ETRI

전자통신동향분석 제24권 제6호 2009년 12월

콘텐츠 분류를 위한 오디오 신호 특징 추출 기술

The Technology of the Audio Feature Extraction for Classifying Contents

임재덕 (J.D. Lim) 지식정보보호연구팀 선임연구원
한승완 (S.W. Han) 지식정보보호연구팀 선임연구원
최병철 (B.C. Choi) 지식정보보호연구팀 선임연구원
정병호 (B.H. Chung) 지식정보보호연구팀 팀장

목 차

음성을 비롯하여 음악, 음향 등을 포함하는 오디오 신호는 멀티미디어 콘텐츠를 구성하는 매우 중요한 미디어 타입이며, 미디어 기록 매체와 네트워크의 발전으로 인한 데이터 양의 급격한 증대는 수동적 관리의 어려움을 유발하게 되고, 이로 인해 오디오 신호를 자동으로 구분하는 기술은 매우 중요한 기술로 인식되고 있다. 다양한 오디오 신호를 분류하기 위한 오디오 신호의 특징을 추출하는 기술은 많은 연구들을 통해 발전하여 왔으며, 본 논문은 오디오 콘텐츠 자동 분류에서 높은 성능을 갖는 오디오 신호 특징 추출에 대해서 분석한다. 그리고 특징 분류기 중에서 안정적인 성능을 가지는 SVM을 사용한 오디오 신호 분류 방법을 알아본다.

Ⅰ. 서론

미디어 기록 매체의 발전과 네트워크의 발전은 오디오 데이터를 포함하여 멀티미디어 콘텐츠 데이터의 생성과 배포에 많은 영향을 미치고 있다. 물론 일부 매니아 층에서도 가능했지만 대부분 해당 분야의 전문가들만의 전유물이었던 멀티미디어 콘텐츠 생성 및 배포 과정이 일반 사용자들도 손쉽게 접근 가능하게 된 것이다. 이러한 환경의 변화는 사용자에게 매우 많은 양의 멀티미디어 콘텐츠를 제공하고 있으며, 이런 대규모의 멀티미디어 콘텐츠는 사용자가 자신이 원하는 콘텐츠를 직접 찾는 것을 매우 어렵게 하고 있다. 비단 사용자 관점에서뿐만 아니라 관리자의 입장에서도 폭발적으로 증가하는 콘텐츠의 양은 기존의 수동적으로 분류 및 관리하던 콘텐츠 관리 방법의 변화를 요구하고 있다. 콘텐츠의 양이 많아질수록 콘텐츠마다 콘텐츠를 설명하는 메타데이터를 통해 관리하는 수동적인 방법은 많은 인적, 시간적 비용을 요구하고 있어 매우 비효율적이기 때문이다. 이를 개선하기 위해 메타데이터가 아닌 멀티미디어 콘텐츠 내용을 기반으로 해당 콘텐츠의 특성을 파악하여 자동으로 분류 및 검색을 할 수 있는 연구가 많이 진행되고 있다[1]-[8]. 멀티미디어 콘텐츠의 자동 분류 기술에 접근하는 방법은 크게 텍스트 기반 특징, 오디오 기반 특징, 시각적 기반 특징, 그리고 이런 특징들의 몇몇 조합에 기반한 특징들을 융합한 특징을 이용하는 방법으로 나누어진다[9],[10].

오디오 기반의 특징만으로 멀티미디어 콘텐츠 내용을 이해하고 분류하는 것은 비효율적이어서, 오디오 정보와 시각적 정보를 모두 이용하는 것이 일반적이다. 하지만, 오디오 기반의 특징에 의한 콘텐츠 분류는 시각적 기반의 특징보다 상대적으로 계산적인 비용이 매우 적으며, 오디오 기반의 특징 공간이 다른 특징 공간보다 상대적으로 작아 특징을 저장하기 위한 공간도 더 적게 요구되기 때문에 시각적 정보를 이용하여 좀 더 상세한 분석이 이루어지기 전의 전처리 단계에서 효과적으로 사용될 수 있다[1],[9].

본 논문은 멀티미디어 콘텐츠 혹은 오디오 데이터를 자동으로 분류하고 검색하기 위해 사용된 콘텐츠 내용 기반의 오디오 관련 특징 및 특징 추출 기술들에 대한 최근 동향을 분석하고, 이들 오디오 특징들의 자동분류를 위해 오디오 특징들을 학습하고 분류하는 데 사용되는 SVM을 적용한 멀티클래스 기반의 분류 방법에 대해 정리한다.

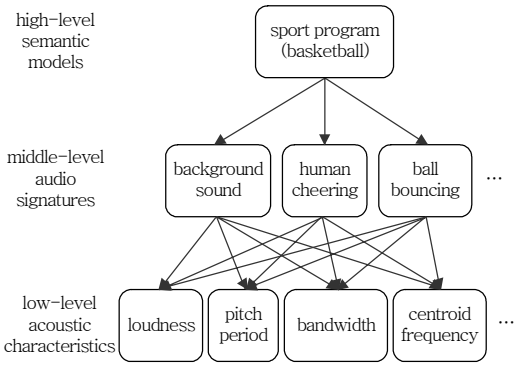
● 용 어 해 설 ●

특징: 임의의 대상을 표현하는 공통적인 성질을 의미하며, 같은 부류의 대상일 경우에는 그 성질이 유사하고 다른 부류의 대상일 경우에는 그 성질이 상이한 것이 좋은 특징이다.

Ⅱ. 오디오 특징 추출 기술

1. 기본적인 오디오 특징

많은 오디오 기반의 특징들은 소리에 대한 인간의 인지능력에 가깝도록 선택되며, 오디오 신호들은 저수준의 음향적 특징(acoustic features) 계층, 서로 다른 신호 객체를 구분할 수 있는 중간수준의 오디오 신호(audio signatures) 계층, 그리고 서로 다른 영역(class)을 구분할 수 있는 고수준의 의미적 모델(semantic model) 계층에 포함된 특징들을 통해 이해될 수 있다[1]. (그림 1)은 이런 오디오 특징 계층의 예를 보여준다. (그림 1)의 예는 농구에 대



(그림 1) 오디오 신호의 특징 계층의 예

한 스포츠 프로그램 콘텐츠를 분류하는 데 다양한 계층의 오디오 신호 특징이 사용되는 예를 보여주는 것으로, 고수준의 의미론적 모델은, 즉 분류하고자 하는 콘텐츠는 농구 프로그램이 된다. 농구 프로그램이라는 고수준의 의미론적 모델은 다음 하위 계층인 중간 수준의 다양한 오디오 객체로 분류될 수 있으며, 예시에는 농구 경기에서의 배경 효과음, 응원 소리, 농구공이 튀는 소리 등을 중간 수준의 오디오 객체의 특징으로 하고 있다. 중간 계층의 각 오디오 객체는 해당 객체와 관련된 특징을 결정지을 수 있는 다음 하위 계층인 저수준의 음향적 특징들을 통해 구성될 수 있으며, 오디오 신호의 에너지, 피치 주기, 대역폭, 주된 주파수 성분 등을 통해 구성할 수 있다.

