논문 2004-41SP-6-30

내용기반 오디오 장르 분류를 위한 신호 처리 연구

(A Study on the Signal Processing for Content-Based Audio Genre Classification)

윤 원 중*, 이 강 규*, 박 규 식**

(Won-Jung Yoon, Kang-Kyu Lee, and Kyu-Sik Park)

요 약

본 논문에서는 디지털 신호처리를 이용하여 Classic, Hiphop, Jazz, Rock, Speech 등 5개의 오디오 장르를 자동적으로 분류하는 내용기반 오디오 장르 분류기를 제안하였다. 20초 분량의 질의 오디오로부터 23ms 크기의 Hamming window를 이동시켜 가며 Spectral Centroid, Rolloff, Flux 등 STFT 기반의 특징 계수들과 MFCC, LPC 등의 계수들을 구하여 총 54차에 해당하는 특징 벡터 열을 추출하였으며 분류 알고리즘으로는 k-NN, Gaussian, GMM 분류기를 사용하였다. 최적의 특징 벡터를 선별하는 알고리즘으로 총 54차의 특징벡터 중 가장 성능이 좋은 특징 계수들을 찾아 순차적으로 재배치하는 SFS(Sequential Forward Selection)방법을 사용하였고, 이를 이용하여 최적화 된 10차의 특징 벡터만을 선정해서 오디오 장르 분류에 사용하였다. SFS를 적용한 실험 결과 약 90% 가까운 분류 성공률을 보이고 있어 기존 연구에 비하여 약 10%~20% 정도의 성능 향상을 꾀 할 수 있었다. 한편 실제 사용자들이 오디오 자동 장르 분류 시스템을 사용할 때 일어날 수 있는 상황을 가정하여 임의 구간에서 질의 데이터를 추출하여 실험을 수행하였으며 실험 결과 오디오 파일의 맨 앞과 맨 뒤 등 worst-case 질의를 제외하고는 약 80%대의 분류 성공률을 얻을 수 있었다.

Abstract

In this paper, we propose a content-based audio genre classification algorithm that automatically classifies the query audio into five genres such as Classic, Hiphop, Jazz, Rock, Speech using digital signal processing approach. From the 20 seconds query audio file, the audio signal is segmented into 23ms frame with non-overlapped hamming window and 54 dimensional feature vectors, including Spectral Centroid, Rolloff, Flux, LPC, MFCC, is extracted from each query audio. For the classification algorithm, k-NN, Gaussian, GMM classifier is used. In order to choose optimum features from the 54 dimension feature vectors, SFS(Sequential Forward Selection) method is applied to draw 10 dimension optimum features and these are used for the genre classification algorithm. From the experimental result, we can verify the superior performance of the proposed method that provides near 90% success rate for the genre classification which means 10%~20% improvements over the previous methods. For the case of actual user system environment, feature vector is extracted from the random interval of the query audio and it shows overall 80% success rate except extreme cases of beginning and ending portion of the query audio file.

Keywords: Audio Genre Classification, Audio Information Retrieval, Audio Signal Processing, SFS, k-NN

I. 서 론

폭 넓은 인터넷의 확산과 정보통신기술의 발달로 사용자들은 단기간 내에 보다 풍부하고 깊이 있는 정보를 접할 수 있게 되었으며, 정보를 구성하는 데이터들도 사용자들의 요구에 맞추어 이미지나 오디오, 비디오등과 같은 멀티미디어 정보의 형식을 갖추게 되었다. 그러나 시시각각 폭발적으로 늘어나는 방대한 데이터의 양으로 인해 앞으로는 정보 자체보다 정보의 관리 및 정보에 관한 정보가 더욱 가치 있는 정보가 될 것이다. 즉, 저장된 수많은 데이터들도 그 내용이 아무리 중요

* 학생회원, ** 정회원, 단국대학교 컴퓨터과학 및 통계학과
(Dept. of Computer Science and Statistics, Dankook University)
※ 본 연구는 한국과학재단 목적기초연구(R01-2004-000-10122-0)지원으로 수행되었음.
접수일자: 2004년7월2일, 수정완료일: 2004년7월18일

하다 할지라도 적시에 검색되어 활용될 수 없다면 가치 있는 정보라 할 수 없다.

내용 기반 정보검색 시스템은 기존의 수작업으로 이루어지던 텍스트 기반의 분류 시스템과 달리 정보의 내용(contents)을 수학적으로 분석하여 구조화된 기준에 따른 대표적인 특성을 추출하고 컴퓨터를 통해 멀티미디어 데이터를 체계적인 구조로 색인화 한다. 이러한 색인화 및 분류, 검색 작업은 멀티미디어 데이터의 신속하고 정확한 정보취득을 제공할 수 있기 때문에 멀티미디어 관련 응용 연구에 필수적인 기반 기술로 사용될 수 있다.

