도록 다양한 특징에 관한 연구도 진행되고 있다. 특히 MPEG-7에서는 방송 컨텐츠의 내용 및 특징을 서술할 수 있도록 다양한 서술자를 정의하고 있다. MPEG-7 서술자는 컨텐츠 내용 기반의 특징을 추출할 수 있기 때문에 사용자가 원하는 정보를 자동으로 검색하는데 효율적으로 이용될 수 있다.

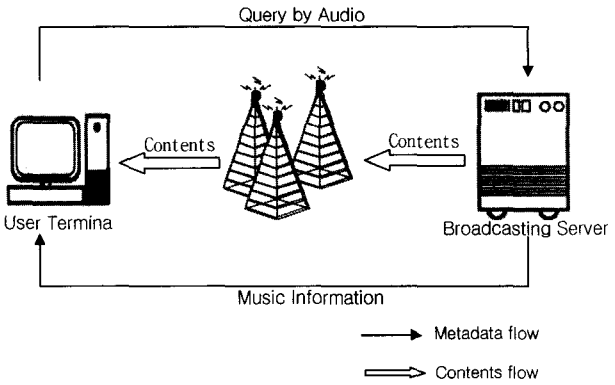


그림 1. 오디오 검색을 이용한 배경음악 정보제공 서비스
Fig. 1. Music information proffer service using content-based audio retrieval technique

그림 1은 오디오 쿼리를 이용한 음악 정보 제공 서비스의 한 예를 나타낸다. 시청자가 현재 방송 중인 컨텐츠의 배경음악의 정보를 원할 경우 O.S.T 앨범 내에서 그 배경음악의 정보를 시청자에게 제공하는 서비스이다. 즉, 시청자가 검색을 요청 할 경우 컨텐츠에서 배경음악에 해당하는 오디오 신호 일부를 추출하여 오디오 검색을

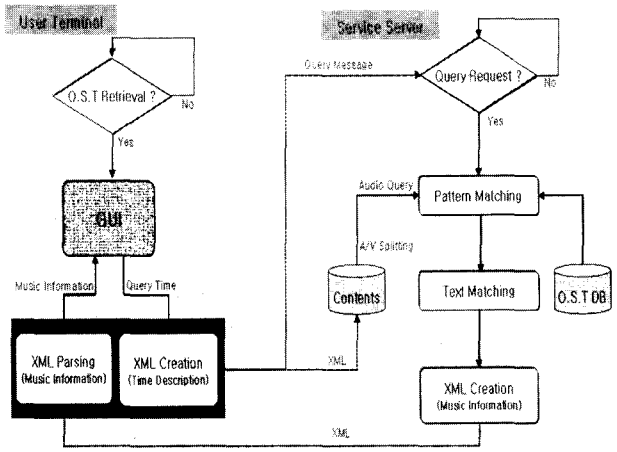


그림 2. O.S.T 검색 시스템 구조
Fig. 2. The system architecture for O.S.T retrieval system

수행하고 그 결과에 해당하는 음악 정보를 시청자에게 제공한다. 이러한 서비스를 제공하기 위한 시스템 구조를 그림 2에서 나타내고 있다. 메타데이터에 의해 양방향 전송이 가능하기 때문에 TV 단말에서 검색요청을 위한 오디오 쿼리와 검색결과에 해당하는 배경음악 정보는 메타데이터 형태로 MPEG-7 표준화에 맞는 XML 형태로 전송된다. 본 논문에서는 방송 서비스에 적용할 수 있는 오디오 검색 시스템을 구성하기 위해 MPEG-7 오디오 서술자[5]의 조합을 통해 오디오 신호의 특징벡터를 추출하였다.