최종적인 오디오 콘텐츠 분류를 위한 오디오 데이터의 내용은 고수준의 의미론적 모델에 가깝지만, 결국 저수준의 음향적 특징으로 구성됨을 알 수 있다. 물론 오디오 콘텐츠의 특성에 따라 최적화된 저수준의 오디오 특징이 달라질 수 있지만, 대체적으로 일반화된 저수준의 오디오 특징들을 기반으로 하고 있으며 본 절에서는 이들 기본적인 저수준의 오디오 특징들을 살펴보고자 한다.

저수준의 오디오 특징들은 크게 시간 영역과 주파수 영역의 특징들로 나누어지며 시간 영역보다는 주파수 영역에서의 특징들이 더 많이 사용된다. 각 영역에서 주로 많이 사용되는 특징들은 다음과 같다.

가. Time-domain Features

오디오 신호 에너지에 대한 RMS는 소음에 대한 인간의 인지 정도 혹은 소리의 볼륨을 유사하게 나타낼 수 있으며, n번째 프레임에 대한 볼륨은 식 (1)과 같이 표현된다[3].

$$v[n] = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} x_n{}^2(i)} \qquad (1)$$

$x_n(i)$는 n번째 프레임의 i번째 샘플을 나타내며, N은 프레임의 길이이다. 이산 신호 $x(i)$에 대한 에너지는 오디오 신호에 대한 정보 측면에서 그리 중요하지 않지만, 오디오 신호 크기의 변화, 즉 에너지의 변화는 유용한 정보가 될 수 있다. STE는 신호 크기의 변화를 표현하는 방법으로 서로 다른 신호의 분류에 사용되어 왔으며, 임의의 프레임에서의 STE는 (2)와 같이 구해진다.

$$STE = \sum_{i=0}^{N-1} x_n{}^2(i) \qquad (2)$$

ZCR은 서로 다른 오디오 신호를 특징짓는 데 유용하며 특히 음성 신호와 음악 신호를 분류하는 알고리듬에 많이 사용되어 왔다. ZCR은 프레임 내에서 신호의 부호가 변하는 횟수를 의미하므로 일반적으로 고주파는 높은 ZCR을 보이며, 저주파는 낮은 ZCR을 보인다. 임의의 프레임에서의 ZCR은 (3)과 같이 구해진다.

$$ZCR = \frac{1}{2(N-1)}\sum_{m=1}^{N-1}\left|sign[x(m+1)] - sign[x(m)]\right| \qquad (3)$$

N의 길이를 가지는 시간 영역의 신호인 $x(m)$, $(0 \le m \le N-1)$에 대해 $sign[x(m)]$는 $x(m) \ge 0$ 때에는 $+1$, 그 외에는 -1의 값을 가지는 부호 함수이다. 음성 신호는 일반적으로 음악 신호보다 더 높은 ZCR을 가진다. ZCR은 볼륨과 더불어 묵음 구간(silent frame) 검출에 많이 사용된다. 예를 들어 볼륨과 ZCR 모두 일정 기준치 이하의 값을 가진다면 묵음 구간으로 정의할 수 있다. 볼륨과 함께 ZCR을 사용할 때의 장점은 낮은 에너지를 가지는 무성음이 묵음 구간으로 분류되는 것을 방지해 준다[1].

볼륨과 ZCR을 통해 검출되는 묵음 구간을 통하여 임의의 오디오 클립 내에서의 묵음비율(silent ratio)을 특징으로 할 수도 있다. 이는 다양한 장르의 콘텐츠를 구분할 수 있는 특징으로 사용될 수 있으며, 예를 들어 일반적으로 음성은 음악보다 묵음 비율이 더 높고, 뉴스 같은 콘텐츠는 광고물 같은 콘텐츠 보다 묵음비율이 더 높다.

음성 신호와 음악 신호의 구분 성능을 높이기 위해 STE와 ZCR을 개선한 연구도 있다[4]. [4]에서는 STE와 ZCR의 값 자체보다 STE의 변화값 그리고, ZCR의 변화값이 더 좋은 특징으로 사용될 수 있

음을 보여주고 있다. 이들은 각각 LSTER과 HZCRR로 정의되어 사용되었다.

LSTER은 1초 동안의 윈도 내에서 평균 STE의 절반보다 낮은 STE를 가지는 프레임 수의 비율로 정의하고 있으며 (4)와 같이 구해진다.

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[sign(0.5avSTE - STE(n)) + 1] \quad (4)$$

N은 전체 프레임의 개수, n은 프레임 인덱스이고, $STE(n)$은 n번째 프레임에서의 STE이고, $avSTE$는 1초 길이의 윈도 내에서의 평균 STE이다. 일반적으로 묵음 구간은 음악 신호보다 음성 신호에서 더 많이 존재하므로 음성 신호의 LSTER은 음악 신호의 LSTER 보다 높게 나타난다.

HZCRR은 1초 동안의 윈도 내에서 평균 ZCR의 1.5배 보다 높은 ZCR을 가지는 프레임 수의 비율로 정의하고 있으며 (5)와 같이 구해진다.

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sign(ZCR(n) - 1.5avZCR) + 1] \quad (5)$$

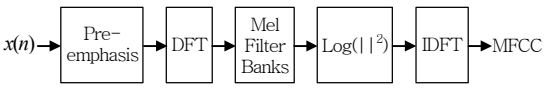
N은 전체 프레임의 개수, n은 프레임 인덱스이고, $ZCR(n)$은 n번째 프레임에서의 ZCR이고, $avZCR$는 1초 길이의 윈도 내에서의 평균 ZCR이다. 일반적으로 음성 신호는 음절 내에서 유성음과 무성음의 상호교차로 구성되어 있으나, 음악 신호는 그렇지 못하기 때문에 음성 신호의 HZCRR이 음악 신호의 HZCRR 보다 높게 나타난다.