특히 오디오는 영상·음향·음성 등 멀티미디어 데이터들이 공통으로 포함하고 있는 정보 매체로서 입력 정보의 내용을 분석하여 장르를 분류하고 검색하는데 핵심적인 역할을 한다. 이러한 내용기반 오디오 기술은 인터넷 검색, 디지털 오디오 라이브러리, 의학 및 법률 연구소에서의 자료 정리 또는 개인용 PC에서 음악이나 비디오 파일 정리, 폭발음, 폭풍우, 지진, 동물들의 소리를 포함시키는 필름 후처리 색인화, 그리고 가상현실 시스템 등을 위해 대규모 오디오 데이터베이스로부터 해당 음향 효과 분류 및 검색 등 다양한 응용 분야에 활용될 수 있다[2,9].

기존의 내용기반 오디오 장르 분류 및 검색에 관련된 연구는 1997년 미국 Audible Magic 사의 MuscleFish[1]가 그 효시로서, 최근에는 ACM ISMIR 학술회를 중심으로 다양한 연구 논문이 발표되고 있으며 크게 DSP(디지털 신호처리) 기술을 이용한 방법과 MIDI 파일 내의 음악 표기 정보를 이용하는 2가지 방법이 있다. DSP를 이용한 기법은 오디오의 Pitch, 음색(Timber), 하모니 등의 특징을 추출하여 DB내의 오디오와 비교 검색하는 방법으로 비교적 구현 방법이 복잡한 반면 모든 오디오 파일 포맷에 적용 가능하다는 장점이 있다. 반면 MIDI 파일 내의 정보를 이용한 방법은 MIDI내의 멜로디 음표를 이용한 것으로 비교적 검색 시간이 짧고 구현이 쉽다는 장점이 있으나 적용범위가 제한되는 단점이 있다.

DSP를 이용한 방법으로 논문 [1]에서는 MuscleFish를 통해 15초 미만의 동물소리, 기계소리, 악기소리, 음성 등의 음향 효과들에 대하여 신호의 크기(loudness), Pitch, 밝음(brightness), 대역폭, 하모니 등의 특징들을 추출하여 DB내에서 유사한 오디오를 검색한다. 그러나 각 오디오 음향의 장르 구분은 사용자가 직접 수동으로 해주어야 하는 불편함이 있다. T. Zhang은 [2]에서 HMM을 이용한 계층적 구조의 오디오 분류 및 검색 시스템을 제안하였다. Coarse-level 분류에서는 음악, 음성, 배경음 등의 개괄적인 분류를 제공하고 Fine-level 분류에서 남성 및 여성의 음성, 박수 소리, 폭발음, 새소리 등의 배경음 등에 대한 세밀한 분류를 하였다. 이 연구는 주로 배경음 분류에 치중하였고, 약 80% 정도의 성공률을 보이고 있다. 한편 논문 [3,4]에서 G. Tzanetakis는 음악 장르의 계층적인 자동 분류를 위하여 STFT 기반의 오디오의 표면적 특성과 Wavelet Transform 기반의 Rhythm, Pitch 등의 오디오 특성 정보들을 추출하여 음성/음악/잡음 분류에서 90%, Cla-ssic, Jazz, Fork, R&B 등 Pop 음악 장르 분류에는 약 74% 정도의 성공률을 보였으며 분류기로는 k-NN, Gaussian, GMM 등을 사용하였다.

MIDI 파일 내 멜로디를 이용한 방법으로는 논문 [5]에서의 QBH(Query By Humming)에 대한 연구가 있으며 이는 멜로디 내의 연속된 음들을 U(높음), D(낮음), S(같음)로 표현하여 DB 내 곡들과 UDS 문자열을 비교하여 검색한다. 그러나 이 경우 문자열 정합을 위한 검색 속도에 대한 문제점이 있어 이를 개선하기 위한 다양한 연구가 발표되고 있다[6,7,8].

본 논문에서는 디지털 신호 처리 기법을 이용하여 Classic, Hiphop, Jazz, Rock, Speech 등 5개의 오디오 장르를 자동적으로 분류하는 내용기반 오디오 장르 분류기를 제안한다. 입력 오디오 질의로부터 Centroid, Rolloff, Flux 등의 STFT 기반의 특징들과 MFCC, LPC 등의 특징 벡터들을 추출한 후 DB와 비교하여 자동으로 오디오 장르를 분류한다. 장르 분류에 사용되는 오디오 특징 벡터는 SFS(Sequential Forward Selec-tion) 기법을 사용하여 최적화되며 전체적으로 약 90% 가까운 성공률을 보이고 있다. 특히 본 논문에서는 실제 사용자들이 시스템을 사용할 때를 가정하여 질의 오디오의 임의의 구간에서 특징 벡터를 추출하여 실험에 사용하였다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 제안 시스템의 전체 구조도를 설명하였으며 Ⅲ장에서는 본 논문에서 사용된 특징 벡터와 오디오 장르 분류에 사용된 분류 알고리즘에 대하여 설명한다. Ⅳ장에서는 컴퓨터 모의실험을 통한 비교 분석을 수행하였고, 마지막으로는 결론으로 끝을 맺는다.