Ⅱ. 특징벡터 구성

1. MPEG-7 오디오 하위 서술자

MPEG-7 오디오 하위 서술자[5]는 오디오 신호의 다양한 특징들을 표현할 수 있다. 그림 3은 MPEG-7 오디오 하위 서술자의 framework[6]을 나타낸다. 그림에서 나타나듯이 오디오 신호의 기본 파라미터 부터 음악이나 악기의 음색을 나타낼 수 있는 timbral spectral 파라미터 까지 다양하다. 총 18개의 temporal 및 spectral 서술자를 포함하고 있고 의미적으로 8개 group으로 구분할 수 있다. 여기에서 'D'는 서술자(Descriptor)를 의미한다. 본 논문에서는 이 그룹 중에 음색 정보를 표현하는 "Timbral Spectral"과 오디오 신호의 멜로디를 표현할 수 있는 "Signal Parameters"를

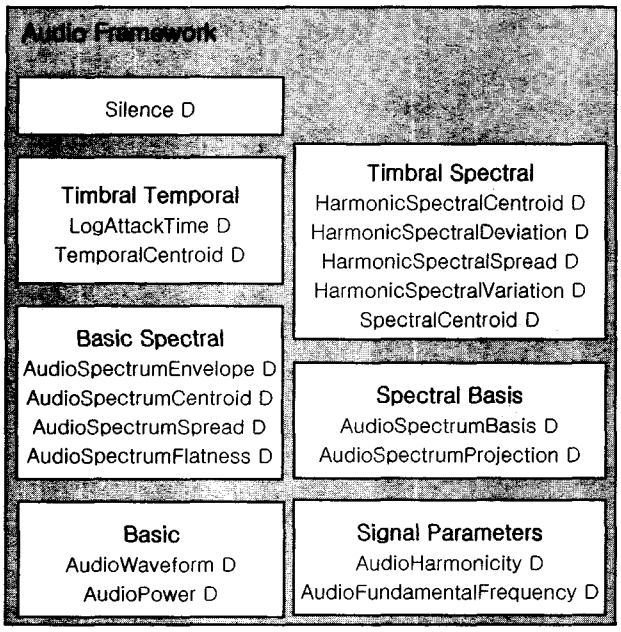


그림 3. MPEG-7 오디오 하위 서술자 framework
Fig. 3. MPEG-7 audio framework (low-level descriptors)

조합하여 오디오 특징으로 사용하였다.

2. 음색을 나타내기 위한 특징벡터 구성

음악 검색 서비스는 여러 장르가 섞여 있고 음악 파일 일부 클립만을 쿼리로 사용하기 때문에 클래스의 구분이 명확하지 않고 각각의 음악을 표현하기가 어렵다. 따라서 본 논문에서는 서로 다른 음악의 특성을 표현하기 위해 음색을 기반으로 하는 서술자를 이용하여 특징벡터를 구성하였다. 장르에 따라서 또는 사용되는 악기에 따라서 음색의 차이가 존재한다. 또한 가수에 따라서 음색의 차이가 존재한다. 그러한 음색의 차이를 표현하기 위해 본 논문에서는 그림 4와 같이 MPEG-7에 정의된 timbral spectral 서술자[4]들을 사용하였다. 그림 4에 나타나듯이 timbral spectral 특징들은 다음과 같은 특징들로 구성되어있다. 파워스펙트럼의 중심 값을 나타내는 SpectralCentroid는 식 (1)과 같다. 그 외의 timbral spectral 특징은 각 프레임에 해당하는 오디오 신호의 기본 주파수와 하모닉 피크 검출을 통해 구할 수 있다. HarmonicSpectralCentroid는 식 (2)와 같이 하모닉 피크 값의 중심이 되는 주파수를 의미하고 HarmonicSpectralDeviation은 식 (3)과 같이 로그단위의 하모닉 피크의 편차 값을 나타

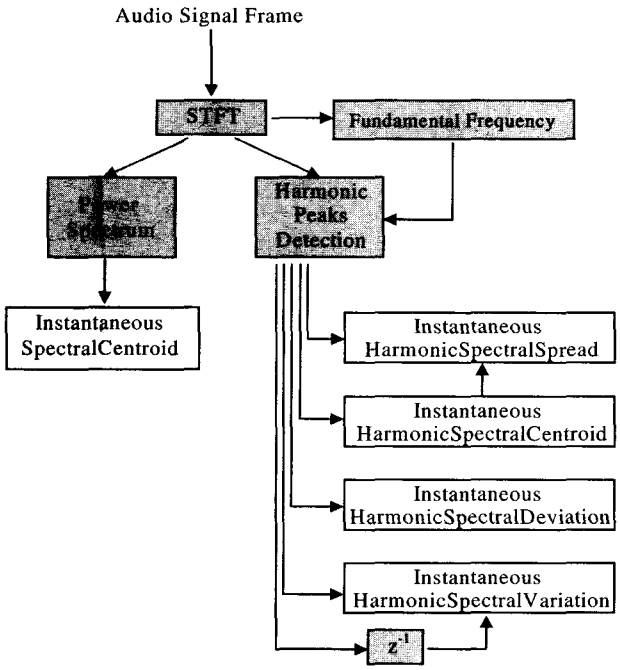


그림 4. 음색 정보를 나타내는 오디오 특징벡터 구성
Fig. 4. The audio feature vector organized by pitch histogram