나. Frequency-domain Features

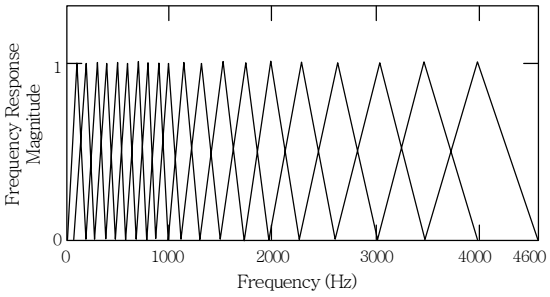
신호에 대한 주파수 영역의 성분은 시간 영역의 성분보다 많은 정보를 제공할 수 있어 시간 영역 기반의 특징보다 주파수 영역 기반의 특징이 상대적으로 많이 연구되었다. 특히 오디오 신호에 있어 주파수 성분의 특징은 인간의 청각을 모델링하여 이를 특징 추출 알고리듬으로 사용하는데, 대표적인 것이 MFCC이다[5]-[7].

MFCC는 음성 인식 영역뿐만 아니라 오디오 콘텐츠 분류 등 오디오 신호 처리에 기본적으로 적용되는 특징 중의 하나이다. MFCC는 인간의 청각 인지 시스템을 모델링한 것으로 (그림 2)와 같은 과정을 거쳐 구해진다[11].

일반적으로 입력된 신호는 고주파 성분의 포만트(formant)가 저주파 성분의 포만트보다 상대적으로 크기가 작아 pre-emphasis 과정을 거쳐, 전 대역의 포만트에 대해 고른 에너지 분포를 가지도록 한다. Pre-emphasis 과정을 거친 신호는 음의 특성이 변하지 않는 안정적이라고 가정할 수 있는 매우 짧은 시간 동안의 프레임 단위로 분석이 이루어지며, 주로 Hamming 윈도나 Hanning 윈도를 통해 프레임이 추출된다. MFCC는 추출된 프레임 단위로 (그림 2)와 같은 과정을 통해 구해진다. 입력된 신호는 DFT를 통해 주파수 대역으로 변환되고, 변환된 신호는 (그림 3)과 같은 Mel Filter Bank를 통과한다. Mel Filter Bank는 인간의 청각 구조에 근접한 여러 개의 band-pass filter들의 filter bank로 구성되어 있으며, 각 filter의 중심 주파수 배열은 인간의 주파수 지각 단위인 Mel 단위를 기반으로 1kHz 이하에서는



(그림 2) MFCC 처리 과정



(그림 3) Mel Filter Bank

● 용 어 해 설 ●

포만트(formant): 복잡한 소리일 경우라도 분석해보면 단순한 음파의 조합으로 나타낼 수 있으며, 음성 신호 중 특정 부분이 강화되는 부분을 가리킨다.

켑스트럼(Cepstrum): 시간 영역 함수의 스펙트럼을 다시 한번 푸리에 변환한 2차 주파수 스펙트럼을 가리키며, 단구간 스펙트럼에 대한 로그 스케일의 크기를 푸리에 역변환한 것을 가리키기도 한다.

균일하게 구성되어 있고 1kHz 이상에서는 로그 스케일 단위로 구성되어 있다. Mel Filter Bank의 처리는 입력 신호를 사람의 청각 시스템에서 인지하는 스펙트럼 신호와 유사하도록 해준다. Mel Filter Bank를 통과한 신호는 logarithm 처리 과정을 거치는데, logarithm dynamic 압축 과정을 통해 신호의 크기 성분은 살리고 중요성이 떨어지는 위상 성분은 버리게 된다. 따라서 로그 계산 후 얻어지는 신호는 동적 측면에서의 변화, 즉 음성일 경우 화자의 입과 입력 장치(마이크 등) 간의 거리 변화에 덜 민감하게 된다. Logarithm 처리 과정을 통한 신호는 IDFT를 거쳐 MFCC를 구하며, k번째 Mel Filter Bank를 통과한 신호를 $S[k]$라고 할 경우, (6)과 같이 표현된다. M은 필터 뱅크 개수이고, L은 MFCC 차수이다.

$$C[n]=\sum_{k=1}^{M}\log(S[k])\cos[(k-0.5)\frac{n\pi}{M}],\ n=1,\ldots,L \quad (6)$$

IDFT 연산은 logarithm 처리 과정으로 구해진 신호가 실수이고, 대칭적이므로 DCT 연산으로 처리가 가능하다.

MFCC 외에도 주파수 특성과 관련된 많은 인지적 특징들(perceptual features)이 있으며, 전체 파워 스펙트럼(total power spectrum), 서브밴드 파워(subband powers), 중심주파수(frequency centroid), 대역폭(bandwidth), 피치주파수(기본 주파수) 등이 그것이며, 내용 기반의 오디오 콘텐츠 분류에 많이 사용되고 있다[5]-[7].

전체 파워 스펙트럼(P)은 프레임 구간 내의 전체 스펙트럼 에너지 분포에 대한 값이며 (7)과 같이 구해진다. f_o는 $f_s/2$ (f_s: 샘플링 주파수)이다.

$$P=\log(\int_0^{f_o}|F(f)|^2df) \quad (7)$$

서브밴드 파워(P_i)는 [0, $1/8f_o$], [$1/8f_o$, $1/4f_o$], [$1/4f_o$, $1/2f_o$], [$1/2f_o$, f_o]와 같은 보통 4개의 서브밴드 구간에서의 스펙트럼 에너지 분포값이다. 그 구간은 오디오 콘텐츠의 주파수 성분 특성에 따라 적절히 나누어 적용할 수 있다. 상한 $H(i)$와 하한 $L(i)$의 영역을 가지는 i번째 서브 밴드 파워는 (8)과 같이 구해진다.

$$P_i=\log(\int_{L(i)}^{H(i)}|F(f)|^2df) \quad (8)$$

중심주파수(F_C)는 스펙트럼 에너지 분포의 중간 지점을 의미하며, 어떤 주파수 성분에서 에너지가 집중되었는지를 나타낸다. 또한 소리의 선명도(brightness)를 표현하는 수단으로 사용될 수 있으며 (9)와 같이 구해진다[12]. p_i는 임의의 주파수 f_i와 관련된 파워 값이다.

$$F_C=\frac{\int_0^{f_0}f_i\times p_i df}{\int_0^{f_0}p_i df} \quad (9)$$

대역폭(F_B)은 파워 스펙트럼의 모양이 중심 주파수 근처에 집중되어 있는지 혹은 전체 스펙트럼에 걸쳐 퍼져 있는지를 나타내며, (10)과 같이 구해진다.