Ⅱ. 제안 시스템의 구조

다음의 그림 1은 제안된 내용기반 오디오 장르 분류 시스템의 구조를 나타낸다. 시스템은 일반적인 오디오 파일로부터 RAW 데이터 형식의 오디오 신호만을 입력받게 되고 최종 사용자들에게 해당 입력 신호에 대한 장르 분류 결과를 출력으로 하며, 시스템의 핵심이 되는 오디오 정보에 대한 장르 분류기는 입·출력단 사이에 위치한다.

오디오 장르 자동 분류기의 동작과정은 전처리 및 본처리의 2개 과정으로 이루어진다. 전처리 단계에서는 장르별 특징 테이블을 생성하기 위한 패턴 학습(Train-ing)을 한다. 패턴 학습은 Classic, Hiphop, Jazz, Rock, Speech 등 5개의 오디오 장르에 해당하는 오디오 데이터의 장르별 특성을 모델링하기 위해 많은 수의 오디오 데이터 샘플로부터 특징 벡터를 추출, 분석하여 통계분석학적인 최종 특성 값을 결정하기 위한 작업이다. 즉 시스템에 오디오 파일이 입력(1~4)되면 Segmenta-tion & Windowing(Hamming) 과정(5)을 거쳐 오디오 특징 벡터를 추출하여 최적의 특징 벡터열을 선정(6)하고 모델 정합과 색인화 과정을 포함하는 학습 단계(7-1)를 거쳐 특징 벡터 테이블을 생성하여 이를 오디오 DB에 저장(7-2)하는 과정으로 이해할 수 있다.

반면 본 처리 과정은 사용자 입장에서의 시스템 동작 과정과 일치하는데, 사용자가 오디오 신호파일을 시스템에 입력(1)시키면 시스템은 이 파일을 받아 원시 파일로 변환(1-4)시킨 후, Segmentation & Windowing (Hamming) 작업(5)을 수행하여 선정된 최적의 특징 벡터를 추출(6)하고 오디오 DB에 이미 구축되어있는 특성 테이블을 참고하여 유사한 특성을 지니는 장르를 결정(8-1~2)하여 사용자에게 분류된 장르의 최종 결과(8-3)를 보여주게 된다.

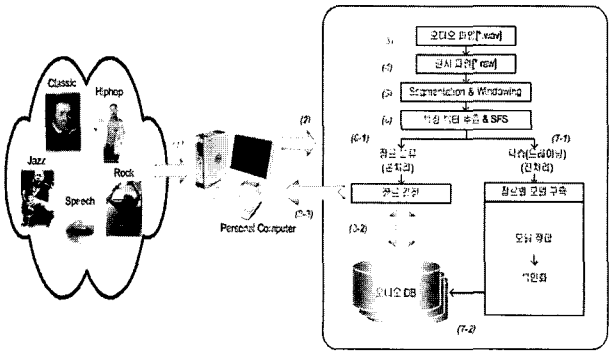

그림 1. 제안된 시스템의 구조
Fig. 1. Structure of proposed system.

Ⅲ. 특징 벡터 추출 및 장르 분류 알고리즘

1. 특징 벡터 추출

각 특징 벡터의 분석 및 추출은 다음과 같은 조건에서 수행하였다. 오디오 신호는 22050Hz, 16bits, mono로 샘플링 되었으며 실험에 사용된 오디오 클립은 20초 분량에 해당한다. 20초 분량의 오디오 신호는 그림 2와 같이 23ms 크기의 Hamming window를 중복되지 않게 이동하면서 각 23ms 프레임으로부터 특징을 추출하여 평균과 분산 값을 조합해서 총 54차의 특징 벡터를 구하게 된다. 본 논문에서는 SFS(Sequential Forward Selection) 기법을 사용하여 총 54차 특징 벡터로부터 최적의 10차 특징 벡터만을 선택하여 사용하게 되며 이는 다음 절에서 상세히 설명하도록 한다.

다음은 본 논문에서 사용된 특징 벡터들을 간략하게 소개하였다.

가. Spectral Centroid[10]

Centroid는 STFT(Short Time Fourier Transform)의 magnitude 스펙트럼의 중심을 뜻한다.

