오디오 신호에 기반한 음란 동영상 판별*

김봉완(SiTEC), 최대림(SiTEC), 이용주(원광대)

<차 례>

<Abstract>

Classification of Phornographic Videos Based on the Audio Information

Bong-Wan Kim, Dae-Lim Choi, Yong-Ju Lee

As the Internet becomes prevalent in our lives, harmful contents, such as phornographic videos, have been increasing on the Internet, which has become a very serious problem. To prevent such an event, there are many filtering systems mainly based on the keyword- or image-based methods.

The main purpose of this paper is to devise a system that classifies pornographic videos based on the audio information. We use the mel-cepstrum modulation energy (MCME) which is a modulation energy calculated on the time trajectory of the mel-frequency cepstral coefficients (MFCC) as well as the MFCC as the feature vector. For the classifier, we use the well-known Gaussian mixture model (GMM).

The experimental results showd that the proposed system effectively classified 98.3% of pornographic data and 99.8% of non-pornographic data. We expect the proposed method can be applied to the more accurate classification system which uses both video and audio information.

* Keywords: Phornographic video classification, Harmful video, MCME, Modulation energy.

* 이 논문은 2006년도 원광대학교의 교비 지원에 의해서 수행되었음.

1. 서 론

최근 대량의 멀티미디어 자료들이 인터넷을 통해 공개 및 유통되면서, 청소년 등이 접근할 수 있는 인터넷 공간에 음란 동영상이 무방비 상태로 노출되는 사례가 증가하고 있다. 2007년 3월에만도 인터넷 포털 사이트 및 사용자제작콘텐츠(User Created Content, UCC) 게시 사이트에 음란 동영상이 게재되어 일반인들에게 공개됨으로써 사회적 파장을 불러일으킨 사례가 언론을 통해 공개된 경우만 2건이 발생한 바 있다[1][2].

따라서 이러한 사례를 방지하기 위한 기술이 지속적으로 개발되어오고 있으며 대표적인 방법들로는 파일이름, 제목 및 본문 내용 등의 키워드를 이용한 검출 방법과 동영상에 포함된 이미지를 분석하여 음란성 여부를 판단하여 검출하는 방법 등이 있다[3]-[5]. 그러나 키워드 기반의 필터링 방법의 경우 관련 키워드를 고의로 회피하여 게시할 경우 이를 방지할 수 없다. 이미지 기반의 필터링 방법의 경우 일반적으로 좋은 성능을 보이고 있으나 배경 및 조명 상태, 인종에 따른 다양한 피부색으로 인한 신체 영역 검출오류로 인하여 판별 성능이 저하되는 문제점이 있다.

따라서, 음란 동영상에서 이미지 이외의 중요한 정보를 포함하고 있는 오디오 신호를 이용하여 음란성 동영상 검출하기 위한 방법에 관한 연구가 필요하다. 관련 사전 연구로는 음란 유해사이트의 차단을 위하여 음란 동영상으로 부터 표준적 음향 신호를 미리 선정해 두고, 판별하고자 하는 음향 신호가 입력되었을 때 각 표준 음향 신호와의 상관계수를 구하고 그 값이 임계치를 넘는지에 따라 판별하는 방법이 제안된 바 있다[6]. 그러나 제안된 방법은 표준적 음향 신호의 개수 및 선정 방법에 따라 판별 성능이 좌우될 수 있으며, 표준적 음향 신호의 개수가 증가함에 따라 상관계수 계산을 위한 시간이 증가하는 단점을 갖고 있다. 또한 다양한 길이의 오디오 신호를 처리하기 위한 방안도 미흡하다고 볼 수 있다.

따라서 본 논문에서는 음성/비음성 판별, 오디오 유형 분류 및 화자 인식 등에서 자주 사용되고 있는 통계적 음향 신호처리 방법을 이용하여 음란 동영상을 판별하는 방법을 제안하고자 한다. 제안된 오디오 기반 음란성 동영상 판별 방법은 오디오 신호만을 이용한 음란 동영상 기반 판별 기술뿐만 아니라, 이미지 기반 판별 기술과 함께 사용됨으로써 판별 성능을 향상시키는 데에도 활용될 수 있으리라 사료된다.

본 논문의 구성은 다음과 같다. 제 2장에서는 음란 동영상에 포함된 오디오 신호의 특징에 대하여 기술하고, 제 3장에서는 본 논문에서 오디오 기반 동영상 판별을 위해 사용하고 있는 특징인 MCME에 대하여 기술한다. 제 4장에서는 실험을 위해 사용된 데이터베이스에 대하여 기술하며, 제 5장에서 실험 결과에 대하여 기술한다. 제 6장에서 최종적으로 결론을 맺는다.