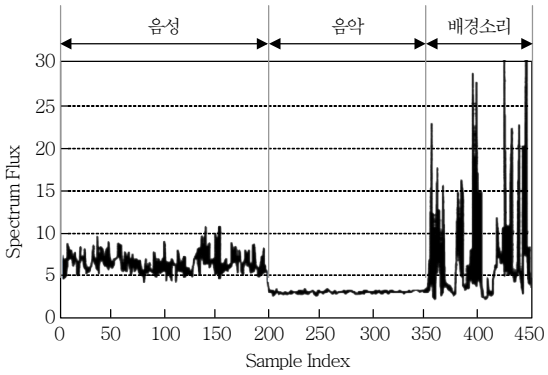
$$F_B=\sqrt{\frac{\int_0^{f_0}(f_i-F_C)^2\,p_i df}{\int_0^{f_0}p_i df}} \quad (10)$$

피치주파수는 정규화된 autocorrelation 함수의 최고점을 검출하는 간단한 방법으로 얻을 수 있으며, 최고점이 임의의 임계치([5]에서는 실험적으로 0.65 값을 사용하였다.) 이상일 경우 피치주파수를 구하며, 그렇지 않을 경우 non-pitched 프레임으로 식별한다.

스펙트럼 플럭스(SF)는 오디오 클립에서 인접한 두 프레임 사이의 스펙트럼의 변화를 보여주는 특징으로, 1초 길이의 오디오 클립 내의 인접한 두 프레임 사이의 스펙트럼에 대한 평균 변화값으로 정의하며 (11)과 같이 구해진다[4],[8].

$$SF=\frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\int_0^{f0}[\log|F_n(f)|+\delta)-\log|F_{n-1}(f)|+\delta)]^2 \quad (11)$$

δ은 계산적인 오버플로를 피하기 위해 사용하는 매우 작은 수이고, N은 오디오 클립 내의 전체 프레임의 수를 나타내며, K는 DFT 차수를 나타낸다. 스펙트럼 플럭스는 정의에서 알 수 있듯이 프레임 기반의 특징이 아닌 다수의 프레임으로 구성된 클립 기반의 특징이다. (그림 4)는 음성, 음악 그리고 배



(그림 4) 음성, 음악, 배경 소리로 구성된 오디오클립에서의 스펙트럼 플럭스 커브

경 소리에 해당하는 오디오 신호에 대한 스펙트럼 플럭스 특징을 보여주는 것으로 그림에서 보는 것과 같이 음성, 음악 그리고 배경 소리의 구분에 좋은 성능을 보임을 알 수 있다.

2. 콘텐츠 분류를 위한 오디오 특징 추출에 대한 연구

본 장에서는 콘텐츠 분류를 위한 기본적인 오디오 신호 기반의 특징 이외에 몇몇 특정 분야의 연구 내용을 정리하면서 각 연구 분야에 맞게 사용된 특징을 중심으로 정리한다.

우선 TV 프로그램의 종류들이 다양해짐에 따라 오디오 정보를 이용하여 TV 프로그램의 종류를 구분하려는 연구들이 있었다[1],[2]. [1]에서는 4개의 볼륨 기반의 특징, 3개의 피치 기반의 특징, 5개의 주파수 기반의 특징들을 통해 총 12개의 특징 집합을 구성하여 광고, 농구게임, 축구게임, 뉴스보도, 일기예보 등 5개 종류의 TV 프로그램 분류에 대한 시도를 하였다. 볼륨 기반의 특징은 오디오 샘플 클립 내 볼륨의 표준편차(VSTD), 오디오 클립 내 최대 볼륨값($max(v)$)과 최소 볼륨값($min(v)$)으로부터 $(max(v)-min(v))/max(v)$를 통해 구해지는 볼륨 동적 범위(VDR), 비묵음비율(NSR), 4Hz 부근에서의 볼륨 파형에 대한 주파수 성분(FCVC4) 등이 사용되었다. 오디오 파형의 기본적인 주기인 피치 기반의 특징은 피치에 대한 표준편차(PSTD), 전체 오디오 클립에서의 유성음 혹은 음악 프레임의 비율인 유성음-음악 비율(VMR), 그리고, 전체 오디오 클립에서의 잡음 혹은 무성음 비율인 잡음-무성음 비율(NUR) 등이 사용되었다. 주파수 기반의 특징은 중심주파수(FC), 대역폭(BW), 0~630Hz 영역에서의 에너지 비율(ERSB1), 630~1720Hz 영역에서의 에너지 비율(ERSB2), 1720~4400Hz 영역에서의 에너지 비율(ERSB3) 등이 사용되었다. 각각의 특징 값은 프로그램 종류에 따라 다르게 나타나는데, VSTD, VDR, FC, ERSB3는 광고, 농구/축구게임, 뉴스보도/일기예보 등에서는 매우 다르게 나타나고, PSTD, VMR, NUR는 뉴스보도와 일기예보 사이에서 다르게 나타나며, FC와 BW는 농구게임과 축구게임에서 다르게 나타난다. 일반적으로 일기예보에서의 발음 속도는 뉴스보도에서의 발음 속도보다 빠르며, 일기예보에서는 주로 한 사람이 보도를 하고 뉴스에서는 여러 명이 보도를 한다. 또한 일기예보는 대체로 끊김 없이 보도가 진행되는데 반해 뉴스 보도는 보도 중간에 끊김이 존재한다. 농구게임과 축구게임은 배경 소리가 시끄럽지만, 하나는 실내 게임이고 하나는 실외 게임이라는 차이로 서로 다른 주파수 구조를 가지고 있으며, 농구게임에서의 오디오 신호는 선수의 신발과 바닥 사이의 마찰로 인한 많은 고주파 성분이 존재한다. 해당 논문에서는 TV 프로그램 분류를 위해서 OCON 구조의 신경망(neural network)을 사용하였으며, 평균 82.5%의 분류 정확도를 보여주고 있다. [2]에서는 [1]에서 사용된 특징에 ZCR에 대한 변화율인 ZCR 표준편차(ZSTD), 볼륨 파형 내에서의 인접한 peak값과 valley값의 차이의 누적 합인 볼륨 파동(VU)이 추가된 14개의 특징 집합을 통해 [1]의 TV 프로그램 분류를 시도하였다. TV 프로그램 분류를 위해 HMM 분류 기법을 사용하였고, 가장 좋은 성능을 보이는 5개의 상태와 128개의 심볼에서 평균 93.4%의 분류 성능을 보여주었다.