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \tag{1}$$

여기서 $M_t[n]$은 프레임 t와 주파수 Bin n에서의 스펙트럼 magnitude에 해당한다. Centroid는 스펙트럼의 형태를 측정하는 방법 중의 하나이다.

나. Spectral Rolloff

Rolloff는 스펙트럼 magnitude 분포의 80%가 집중되어 있는 주파수 R_t이하를 나타낸다.

$$\sum_{n=1}^{R} M_t[n] = 0.8 * \sum_{n=1}^{N} m_t[n] \tag{2}$$

Rolloff는 스펙트럼의 형태와 낮은 주파수 영역에 신

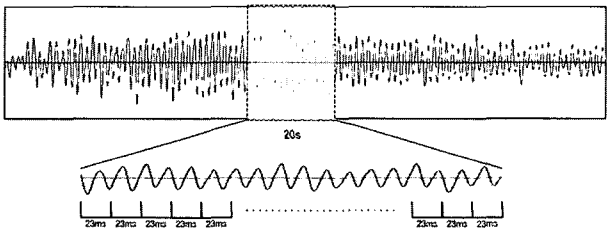

그림 2. 특징 벡터 추출
Fig. 2. Feature vector extraction.

호의 에너지가 얼마나 집중되어 있는지를 보여준다.

다. Spectral Flux[10]

Flux는 연속된 스펙트럼 분포에서 정규화 된 magnitude들 간의 차이를 제곱해서 구할 수 있다.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \quad (3)$$

여기서 $N_t[n]$, $N_{t-1}[n]$은 각각 현재 프레임 t와 이전 프레임 t-1 에서의 FT의 정규화 된 magnitude이다. Flux는 스펙트럼 변화의 양을 계산할 수 있다.

라. ZCR(Zero Crossing Rate)[11]

영 교차율은 오디오 신호 파형의 위상이 중심축을 통과하는 회수를 나타낸다. 영 교차율은 신호의 주파수 내용을 측정하는 가장 간단한 특징으로 음성인식에서 유·무성음의 판별에 사용된다.

마. LPC(Linear Predictive Coding)[12]

선형 예측 계수는 인간의 발성 모델에 입각해서 음성 신호를 부호화하는 방법으로 오디오 파형의 샘플 값에서 필터 계수를 구하여 성대에서 입, 코까지의 성도 특성을 8~12차의 전극형(All-pole) 필터에 근사 시키는 방법이다. 본 논문에서는 10차 계수를 사용하였다.

바. MFCC(Mel Frequency Cepstrum Coefficient)[13]

MFCC는 인간의 청각 특성을 모델링 하는 방법으로 오디오 신호의 magnitude 스펙트럼을 log scale 한 후 FFT bin을 그룹화하여 인간의 청각 특성에 맞는 Mel-Frequency 스케일로 변환한 것이다. 본 논문에서는 13차 계수를 사용하였다.

2. SFS (Sequential Forward Selection)를 이용한 특징 벡터의 최적화

본 절의 각 23ms 프레임에서 추출하게 되는 각 특징 계수는 Spectral Centroid, Rolloff, Flux, ZCR 각 1개, 그리고 10차 LPC, 13차 MFCC로서 총 27개 계수를 20초 오디오 클립 내에서 각 프레임 별 특징 계수의 평균과 분산을 조합하게 되면 총 54차 특징 벡터 계수가 존재하게 된다. 54차 특징 벡터는 모의실험을 위한 소규모 DB에서는 큰 문제가 되지 않지만, 10만~100만 곡 정도 되는 방대한 규모의 DB에서는 분류 속도에 큰 영향을 줄 수 있다. 따라서 본 논문에서는 SFS 기법을 사용하여 특징 벡터의 차원을 줄이고 약 1/5에 해당하는 10차의 최적 특징 벡터만을 선정하여 오디오 장르 분류에 사용하게 된다.

SFS는 먼저 각 특징 계수들을 개별적으로 사용하여 장르 분류를 한 후, 가장 좋은 성공률을 나타내는 특징부터 순차적으로 나열하여 새로운 54차의 특징 벡터 열을 만든다. 다음으로는 첫 번째 특징 벡터의 성분부터 순차적으로 특징 계수의 수를 늘려가면서 혹은 전체 54개의 특징 벡터 열에서 하나씩 특징 계수를 감소시키면서 최적의 성공률을 나타내는 특징 벡터 열만을 찾아낸다. SFS를 사용한 최적 특징 벡터의 선정과 분류기의 성능은 4절의 모의실험에서 자세히 다루기로 한다.