2. 음란 동영상에 포함된 오디오 신호의 특징

음란 동영상에 포함된 오디오의 주요 특징으로는 교성, 신음 소리, 거친 호흡음 및 접촉음 등이 주된 내용을 이루며 특정 구간내에서 이러한 소리들이 일정한 주기를 가지고 반복적으로 나타난다는 점을 들 수 있다. 아울러 흥미를 끌기 위한 부가적 요소로 배경 음악, 시나리오에 따른 배우들간의 음성 대화, 주변 환경 소음 등이 포함된 경우가 많다.

특히 최근 음악 관련 멀티미디어 콘텐츠가 증가함에 따라, 분석 대상 오디오 신호가 음성 위주의 신호인지, 음악 위주의 신호인지 아니면 음란성있는 요소를 주된 내용으로 하는 신호인지 판별하는 것은 매우 중요하다고 할 수 있다. 이를 위해 본 논문에서는 음란 오디오를 위한 음란 모델과, 안티 모델로서 음성 위주의 일반 모델 및 음악 모델을 사용하여 판별하고자 한다.

음성 신호, 음악 신호 및 음란 동영상에 포함된 오디오 신호의 특징을 살펴보기 위해 다음 <그림 1>에 10초 분량의 음악 방송의 아나운서 멘트, 음란 동영상에 포함된 오디오 신호 및 가수의 노래가 포함된 락음악에 대하여 파형과 스펙트로그램을 나타내었다.

(a) 음악 방송 아나운서 멘트

(b) 음란 동영상 오디오 신호

(c) 가수의 노래가 포함된 락음악

<그림 1> 음성 신호, 음란 동영상 오디오 신호 및 음악 신호의 파형과 스펙트로그램

위의 그림에서 보는 바와 같이 음성의 경우 자음과 모음의 연속적 발성으로 인해 스펙트럼 포락선의 변화가 다른 신호에 비해 빠르다는 것을 알 수 있다. 락 음악의 경우 비록 빠른 음악임에도 불구하고 스펙트럼 포락선의 변화 속도는 음성에 비해 빠르지 않은 것을 볼 수 있다. 음란 오디오 신호의 경우 유사한 음향적 특징이 일정한 주기로 매우 분명하게 반복되고 있음을 볼 수 있다. 음란 오디오 신호의 스펙트럼 포락선의 변화 속도가 음성에 비해 느린 것은 조음 기관의 움직임 속도와 신체의 주기적 움직임 속도의 차이에 기인한 것으로 판단된다.

따라서 오디오 신호를 이용하여 음란 동영상 여부를 판별하기 위해서는 교성, 신음 소리 및 접촉음 등의 주요 요소의 음향적 특징을 반영하기 위한 MFCC 이외에, 주기성 및 변화의 빠르기에 대한 특성을 반영한 특징이 추가되어야 한다고 생각된다. 이를 위해 본 논문에서는 모듈레이션 분석 결과를 판별을 위한 특징으로 사용하고자 한다. 모듈레이션 분석이란 단구간에서 추출한 음향 특징이 시간에 따라 얼마나 빠르게 변화하는 지를 DFT 등의 주파수 분석을 통하여 측정하는 것이다.

3. MCME

오디오 신호의 n번째 프레임으로부터 구한 1차 DFT를 $X[n,k]$라고 하면, 다음 식 (1)과 같이 k번째 계수의 프레임별 결과에 대하여 2차 DFT를 취함으로써 크기 모듈레이션 스펙트럼(Magnitude Modulation Spectrum, MMS)을 얻을 수 있다.

$$MMS[n,k,q] = \sum_{p=0}^{P-1} |X[n+p,k]| e^{-j2\pi qp/P} \tag{1}$$

여기에서 n은 프레임 인덱스, k는 1차 DFT 결과의 주파수축 인덱스, q는 2차 DFT의 주파수축 인덱스, P는 2차 DFT의 포인트 수이다.

모듈레이션 주파수가 낮다는 것은 시간에 따른 스펙트럼의 변화가 느림을 의미하며, 높다는 것은 스펙트럼의 변화가 빠르게 나타난다는 것을 의미한다. 따라서 음성의 경우 유성음과 무성음의 연이은 발음으로 인해 스펙트럼의 변화가 자주 발생하는데 비해 음악의 경우 스펙트럼의 변화가 크지 않은 특징이 있으므로 모듈레이션 스펙트럼은 음성 및 음악의 판별을 위한 특징으로 사용된다.