[6]의 연구에서는 전체 파워 스펙트럼, 4개 구간의 서브밴드 파워, 중심주파수, 대역폭, 그리고 피치 주파수로 이루어진 인지적 특징과 MFCC를 사용하

여 다양한 특징 집합을 구성하여 오디오 신호의 분류 및 검색에 사용하였다. 특징 집합은 8개의 인지적 특징의 평균값과 표준편차로 구성된 16차원 특징 벡터에 묵음 비율(묵음 프레임의 수/전체 프레임의 수)과 피치 비율(피치가 검출된 프레임의 수/전체 프레임의 수)을 더한 18차원의 특징 벡터에 대해 정규화된 '*Perc*'로 명명된 특징 집합과 L차 MFCC들의 평균과 표준 편차로 구성된 '*CepsL*'로 명명된 2L 차원의 켑스트럼 특징 집합들의 조합으로 구성된다. 이들 특징조합은 *Perc* 집합 단독, *CepsL*(L = 5, 8, 10, 15, 20, 40, 60, 80, 100, 120) 그리고, 둘 사이의 조합인 *PercCepsL*을 통해 오디오 콘텐츠의 분류 및 검색 연구가 진행되었고, 시험에 사용된 오디오 샘플은 altotrombone(13), animals(9), bells(7), cello-bowed(47), crowds(4), female(35), laughter(7), machines(11), male(17), oboe(32), percussion (99), telephone(17), tubularbells(19), violin-bowed (45), violinpizz(40), water(7) 등 16 종류, 409개의 샘플로 구성되어 있으며, 학습을 위해 211개의 샘플이, 시험을 위해 198개의 샘플이 사용되었다. 분류 시험은 위의 특징 집합의 조합에 대해 각각 SVM, NN, 5-NN, NC 등의 분류 방법으로 이루어졌으며, *PercCeps5* 특징 조합으로 SVM 분류 방법을 사용했을 때, 8.08~11.00%의 오류율로 가장 좋은 성능을 보여 주었다.

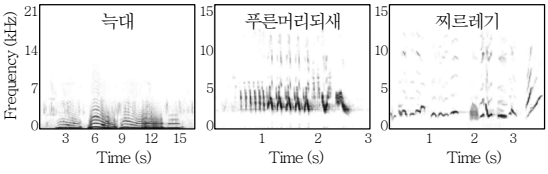
[7]에서는 오디오 특징 집합을 개별 특징의 평균과 표준편차를 통해 파일 단위로 가지는 [5]의 연구와 달리 프레임 단위의 특징 집합을 통해 오디오 신호에 대한 분류를 시도하였다. [7]에서는 기본적인 MFCC와 전체 파워 스펙트럼, 4개 구간의 서브밴드 파워, 중심주파수, 대역폭, 그리고 피치 주파수를 특징으로 하고, 여기에 Mel-Frequency ICA 기반의 특징을 추가로 사용하고 있다. Mel-Frequency ICA 기반의 특징은 MFCC 처리 과정 중 Mel Filter Bank 결과값에 대한 로그 파워값을 DCT 변환 대신 ICA에 의해 변환시킨 값으로 (12)와 같이 표현된다. ICA에 의해 변환된 값들은 이론적으로 최대의 통계적인 독립성을 가진다.

$$\mathbf{f}_{ICA} = \sum_{k=1}^{M} \log(S[k])\mathbf{w}_k \qquad (12)$$

$\mathbf{w}_k$ 는 ICA 변환에 사용되는 일련의 과정을 통해 얻어지는 변환 기저(transform basis)이고, $S[k]$는 Mel Filter Bank를 통해 얻어진 값이며, M은 Mel Filter Bank 개수이다. 시험에 사용된 오디오 샘플은 male speech(50), female speech(50), cough(50), laughing(49), screaming(26), dog barking(50), cat mewing(45), frog wailing(50), piano(40), glass breaking(34), gun shooting(33), knocking(50) motorcycle(50), doorbell(50), telephone(50) 등 15종류, 677개의 샘플로 구성되어 있고, 학습과 시험에 각각 절반씩 사용되었다. 본 연구에서 제시한 특징 집합과 프레임 기반의 SVM 분류 방법을 통해 정확도(정상 분류 파일 수/전체 시험 파일 수)는 91.7% 정도 보이고 있다.

[13]은 스펙트로그램(spectrogram) 상에서 특정한 패턴을 가지는 동물들의 소리를 분석하고, 이를 고려한 각 동물들의 특징 추출 방법과 효율적인 검색 기법을 제안하였다. 해당 연구에서는 특정 동물들의 소리에 대한 스펙트럼 분포가 curve-like한 형태를 보이며 시각적인 차이가 분명하다는 것에 착안하여 스펙트럼 분포의 패턴에 대해 이미지 처리 기법을 적용하여 특징 클래스를 구성한다. (그림 5)는 본 연구에 사용한 몇몇 동물 소리의 예로 스펙트럼 분포가 각 동물마다 매우 뚜렷하다는 특징을 볼 수 있으며 특히 새 종류에서 더욱 뚜렷한 curve-like한 특징을 볼 수 있다.

본 연구에서 제안한 특징은 다음과 같이 구해진다. 시간 축과 주파수 축에서의 스펙트로그램의 변화율을 고려한 eigen analysis를 수행하고 변화율이 급격한 부분을 eigenvalue의 크기를 통해 POI로

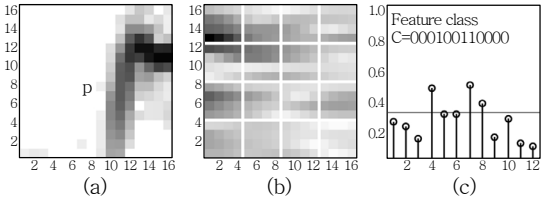


(그림 5) Curve-like한 특징을 보이는 동물 소리의 예

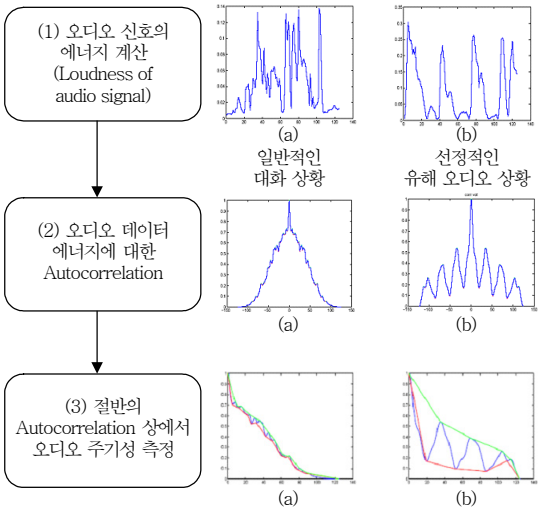
규정하여 선택한다. 다음 단계는 (그림 6)과 같이 수행된다. 선택된 POI의 주변 픽셀(a)들을 고려한 2차원 DFT를 수행하고, DFT 결과(b)를 가로와 세로 각각 4개의 서브밴드로 나누어 인접한 주파수와의 내적을 통해 12bit의 값을 가지는 특징 클래스를 구성한다(c).