3. 오디오 장르 분류 알고리즘

오디오 장르 분류 알고리즘은 질의 오디오를 Classic, Hiphop, Jazz, Rock, Speech 등 5개의 오디오 장르 중 하나로 분류하는데 사용되며 크게 인공 신경망을 이용한 방법과 패턴 학습 및 인식 기법을 이용한 통계적 방법 등이 있다. 일반적으로 인공 신경망을 이용한 방법은 통계적 방법의 분류 알고리즘 보다 좋은 성능을 보여주지만 복잡도가 높은 반면 통계적 패턴 인식 기법은 연산 속도가 빠르다는 장점이 있다.

본 논문에서는 일반적으로 많이 사용되고 있는 패턴 학습 및 인식 기법으로서 k-NN, Gaussian, GMM 등 3가지 방법을 분류기로 사용한다.

가. k-NN(Nearest Neighbor) 분류기

k-NN 분류기는 임의의 특징 벡터가 어떤 장르와 더 유사한가를 결정하여 분류하는 방법이다. 특징 벡터간의 유사도를 계산하기 위해 비교적 단순하고 직관적인 접근 방법인 Euclidean 거리 함수를 이용하여, 미지의 입력 오디오의 특징 벡터와 장르를 미리 알고 있는 표준 특징 벡터와의 거리가 가장 가까운 장르로 결정되는 최소거리분류 규칙을 따른다. 각 장르별로 DB에 미리 준비된 표준 특징 벡터들과의 거리를 계산하여 가장 가까운 k개의 특징 벡터가 속해있는 장르로 질의 데이터를 분류한다. k-NN은 특징 계수들의 확률분포가 알려지지 않았을 때 유용한 방법이다.

나. Gaussian 분류기

Gaussian 분류기는 각 오디오 장르의 확률밀도함수가 Gaussian 분포를 갖는다고 가정한다. Gaussian 분류기는 DB내 오디오 장르별 평균과 공분산을 미리 계산

하여 Query 오디오가 입력되었을 때 장르별 수식 (4)의 확률 밀도 함수를 계산하여 이 값이 가장 높은 장르를 해당 장르로 결정한다.

$$p(x)=\frac{1}{(2\pi)^{d/2}|C|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'C^{-1}(x-\mu)\right] \quad (4)$$

다. GMM(Gaussian Mixture Model) 분류기

GMM은 각 장르별 확률 밀도 함수를 몇 개의 Gaussian 확률 밀도 함수의 선형 결합으로 정의한다. Gaussian 분류기가 각 장르별 확률 밀도 함수를 단일 Gaussian으로 모델링하는 반면 GMM 분류기는 여러 Gaussian의 선형 결합으로 모델링함으로써 각 장르의 패턴 데이터들을 좀더 정확하게 묘사할 수 있는 장점이 있다. 그러나 패턴 학습 시 GMM 파라 메타 계산을 위한 EM(Expectation-Maximization) 알고리즘이 추가적으로 사용되어 높은 연산속도가 필요하며 알고리즘 구현 시 반복 루프로 인한 연산 정확도가 민감해지는 단점이 있다.

Ⅳ. 실험 결과 및 분석

실험에 사용된 오디오 DB는 인터넷 전문 음악 사이트나 음악 CD로부터 Classic, Hiphop, Jazz, Rock 등 4개 음악 장르에 대해 각 장르별로 60곡을 선정하였으며, Speech는 라디오 방송을 인터넷을 통해 실시간으로 녹음하여 60개 클립을 준비하였다. 실험에서 사용된 모든 오디오 파일은 22050Hz, 16bits, mono의 wave 파일로 변환하여 사용하였다.

음악은 주로 일정한 시나리오를 가지고 연주되기 마련인데 한 곡 내에서도 음악이 속한 장르의 특성이 잘 나타나는 부분이 있고, 그렇지 못한 부분이 있을 것이라 사료되어, 모의실험을 통하여 음악의 장르별 특성이 가장 잘 나타나는 부분을 실험한 결과 곡의 처음 시작 부분부터 약 40%~45% 지점에서 가장 좋은 성능을 보였다. 위 결과를 토대로 본 논문에서는 실험을 위한 학습 데이터를 각 음악의 40% 지점에서부터 20초를 추출하여 사용하였다.

오디오 장르 분류를 위한 패턴 학습 및 인식기는 k-NN, Gaussian, GMM 3가지를 사용하였다. 분류기의 성능 평가를 위해서는 각 장르별로 랜덤하게 90%를 학습 데이터로 사용하고 나머지 10%를 오디오 질의 데이터로 사용하였으며 학습 데이터와 질의 데이터 사이에 동일한 곡이 있는 상황은 배제하였다. 또한 제한된 DB에서 신뢰성 있는 결과를 도출하기 위해서 위 과정을 100번 반복하여 실험 결과를 도출하였다.