낸다. 그리고 HarmonicSpecralSpread는 식 (4)와 같이 하모닉 피크의 표준편차를 의미하고 HarmonicSpectral-Variation은 식 (5)와 같이 시간적으로 인접한 프레임간의 variation 값을 나타낸다.

$$ISC(i) = \frac{\sum_{k=1}^{N} f(k) \cdot S(i,k)}{\sum_{k=1}^{N} S(i,k)} \quad (1)$$

$$IHSC(i) = \frac{\sum_{h=1}^{H} f(i,h) \cdot A(i,h)}{\sum_{h=1}^{H} A(i,h)} \quad (2)$$

$$IHSD(i) = \frac{\sum_{h=1}^{H} \left| \log_{10}(A(i,h)) - \log_{10}(SE(i,h)) \right|}{\sum_{h=1}^{H} \log_{10}(A(i,h))} \quad (3)$$

$$IHSS(i) = \frac{\sqrt{\sum_{h=1}^{H} A^2(i,h) \cdot [f(i,h) - IHSC(i)]^2}}{\sum_{h=1}^{H} A^2(i,h)} \quad (4)$$

$$IHSV(i) = 1 - \frac{\sum_{h=1}^{H} A(i-1,h) \cdot A(i,h)}{\sqrt{\sum_{h=1}^{H} A^2(i-1,h)} \cdot \sqrt{\sum_{h=1}^{H} A^2(i,h)}} \quad (5)$$

여기에서 i는 프레임, N은 파워스펙트럼 사이즈, $S(i,k)$는 i번째 프레임의 k번째 파워스펙트럼 계수, 그리고 $f(k)$는 k번째 파워스펙트럼 계수의 해당하는 주파수 값을 나타낸다. $A(i,h)$는 i번째 프레임의 h번째 하모닉 피크 값을 의미한다. 하모닉 피크 값은 오디오의 기본 주파수를 이용하여 구할 수 있다. $f(i,h)$는 i번째 프레임의 h번째 하모닉 피크에 해당하는 주파수 값을 의미하고 $SE(i,h)$는 i번째 프레임의 h번째 하모닉 피크 주변의 스펙트럼 envelope 값을 나타낸다. 스펙트럼 envelope 값은 Basic Spectral 그룹의 AudioSpectrumEnvelope 서술자를 이용하여 구한다.