이렇게 구성된 12bit의 특징 클래스는 적은 용량과 빠른 검색 속도를 보장할 수 있지만, 스펙트로그램의 패턴에 매우 의존적이므로 패턴이 일정한 형태를 보이는 신호가 아닌 경우에는 그리 좋은 성능을 보이지 않는다. 실제 시험 결과에서도 다른 종류의 동물 소리보다 curve-like 특성이 상대적으로 더 뚜렷한 새 종류의 동물 소리에서 더 좋은 성능을 보이고 있다.

오디오 특징을 사용하여 비디오 영상물에서의 유해 콘텐츠 검출을 시도한 연구 결과도 있으며, 이 연구에서의 유해 콘텐츠는 음란 콘텐츠를 가리킨다



(그림 6) p 픽셀 주변으로부터의 특징 클래스 구성



(그림 7) 오디오 신호의 주기성 측정 방법

[10]. [10]에서는 현재까지 대부분의 유해 콘텐츠 검출 연구가 주로 이미지 및 텍스트 기반 검출 기술에 집중되어 있지만, 비디오 영상물에는 오디오와 모션 정보와 같은 보조적인 미디어 특징이 있음을 지적하며, 유해 콘텐츠 검출의 보조적인 기술로 오디오 스트림의 주기성에 기반한 유해 콘텐츠 검출 기법을 소개하였다. 오디오 스트림의 주기성을 측정하기 위한 과정은 (그림 7)과 같으며, 다음과 같이 진행된다.

(1) 오디오 신호의 에너지를 계산

5초 길이의 오디오 클립에 대해 0.04초 길이의 윈도 크기 단위로 125 레벨로 소리의 크기가 계산된다. 일반적인 상황에서의 에너지 분포(a)와 유해한 상황에서의 에너지 분포(b)는 (그림 7)의 (1) 과정의 옆 그래프와 같다. 유해한 상황에서 좀 더 주기성(periodicity)이 강하게 나타남을 알 수 있다.

(2) Autocorrelation 수행

계산된 각 에너지 분포에 대해 autocorrelation을 수행한다. 신호의 주기성은 주로 autocorrelation, circular correlation 혹은 periodogram에 의해 분석될 수 있으며, 여기서는 autocorrelation 기법을 통해 주기성을 측정한다[14].

(3) 주기성 측정

Autocorrelation의 절반 부분에서 최대값(peak values)과 최소값(valley values)의 차이를 계산함으로써 주기성을 계산하고, 유해 콘텐츠와 일반 콘텐츠를 구분할 수 있는 경계값을 통해 유해 콘텐츠를 구분한다.

[10]에서는 일반 콘텐츠와 유해 콘텐츠의 구분이 가능한 주기성 경계값으로 '4'를 채택하였으며, 시험에 사용된 비디오 콘텐츠(When Harry met Sally)에서는 거의 정확하게 동작하고 있음을 실험 결과로 보여주고 있다. 해당 논문은 오디오 신호가 음성 신호인지 아닌지를 구분하지 않고 주기성에 의

〈표 1〉 오디오 신호의 특징을 기반으로 한 콘텐츠 분류 응용 예제

종류	관련 연구	목적	특징 집합	분류기	성능
TV 프로그램 분류	Liu et al[1]	5가지 TV 프로그램 분류: 광고, 농구게임, 축구게임, 뉴스 보도, 일기 예보	- Vol. contour based: VSTD, VDR, NSR, FCVC4 - pitch contour based: PSTD, VMR, NUR - freq. based: FC, BW, ERSB1, ERSB2, ERSB3	OCON 구조의 Neural Network	분류 정확도: 평균 82.5%
	Liu et al[2]		[1] + ZSTD, VU	HMM	분류 정확도: 평균 93.4%
다양한 오디오 클립 분류	G.Guo et al [6]	다양한 오디오 클립(남성, 여성, 악기, 총, 웃음 등)에 대한 분류 및 검색 - 16종류, 409개 클립[6] - 15종류, 677개 클립[7]	Total power of spectrum, Sub-bands powers of spectrum, FC, BW, pitch frequency, MFCCs	Binary tree based multiclass SVM	Lowest error rate: 약 11.00%
	J.Wang et al [7]		[6] + Mel Frequency-based ICA	Frame-based multiclass SVM	분류 정확도: 약 91.7%
새소리 검색	Rolf Bardeli et al[13]	Curve-like한 스펙트럼 패턴을 가지는 동물. 특히 새소리 검색 (264종, 1000개의 샘플)	Eigen 분석을 통해 얻어진 POI로부터 계산된 12bit class	HMM	검색 질의에 대한 랭킹 스코어 확인: curve-like한 패턴이 그렇지 않은 패턴보다 검색 성능이 좋음
유해 콘텐츠 검출	N. Rea et al [14]	오디오 에너지의 주기성에 기반한 유해 장면 검출	에너지에 대한 autocorrelation 과정을 통해 얻어진 periodicity value	Periodicity value에 대해 정해진 한계값(4) 이상의 값을 가지는 콘텐츠를 유해로 판정	무해 콘텐츠에 대한 False alarm rate: 2% 이하. 유해 콘텐츠에 대한 검출률: 62%

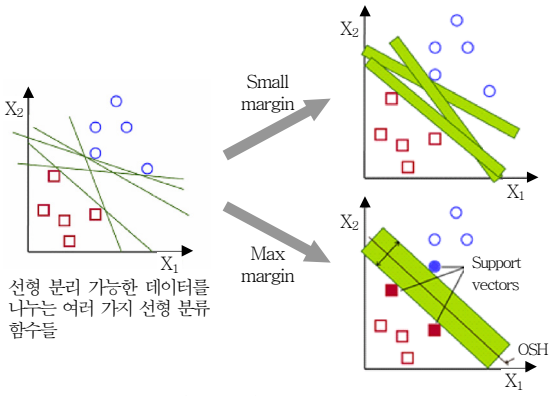
존해 해당 신호를 검출할 수 있는 장점이 있는 반면, 주기적인 반복이 있는 테니스나 탁구 같은 스포츠에 기인한 오디오 신호에 대해서는 그다지 좋은 성능을 보여주지 못하고 있다. 이는 주기성에만 의존하여 해당 오디오 신호를 검출하기 때문이다.

유해 콘텐츠 검출을 위해서 음성 인식 기술이 적용될 수도 있지만 음성 인식의 한계, 즉 언어적 의존성과 유해 콘텐츠 검출에 사용될 음성 언어 DB 관리, 그리고 상대적으로 많은 계산 비용 등의 약점으로 실용화하기는 어려운 점이 있다.