먼저 본 논문에서는 학습 데이터와 질의 데이터를 모두 음악적 특성이 비교적 잘 나타나 있는 오디오 파일의 처음 시작 부분부터 40% 지점에서 3절에 언급한 특징 계수들을 추출하여 54차의 특징 벡터를 만들어 실험을 해 보았다. 그림 3은 총 54차의 특징 벡터 전체를 사용하였을 때 각 5개 장르별 분류 성공률을 평균하여 나타내고 있으며 그림에서 보듯이 전체적으로 약 70%를 하회하는 결과가 나와 논문 [3]의 68%에 해당하는 분류 결과와 별 차이가 없음을 알 수 있다. 각 장르별 분류 성공률은 다음의 표 1에 명시하였다. 표에서 보듯이 Jazz의 분류 성공률이 현저히 낮게 나오고, 그 중 Classic, Hiphop, Rock으로의 오분류율이 50%정도 이며 이는 Jazz라는 음악 특성이 Classic, Hiphop, Rock 등 비교적 여러 장르의 음악 특성을 골고루 가지고 있기 때문으로 분석된다. Hiphop은 Rock에 대해서 약 20%의 오분류율을 나타내고, Rock 또한 Hiphop에 대하여 비슷한 오분류율을 나타내고 있다. 인간이 듣기에는 랩

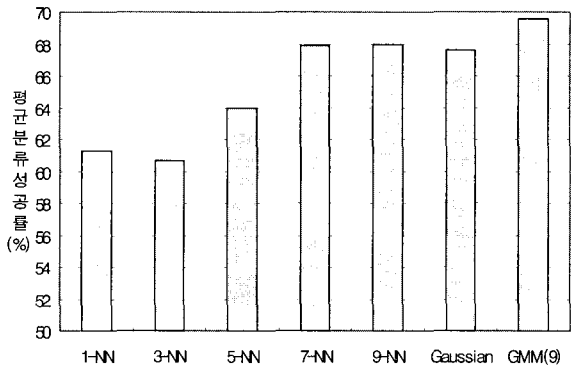

그림 3. 54차 특징 벡터를 사용한 결과
Fig. 3. The result of using 54 dimensional feature vectors.

표 1. 오디오 장르별 분류 성공률(54차 특징 벡터 사용)
Table 1. Audio genre classification accuracy. (Using 54 dimensional feature vectors)

	Classic	Hiphop	Jazz	Rock	Speech	분류 성공률
Classic	51	0	5	2	2	85%
Hiphop	0	40	6	14	0	67%
Jazz	11	12	27	7	3	45%
Rock	1	15	8	33	3	55%
Speech	1	1	2	2	54	90%

과 중저음, 강렬한 비트가 주를 이루고 있는 Hiphop과 Rock은 확연히 비교가 되지만, Rock의 특징으로 역시 강렬한 비트를 들 수 있으므로 이러한 부분에서 두 장르간의 오분류가 일어나는 것으로 판단되어진다. 이러한 음악적 특징에 따른 오분류율 등은 논문 [4], [9] 에서도 유사하게 지적되고 있다. 반면, Classic과 Speech는 다른 장르에 심하게 간섭받지 않고 대부분 정확하게 분류하는 모습을 볼 수 있다.

다음의 그림 4는 3절에서 언급한 SFS를 이용한 최적의 특징 벡터 선정 과정을 보이고 있다. SFS는 총 54차 특징 벡터를 대상으로 먼저 각 특징 계수들을 개별적으로 사용하여 장르 분류를 한 후, 가장 좋은 성공률을 나타내는 특징부터 순차적으로 나열하여 새로운 54차의 특징 벡터 열을 만든다. 다음으로는 첫 번째 특징 벡터의 성분부터 순차적으로 특징 계수의 수를 늘려가면서 최적의 성공률을 나타내는 특징 벡터 열만을 찾아낸다. 그림에서 보듯이 특징 벡터의 수가 10~13개 사이에서 약 90%의 성공률을 보이는 반면 그 이상에서는 오히려 분류 성공률이 떨어지는 결과를 보이고 있다. SFS 실험 결과 최적의 특징 벡터 열은 {Ceps2_m, Roll_m, LPC2_m, Cent_m, LPC1_m, Ceps1_m, LPC5_m, Ceps3_m, Flux_m, LPC3_m}의 10차 벡터이다. 여기서 Ceps2는 MFCC 2차 계수, Roll은 Spectral Rolloff 계수, Cent는 Spectral Centroid 계수, LPC1은 LPC 1차 계수 등을 말하며 _m 은 각 계수의 평균값을 의미한다.