3. 멜로디를 표현하기 위한 오디오 특징벡터 구성

위와 같이 음색 정보만으로 오디오 특징벡터를 구성할 경우 음색 정보는 주파수 특성에 민감하기 때문에 오디오 쿼리 신호의 포맷이 reference 음악과 다르거나 순수 음악에 해당하는 신호 외에 주변 잡음과 같은 다른 신호가 섞여 있는 경우 음색 정보는 왜곡 될 수 있다. 그러므로 오디오 신호의 포맷 및 주변 잡음에 따른 주파수 특성의 왜곡에 민감하지 않은 멜로디를 효율적으로 표현할 수 있는 오디오 특징을 구성하여야 할 것이다. 이러한 멜로디를 표현하기 위해 QBH에서는 음의 고저/음의 길이를 이용한 UDR(Up, Down, Repeat)/LSR(Longer, Shorter, Repeat) 방식을 고려한다. 하지만 humming의 경우 음의 발생 기관(인간의 vocal)이 하나에 해당하고 음악의 경우 다양한 악기뿐만 아니라 인간의 vocal 까지 어우러져 있기 때문에 표현하기가 상대적으로 어렵다. 또한 단 수초에 해당하는 오디오 쿼리만으로 전체 음악신호에서 어느 부분에 해당하는지 알 수 없기 때문에 UDR/LSR 방식만으로는 음악 검색이 용이하지 않을 것이다. 이에 본 논문에서는 음의 고저와 음의 길이가 모두 포함 될 수 있는 pitch 히스토그램 방식을 제안한다. 즉, 오디오 신호의 기본 주파수를 최소 62.5Hz에서 최대 1.5kHz 범위 내에서 bilinear scale로 등분하여 총 72차로 구성된 pitch의 히스토그램을 오디오 특징으로 사용하였다. 한 음악에서는 특성상 확률적으로 유사한 음표에 해당하는 pitch가 반복해서 나오는 특성이 존재한다. Pitch 히스토그램은 이러한 패턴을 반영할 수 있기 때문에 상대적으로 음악을 검색하는데 용이할 것이다. 본 논문에서는 MPEG-7에 정의된 AudioFundamentalFrequency 서술자를 이용하여 pitch를 검출하였고 pitch 히스토그램을 보다 정교하게 하기 위해 그림 5에서 보듯이 각 프레임마다 주기성을 판단하여 하모닉 특성이 존재 할 경우만 pitch 히스토그램에 반영하였다. 또한 효율적인 검색을 위해 주기성을 판단하는 복잡한 알고리즘 대신 오디오 검색 시간을 효과적으로 줄일 수 있도록 그림 5와 같이 기본 주파수가 1.5kHz 이상인 경우는 비 주기성 프레임으로 판단하여 pitch 히스토그램에 반영하지 않았다.

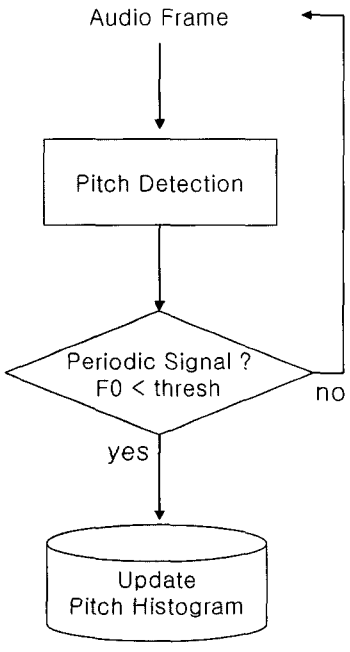


그림 5. Pitch 히스토그램을 이용한 오디오 특징 구성
Fig. 5. The audio feature vector organized by pitch histogram

Ⅲ. Distance 계산

패턴 매칭을 하기 위한 대표적인 vector distance 계산 방식인 ED(Euclidean Distance) 방법은 특징벡터 성분간의 scale 차이로 인해 scale이 큰 특징 성분의 영향을 많이 받는다. 이러한 문제를 보상하고 원하지 않는 신호(잡음)에 의해 음악 신호의 음색 정보의 왜곡이나 pitch의 왜곡에 의해 야기되는 문제를 보상하기 위해 vector distance 계산 방식으로 본 논문에서는 IFCR(Intra-Feature Component Ratio) 방식을 제안한다. IFCR은 식 (6)과 같이 특징벡터의 성분 간의 비율의 곱을 나타내어 이 값이 '1'에 가까울수록 같은 클래스에 근접한다.

$$IFCR = \prod_{n=1}^{D} \frac{\min[o_n, y_n]}{\max[o_n, y_n]} \qquad (6)$$

여기에서 D는 특징벡터의 차원, o_n는 입력(오디오 쿼리) 특징벡터의 n번째 성분, 그리고 y_n는 템플릿(레퍼런스 패턴) 특징벡터의 n번째 성분을 나타낸다. 레퍼런스 템플릿(패턴)을 생성하기 위해 각 음악에 대해 texture window 단위로 모델을 구성하였다. 즉, 음색의 경우 texture window 마다 특징벡터의 평균을 적용하였고 pitch 히스토그램은 texture window 마다 개별적으로 구성하였다. Pitch 히스토그램의 경우 '0'일 확률이 존재하기 때문에 IFCR 적용시 오디오 특징으로 보수 확률($q_{pitch_i} = 1 - p_{pitch_i}$)을 적용하였다.