<표 1>은 오디오 특징을 기반으로 한 콘텐츠 분류 응용에 대한 내용을 정리한 것이다.

Ⅲ. 오디오 분류 기술

오디오 신호를 표현하는 특징들은 다양한 분류기를 통해 기계 학습 과정을 거쳐 해당 클래스들로 분류된다. 오디오 신호의 분류를 위해 사용되는 분류기의 종류는 ANN, HMM, GMM, SVM 등 매우 다양하지만, 본 논문에서는 데이터 분류에 있어 유용



(그림 8) SVM 개념

한 기술로 다양한 패턴 인식 응용 분야에 사용되는 SVM 분류 기술을 통한 멀티클래스 기반의 분류 방법에 대해 분석한다[6]-[8].

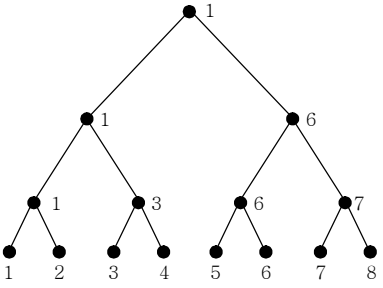
SVM은 최대 마진으로 두 개의 클래스를 구분하는 초평면(hyperplane)을 구성하는 이진 분류기의 한 종류로, 선형 구분자를 통해 데이터를 구분하기 힘들 때 해당 데이터를 보다 높은 차원의 공간으로 사상시켜 선형적인 초평면을 통해 데이터를 구분할 수 있게 한다. (그림 8)은 SVM의 개념을 보여주며, 그림과 같이 두 클래스 사이의 간격을 최대화시키는,

즉 최대 마진을 가지는 경계 상의 데이터를 support vector라고 하고, 이들 경계의 중심에 위치한 구분자를 최적 분류 초평면(OSH)이라고 한다[15].

멀티 클래스의 분류는 이진 클래스 SVM을 이용하여 이루어질 수 있으며, 보통 두 가지 구조를 이용한다. 하나는 one-against-all 구조로 각 클래스와 나머지 클래스를 구분하는 방법이고, 다른 하나는 one-against-one 구조로 클래스별 구분을 수행하는 방법이다. One-against-one 방법은 최종 결정을 위해 조합 가능한 모든 클래스 쌍에 대해 이진 분류 방법이 필요하며, 일반적으로 보팅(voting) 구조를 이용한다. 보팅 구조는 c개의 클래스가 있을 경우 $c(c-1)/2$번의 분류 과정이 필요하며, c가 클수록 매우 많은 계산상의 부하를 유발하여 처리 성능을 떨어뜨릴 수 있다.

상향식 이진 트리(bottom-up binary tree) 기반의 멀티 클래스 분류 방법은 계산상의 부하를 줄일 수 있다[6]. (그림 9)는 상향식 이진 트리 기반의 멀티클래스 분류 방법을 보여준다. 그림 상의 숫자는 클래스 식별자이며, 아무런 의미가 없다.

분류는 상향식으로 이루어지므로 가장 아래쪽부터 분류 과정이 진행되어 위로 올라간다. 이 때 아래 부분에서의 분류 과정에서 분류가 된 클래스가 다음 분류를 위한 클래스로 결정되고, 결정된 클래스들 간에 다시 분류 과정을 거쳐 최종적인 클래스가 결정되는 방식이다. 이 방식에서는 c개의 클래스가 있을 경우 $(c-1)$번의 분류 과정만이 필요하다는 것이다. 하지만, 분류 모델을 생성하는 학습 단계에서는 이진 클래스 구분 단계에서 어떤 클래스들 간의 분류가 이루어질지 또한 상위 단계로 어떤 클래스들이 선택되어 전달될지 알 수 없기 때문에 보팅 구조와 같이 모든 클래스 쌍의 경우에 대해 학습이 필요하며, $c(c-1)/2$개의 분류 모델을 생성해야 한다.

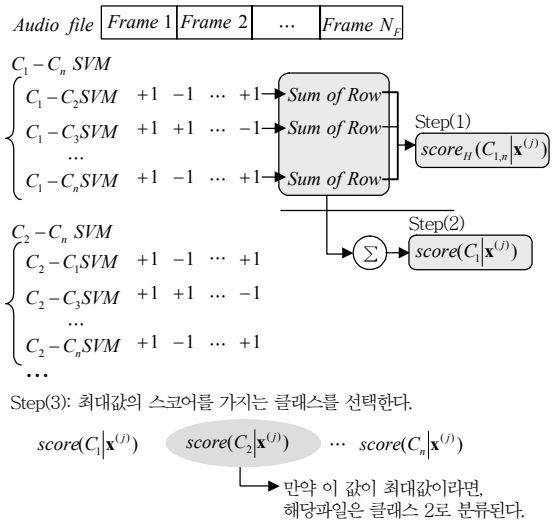


(그림 9) 8개의 클래스를 분류하기 위한 이진 트리 구조

특징 벡터가 오디오 파일 단위로 생성되어 학습 및 분류에 사용되어지기도 하지만, 세분화된 단위인 프레임 단위로 생성되어 학습되고 분류에 사용되는 경우도 있다[7]. 즉, 프레임 기반의 멀티클래스 분류 방법으로, 오디오 파일 분류시 파일 내의 모든 프레임별로 분류 결과를 도출하여 이들의 조합으로 오디오 파일의 분류를 결정하는 방법이다. 오디오 파일 $\mathbf{x}^{(j)},(j=1,...,N_F)$는 N_F개의 프레임으로 구성되어 있고, 각 오디오 파일은 $C_m(m\in\{1,2,...,M\})$ 클래스에 속한다고 할 경우, 다음과 같은 과정으로 최종 클래스가 결정된다.

(1) 클래스 C_m과 그 외의 모든 클래스 $C_n(n\neq m)$에 대해 C_m-C_n 이진 클래스 SVM을 통해 다음과 같은 스코어를 구한다. H(•)는 Heaviside step 함수로 unit step 함수라고도 한다. $(\mathbf{wx}^{(j)}+b)$는 SVM 결정함수로 {-1, 1}의 값을 반환한다. 즉, 모든 프레임에서 m 클래스와 관련된 분류 결과를 얻어낸다.

$$score_H(C_{m,n}|\mathbf{x}^{(j)}) = \sum_{j=1}^{N_F} H(\mathbf{wx}^{(j)}+b) - \sum_{j=1}^{N_F} H(-(\mathbf{wx}^{(j)}+b)) \quad (13)$$



(그림 10) SVM을 이용한 프레임 기반의 멀티클래스 분류 방법

(2) 클래스와 관련된 분류 결과를 합함으로써, m 클래스의 스코어를 계산한다.

$$score(C_m|\mathbf{x}^{(j)}) = \sum_n score(C_{m,n}|\mathbf{x}^{(j)}) \quad (14)$$

(3) 계산된 각 클래스의 스코어 중 가장 큰 값을 가지는 클래스를 해당 오디오 파일의 클래스로 분류한다. 즉, m^*가 최종 분류된 클래스가 된다.

$$m^* = \arg\max_m score(C_m|\mathbf{x}^{(j)}) \quad (15)$$

이 과정을 그림으로 표현하면 (그림 10)과 같다.