이러한 SFS 최적 특징 벡터 선정 방법은 오디오 특징들의 통계적 특성을 감안하지 않은 방법으로서 k-NN 분류기에서 가장 잘 동작이 되며 Gaussian, GMM 등의 통계적 기반 분류기에서는 현저하게 저하된 성능을 나타내었다. 표 2에서는 SFS를 통해 얻은 최적의 10차 특징 벡터를 이용한 장르 분류 성공률을 보여주고 있다. Classic, Hiphop 그리고 Speech에서의 분류 성공률은 아주 만족할 만한 결과를 나타내고 있다. Jazz와 Rock에서의 성공률도 표 1과 비교했을 때 10~20%정도 좋아졌지만 여전히 다른 장르들에 비하여 30%정도 낮은 성공률로 분류기 전체의 성공률을 떨어뜨리고 있다.

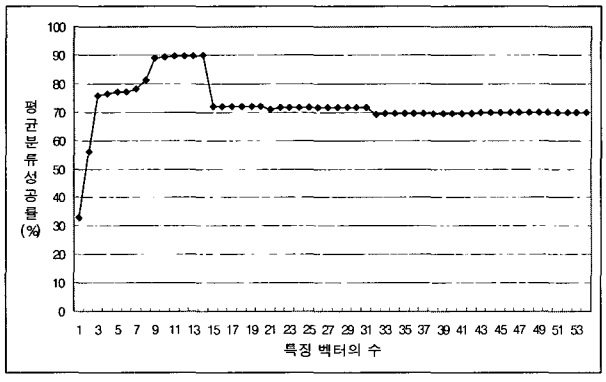

그림 4. SFS를 이용한 최적의 특징 벡터 선정
Fig. 4. Optimum feature vector selection using SFS.

앞서 그림 3과 4의 실험에서는 해당 오디오 장르 특징이 비교적 잘 나타나고 있는 오디오의 도입 부분에서부터 40% 지점에서만 20초 분량의 특징 벡터를 추출하여 질의 오디오로 실험에 사용하였다. 하지만 실제 사용 환경에서는 사용자들이 오디오의 어느 부분을 질의로 사용할지는 알 수 없다. 따라서 실제 응용 가능한 시스템 구축을 위해서는 오디오 파일의 처음 도입 부분부터 끝 구간까지 임의의 20초 분량의 오디오 클립 질의에 대하여 신뢰성 있게 동작할 수 있는 시스템 구축이 중요하다. 이를 위해 본 논문에서는 20초 분량의 질의 오디오를 오디오 파일의 처음 시작부분, 시작 부분으로부터 +20초, 오디오 파일 전체의 10%, 20%, 30%, 40% 지점 그리고 오디오 파일의 끝 지점에서 -40초, -20초로 질의 오디오 클립의 시작 지점을 설정하여 실험을

표 2. 오디오 장르별 분류 성공률(최적의 10차 특징 벡터 사용)
Table 2. Audio genre classification accuracy. (Using optimum 10 dimensional feature vectors)

	Classic	Hiphop	Jazz	Rock	Speech	분류 성공률
Classic	55	1	2	2	0	92%
Hiphop	0	52	6	2	0	87%
Jazz	10	4	38	8	0	63%
Rock	2	6	12	40	0	67%
Speech	1	0	0	0	59	98%

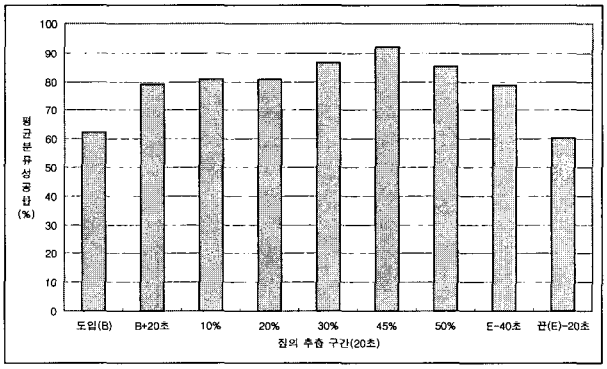

그림 5. 임의 구간 질의 데이터에 대한 실험 결과
Fig. 5. Simulation result of query data for random portion.

수행 하였다. 다음의 그림 5는 이러한 실험 결과를 나타낸다. 그림에서 보듯이 전체적으로 약 80% 이상의 평균 분류 성공률을 보이고 있지만 예상한 바와 같이 오디오의 맨 처음 구간이나 맨 끝 구간 등 아주 극단적인 질의 오디오 구간에서의 성공률은 60% 정도로 저조하게 나타났다.