Ⅳ. 실험 및 결과

본 논문에서 사용한 오디오 특징의 성능을 알아보기 위해 classifier로는 잘 알려진 k-NN rule을 사용하였다. 본 논문에서 12곡으로 구성된 영화 '코요테 어글리'의 O.S.T 앨범(CD)을 이용하여 오디오 쿼리와 reference 사이의 포맷과 음질이 동일한 경우에 대해 성능 평가를 수행하였다. 또한 실제 환경에서 성능을 평가하기 위해 드라마 '다모'의 14부작 비디오로부터 오디오 쿼리를 추출하고 14곡으로 구성된 '다모 O.S.T 앨범(CD)'을 reference로 구성하여 성능을 평가하였다. 이 경우 오디오 쿼리와 reference 사이의 포맷과 음질은 서로 다르다. 즉, 비디오로부터 추출된 오디오 쿼리는 오디오 CD로 구성된 reference와 오디오 포맷뿐만 아니라 배우들의 대사와 배경 잡음에 의해 음질이 서로 다르다. 입력에 해당하는 오디오 쿼리는 '코요테 어글리'의 경우 5초에서 10초 사이의 임의의 구간을 선택하여 각 음악당 30개의 오디오 클립을 추출하여 사용하였고 '다모'의 경우 14부작 비디오로부터 레이블링 후 각 8초의 duration을 갖는 총 3,581개의 오디오 클립을 추출하여 사용하였다. 레퍼런스 템플릿의 생성을 위해 10초와 15초 texture window를 5초 간격으로 overlap을 적용하였다.

표 1. Texture window 길이에 따른 성능변화
Table 1. Performance evaluation according to texture window size

Distance	Texture Window	
	10 sec	15 sec
ED	64.2 %	60.1 %
IFCR	81.9 %	75.8 %

템플릿 생성을 위한 texture window의 길이에 따른 성능을 '코요테 어글리' DB를 사용하여 평가한 결과를 표 1에서 나타내고 있다. 이 경우 음색 정보(timbral spectral)만으로 구성된 오디오 특징을 사용한 결과 오디오 쿼리의 길이가 texture window에 근접할수록 성능이 우수하게 나타난다. 또한 vector distance 방식으로는 IFCR이 ED방식보다 우수한 성능을 나타내고 있다. 즉, IFCR 방식이 각 특징 성분들의 scale 차이에서 오는 문제를 보상해 주고 있다.

표 2. k-NN의 성능 변화 (Texture Window: 10 sec)
Table 2. Performance evaluation by k-NN (Texture Window: 10sec)

k-NN	IFCR
1	81.9 %
3	76.9 %
5	75.3 %
9	72.8 %

k-NN classifier의 k 값에 의한 성능을 '코요테 어글리' DB를 사용하여 평가한 결과를 표 2에서 나타내고 있다. 이 경우 texture window 길이는 10초를 사용하였고 distance 계산은 성능이 우수한 IFCR 방식을 사용하였다. k-NN을 적용할 경우 k 값이 작을수록 우수한 성능을 보이고 있다.

O.S.T의 경우 유사한 음색의 음악들이 존재하기 때문에 본 논문에서 사용한 레퍼런스 템플릿 방식의 경우는 k 값이 증가할수록 성능이 저하된다.

표 3. 오디오 특징 및 오디오 쿼리 구성에 따른 성능변화
Table 3. Performance evaluation according to different audio feature set & audio query data

오디오 특징 구성	오디오 쿼리 구성	
	Audio CD (코요테 어글리)	Video File (다모)
Timbral Spectral	81.9 %	32.4 %
Pitch Histogram	99.7 %	43.1 %
Pitch Histogram + Periodic Signal Detection	100 %	81.4 %

표 3은 표 1과 표 2의 결과를 이용하여 texture window 10초와 k가 1일 때 k-NN을 적용하고 실제 환경에서 오디오 특징 구성방식에 따른 성능을 알아보기 위해 '다모' DB와 '코요테 어글리' DB를 사용한 결과를 나타낸다. 여기에서 '다모' DB는 실제 비디오 파일에서 오디오 쿼리를 추출하였고 '코요테 어글리' DB에서는 오디오 CD로부터 오디오 쿼리를 추출하였다. 표 3에 나타나듯이 음색정보(timbral spectral) 만으로는 실제 환경에서 적용하기에는 무리가 있다. 또한 하모닉 성분의 유무에 관계 없이 무작위의 pitch 히스토그램의 경우에도 실제 환경에서는 성능이 저조했다. 하지만 주기성을 고려한 pitch 히스토그램의 경우 상대적으로 우수한 성능을 나타내어 실제 환경에서도 적용 가능하다.