Ⅳ. 결론

폭발적으로 증가하는 대규모 멀티미디어 콘텐츠는 수동적인 관리 형태에서 벗어나 자동적인 관리 구조를 요구하고 있다. 멀티미디어 콘텐츠에 대한 자동적인 관리를 위해서는 사람이 제공하는 콘텐츠의 메타 정보에 의존하기에는 한계가 있다. 따라서 콘텐츠 내용을 기반으로 자동으로 분류 및 검색이 가능하도록 해야 한다.

본 논문은 멀티미디어 콘텐츠 혹은 오디오 데이터를 자동으로 분류하고 검색하기 위해 사용된 오디오 콘텐츠 내용 기반의 오디오 특징 기술들에 대한 최근 동향을 분석하고, 또한 패턴 분류에 많이 사용되는 SVM 기반의 멀티클래스 분류 방법을 살펴보았다. 분석 결과 오디오 신호의 분류를 위해 사용되는 특징은 기본적인 특징들의 다양한 조합을 통해 특정 영역의 목적에 맞게 최적화된 특징 조합을 찾아가는 방법과 오디오 신호의 특성을 분석하여 특성에 의존적인 최적화된 특징을 개발하는 방법을 통해 구성되었다.

오디오 신호의 특징은 크게 시간 영역의 특징과 주파수 영역의 특징으로 구분되며, 대체로 이 두 영역의 특징의 조합으로 오디오 콘텐츠의 특징이 표현되었다. 또한 오디오 신호 분류의 목적에 따라 신호에 대한 패턴 자체가 특징으로 사용되는 경우도 있었다. 분석 결과 오디오 콘텐츠는 시간 영역 및 주파수 영역의 기본적인 특징들의 조합으로도 어느 정도 분류가 될 수 있지만, 최적화된 성능을 위해서는 분류 목적에 해당하는 오디오 콘텐츠 자체에 대한 분석을 통한 최적의 특징을 찾아내는 것이 필요하다.

약어 정리

ANN	Artificial Neural Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
FCVC4	Frequency Component of the Volume Contour around 4Hz
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HZCRR	High Zero Crossing Rate Ratio
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
LSTER	Low Short Time Energy Ratio
MFCC	Mel-Frequency Cepstrum Coefficient
NN	Nearest Neighbor
NSR	Non-Silence Ratio
NUR	Noise-or-Unvoice Ratio
OCON	One-Class-One-Network
OSH	Optimized Separating Hyperplane
POI	Point Of Interest
PSTD	Pitch Standard Deviation
RMS	Root Mean Square
SF	Spectrum Flux
STE	Short Time Energy
SVM	Support Vector Machine
VDR	Volume Dynamic Range
VMR	Voice-or-Music Ratio
VSTD	Volume Standard Deviation
VU	Volume Undulation
ZCR	Zero Crossing Rate
ZSTD	ZCR Standard Deviation

참 고 문 헌

[1] Z. Liu, Y. Wang, and T. Chen, “Audio Feature Extraction and Analysis for Scene Segmentation and Classification,” *Journal of VLSI Signal Process Systems*, Vol.20, No.1/2, 1998,

pp.61-79.

[2] Z. Liu, J. Huang, and Y. Wang, "Classification of TV Programs Based on Audio Information Using Hidden Markov Model," *in Proc. IEEE Signal Process. Soc. Workshop Multimedia Signal Process.*, 1998, pp.27-32.

[3] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio," *IEEE MultiMedia*, Vol.3, No.3, 1996. pp.27-36.

[4] L. Lu, H.J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. on Speech and Audio Proc.*, Vol.10, No.7, Oct. 2002, pp.504-516.

[5] S.Z. Li, "Content-based Audio Classification and Retrieval Using the Nearest Feature Line Method," *IEEE Trans. on Speech and Audio Proc.*, Vol.8, No.5, Sep. 2000, pp.619-625.

[6] G. Guo and S.Z. Li, "Content-based Audio Classification and Retrieval by Support Vector Machine," *IEEE Trans. on Neural Networks*, Vol.14, No.1, Jan. 2003, pp.209-215.

[7] J.C. Wang, J.F. Wang, C.B. Lin, K.T. Jian, and W.H. Kuok, "Content-based Audio Classification Using Support Vector Machines and Independent Component Analysis," *18th Int'l Conf. on Pattern Recognition(ICPR'06)*, Hong Kong, Aug. 2006.

[8] L. Lu, H.-J. Zhang, and S.Z. Li, "Content based Audio Classification and Segmentation by Using Support Vector Machines," *ACM Multimedia Systems Journal*, Vol.8, No.6, Mar. 2003, pp.482-492.

[9] D. Brezeale and D.J. Cook, "Automatic Video Classification: A Survey of the Literature," *IEEE Trans. on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol.38, No.3, May 2008, pp.416-430.

[10] N. Rea, G. Lacey, C. Lambe, and R. Dahyot, "Multimodal Periodicity Analysis for Illicit Content Detection in Video," *The 3rd European Conf. on Visual Media Production(IET CVMP 2006)*, London, Nov. 2006.

[11] 이건상, 양성일, 권영헌, 음성인식(Speech Recognition), 한양대학교 출판부, 2001, pp.29-37.

[12] U. Srinivasan, S. Pfeiffer, S. Nepal, M. Lee, L. Gu, and S. Barrass, "A Survey of MPEG-1 Audio, Video and Semantic Analysis Techniques," *Multimedia Tools Appl.*, Vol.27, No.1, 2005, pp.105-141.

[13] Rolf Bardeli, "Similarity Search in Animal Sound Databases," *IEEE Trans. on Multimedia*, Vol.11, No.1, Jan. 2009, pp.68-76.

[14] M. Vlachos, P. Yu, and V. Castelli, "On Periodicity Detection and Structural Periodic Similarity," *In SIAM Int'l Conf. on Data Mining*, 2005.

[15] 한학용, "패턴인식 개론: MATLAB 실습을 통한 입체적 학습," 한빛미디어, 2005, pp.518-522.