V. 결론 및 향후 연구과제

본 논문에서는 디지털 신호처리 기법을 이용하여 Classic, Hiphop, Jazz, Rock, Speech 등 5개의 오디오 장르를 자동적으로 분류하는 내용기반 오디오 장르 분류기를 제안하였다. 입력 질의 오디오로부터 총 54차에 해당하는 특징 벡터를 추출하였으며 SFS 기법을 이용하여 최적화 된 10차의 특징벡터만을 구해서 오디오 장르 분류에 사용하였다. 실험 결과 약 90% 가까운 분류 성공률을 보이고 있어 기존 연구에 비하여 약 10%~20% 정도의 성능 향상을 꾀 할 수 있었다. 한편 사용자들의 실제 시스템 사용 환경을 감안하여 질의 오디오의 임의 구간에서 특징 벡터를 추출하여 본 제안 알고리즘의 성능을 측정한 결과 질의 오디오 파일의 맨 처음 구간과 끝 구간에서의 극단적인 경우를 제외하고는 전체적으로 약 80% 이상의 안정적인 분류 성공률을 보이고 있다.

향후 연구 과제로는 본 논문의 실험 결과를 토대로 마이크로폰을 통한 오디오 질의 입력, 잡음 및 반향 제거 등 다양한 방식의 질의 오디오에 대한 연구가 필요하며, 아울러 오디오 파일의 처음과 끝 등 worst-case 질의 입력에 대해서도 안정적인 성능을 발휘할 수 있는 강인한 분류기에 대한 연구를 계획하고 있다. 또한, 보다 심도 깊은 분석을 통하여 현재 20초인 질의 입력 길이를 최소화하는 연구를 진행할 것이다.

참 고 문 헌

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrie -val of audio", IEEE Multimedia, 3(2), 1996.

[2] T. Zhang and C. -C. Jay Kuo, "Hierarchical System for Content-based Audio Classification and Retrieval", Proceedings of SPIE's Conference on Multimedia Storage and Archiving Systems III, SPIE Vol.3527, pp. 398-409, Boston, Nov. 1998.

[3] G. Tzanetakis and P. Cook. "Multifeature audio segmentation for browsing and annotation", In Proc. Workshop on applications of signal proce -ssing to audio and acoustics(WASPAA), New Paltz, NY, 1999. IEEE.

[4] G. Tzanetakis and P. Cook, "Musical Genre Classification of audio Signals", IEEE Transac -tions on Speech and Audio Processing, 2002.

[5] A. Ghias, J. Logan, D. Chamberlin, and B. Smith, "Query by Humming: Musical Information Retrie -val in an Audio Database", ACM Multimedia, pp. 213-236, 1995.

[6] M. Melucci and N. Orio, "Musical Information Retrieval using Melodic Surface", Proceedings of the fourth ACM conference on Digital libraries, pp. 152-160, August 1999.

[7] R. J. McNab, L. Smith, I. H. Witten, C. L. Henderson, "Tune Retrieval in the Multimedia Library", Multimedia Tools and Applications, vol.10, pp. 113-132, 2000.

[8] Lutz Prechelt and Rainer Typke, "An Interface for Melody Input", ACM Transactions on Compu ter-Human Interaction, Vol. 8, No.2, pp. 133-149, June 2001.

[9] S. R. Subramanya, A. Youssef, B. Narahari, and R. Simha, "Automated Classification of Audio Data and Retrieval Based on Audio Classes", International Conference on Computers and Their Applications(ISCA), Cancun, Mexico, April 1999.

[10] J. M. Gray. An Exploration of Musical Timbre. PhD thesis, Dept. of Psychology, Stanford Uni -versity, 1975.

[11] M. J. Carey, E. S. Parris, and H. Lloyd -Thomas, "A comparison of features for speech, music discrimination", In Proc. ICASSP, pp. 1432-1436, March 1999.

[12] J. Makhoul, "Linear prediction: A tutorial over -view", Proceedings of the IEEE, Apr. 1975.

[13] M. Slaney, "A critique of pure audition", Compu -tational Auditory Scene Analysis, 1997.

저 자 소 개

윤 원 중(학생회원)

2003년 상명대학교 정보통신학과 학사 졸업.

2003년~현재 단국대학교 컴퓨터과학 및 통계학과 석사과정

<주관심분야: 음성 및 음향신호처리, 멀티미디어 신호처리, DSP 시스템 구현>

이 강 규(학생회원)

2003년 상명대학교 정보통신학과 학사 졸업.

2003년~현재 단국대학교 컴퓨터과학 및 통계학과 석사과정

<주관심분야: 음성 및 음향신호처리, 멀티미디어 신호처리, DSP 시스템 구현>

박 규 식(정회원)

1986년 Polytechnic University 전자공학과 학사 졸업.

1988년 Polytechnic University 전자공학과 석사 졸업.

1993년 Polytechnic University 전자공학과 박사 졸업.

1994년~1996년 삼성전자 마이크로사업부, 선임 연구원

1996년~2001년 상명대학교 컴퓨터·정보통신공학부 조교수

2001년~현재 단국대학교 정보컴퓨터학부 부교수

<주관심분야: 음성 및 음향신호처리, 멀티미디어 신호처리, DSP 시스템 구현>