V. 결 론

본 논문은 오디오 신호의 일부 클립만을 사용하는 오디오 검색 알고리즘을 제안하였고 템플릿 구성시 texture window 길이와 vector distance 계산 방식의 차이에 대한 성능 비교 실험을 수행 하였다. 표 1과 표 2에 보여진 바와 같이 본 논문에서 제안한 IFCR 방식이 우수한 성능을 보였고 k-NN 적용시 k가 1인 경우, 그리고 texture window 길이는 실제 입력 오디오 쿼리 길이에 근접할 경우에 우수한 성능을 보였다. 또한 표 3에서 보여주듯 본 논문에서 제안한 하모닉 성분의 유무에 따른 pitch 히스토그램의 경우가 우수한 성능을 나타내었다.

본 논문의 결과를 기반으로 향후 오디오 검색에 좀 더 적합한 특징 및 MPEG-7 오디오 서술자 와의 조합을 통한 특징벡터 구성에 관한 연구를 지속할 것이다. 또한 오디오 레퍼런스 템플릿 구성을 다양한 VQ(Vector Quantization) 방식에 적용하고 오디오 검색에 적합한 classification 방식을 채택하여 성능향상을 추구할 것이다.

참 고 문 헌

[1] Yibin Zhang, Jie Zhou, "A Study On Content-Based Music Classification," IEEE Proc. 7th International Symposium on Signal Processing and Its Applications, vol. 2 , pp. 113-116, July, 2003.

[2] Tong Zhang, C.-C. Jay Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," IEEE Trans. On Speech and Audio Processing, vol. 9, no. 4, pp. 441-457, May 2001.

[3] Lie Lu, Hong You, H. J. Zhang, "A New Approach to Query by Humming in Music Retrieval," ICME 2001, Aug. 2001.

[4] K. Kashino, H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," Speech Communication, vol. 27, pp. 337-349, 1999.

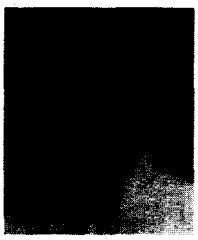
[5] Information Technology Multimedia Content Description Interface Part 4: Audio, ISO/IEC FDIS 15938-4.

[6] Overview of the MPEG-7 Standard (version 6.0), ISO/IEC JTC1/SC29/WG11/N4509.

저 자 소 개

박 만 수

- 2000년 2월 : 인하대학교 전자공학과 졸업(학사)
- 2002년 2월 : 한국정보통신대학교 공학부 졸업(석사)
- 2002년 3월~현재: 한국정보통신대학교 공학부 박사과정
- 주관심분야 : 내용기반 오디오 인덱싱 및 검색, 음성인식



박 철 의

- 2003년 2월 : 전남대학교 정보통신 공학부 졸업(학사)
- 2003년 3월~현재 : 한국정보통신대학교 공학부 석사과정
- 주관심분야 : 내용기반 오디오 인덱싱 및 검색, 음성인식

김 회 린

- 1984년 2월 : 한양대학교 전자공학과 졸업(학사)
- 1987년 2월 : 한국과학기술원 전기및전자공학과 졸업(석사)
- 1992년 2월 : 한국과학기술원 전기및전자공학과 졸업(박사)
- 1987년~1999년 : 한국전자통신연구원 선임연구원
- 1994년~1995년 : 일본 ATR-ITL 방문연구원
- 2000년 2월~현재 : 한국정보통신대학교 공학부 조교수
- 주관심분야 : 음성인식, 화자인식, 음성언어 번역, 내용기반 오디오 인덱싱 및 검색

강 경 옥

- 1985년 2월 : 부산대학교 물리학과 졸업(이학사)
- 1988년 2월 : 부산대학교 물리학과 졸업(이학석사)
- 2004년 2월 : 한국항공대학교 항공전자공학과 졸업(공학박사)
- 1991년 2월~현재 : 한국전자통신연구원 방송미디어연구그룹 3D미디어연구팀장
- 주관심분야 : MPEG-7, TV-Anytime, 오디오 부호화 알고리즘, 음향신호처리, 3-D 오디오