음성 및 음악 판별을 위해 모듈레이션 주파수 분석을 수행할 때, 대부분 1차 DFT의 결과를 그대로 이용하지 않고, 청각 특성을 반영한 멜(Mel) 주파수 대역으로 분할하는 필터 뱅크 분석 결과를 이용하여 2차 DFT 분석을 수행한다. 또한 음성의 경우 음절 발음의 영향으로 인해 약 4 Hz에서 모듈레이션 에너지(Modulation

Energy, ME)의 피크가 발생한다는 연구 결과[7]에 따라, 모듈레이션 에너지 분석에서 모든 주파수를 사용하지 않고 4 Hz (또는 4~8 Hz)의 모듈레이션 에너지만을 계산하여 이를 음성 및 음악 판별을 위한 특징으로 사용하여 왔다[8]-[10]. 이처럼 특정 모듈레이션 주파수 정보만을 사용할 때에는 2차 분석에서 계산량 절감을 위해 모든 주파수에 대하여 DFT 분석을 수행하지 않고 중심 주파수가 4 Hz인 대역필터(bandpass filter)를 이용하여 에너지를 계산한다. 특징의 차수를 줄이기 위해 필터 뱅크의 각 채널의 4 Hz 모듈레이션 에너지를 합산하여 1차의 특징을 추출하고, 정규화(normalization)를 수행한 후 이를 음성 및 음악 판별을 위한 특징으로 사용한다. 그러나 전통적인 ME 분석 방법의 경우 각 채널별 강한 상관을 갖고 있는 스펙트럼을 기반으로 하여 계산함으로써 그 성능이 음성/음악 판별을 위한 켑스트럼 기반의 다른 특징들보다 좋은 편은 아니다.

켑스트럼 기반의 모듈레이션 정보를 이용한 방법으로는 Tyagi 등이 잡음 환경에서의 음성 인식 성능 향상을 위하여 제안한 MCMS(Mel-frequency Cepstrum Modulation Spectrum)가 있다[11][12]. Tyagi 등은 MFCC 영역에서 구한 MCMS를 MFCC의 다이나믹 특징(dynamic feature)으로 사용할 경우, 차분 파라메터 및 RASTA PLP와 비교하여 가산 잡음 환경에서 음성 인식 시스템의 성능을 향상시킬 수 있음을 보였다. MCMS의 정의는 다음 식과 같다.

$$MCMS[n,l,q]=\sum_{p=0}^{P-1} C[n+p,l]e^{-j2\pi pq/P} \tag{2}$$

여기에서 n은 프레임 인덱스, l는 MFCC 계수 인덱스, q는 모듈레이션 주파수 인덱스를, P는 주파수 분석을 위한 DFT 포인트 수를, 그리고 C는 MFCC 계수를 의미한다. MFCC 계수들이 필터 뱅크 계수들보다 서로 상관이 적다는 것을 감안할 때, MCMS가 MMS보다 음성 인식에 있어 더 좋은 성능을 보일 것으로 예상할 수 있다. 즉 MMS가 전체 스펙트럼의 시간에 따른 변화를 모델링하는 데 비하여, MCMS의 경우 스펙트럼 포락선에 영향을 미치는 요소들의 변화만을 모델링한다는 차이가 있다. 그러나 MCMS의 경우에는 MFCC의 각 계수별 스펙트럼을 모두 별도로 취급함으로써 특징 차수가 커지는 단점 (즉, B개의 대역 필터를 통하여 MCMS를 계산할 경우 추출되는 특징 벡터의 차수는 $B \times$ MFCC 차수가 됨)을 갖고 있다.

따라서 우리들은 MCMS의 장점을 취하면서 특징의 차수를 줄이기 위해 다음 식과 같이 멜-켑스트럼 모듈레이션 에너지(Mel-Cepstrum Modulation Energy, MCME)를 정의하여 이를 음성/음악 판별을 위한 특징으로 사용할 것을 제안한 바 있다[13].

$$MCME[n, q] = \frac{\frac{1}{L}\sum_{l=0}^{L-1}|MCMS[n, l, q]|^2}{\frac{1}{P}\sum_{p=0}^{P-1}\log(E[n+p])} \quad (3)$$

여기에서 E는 오디오 신호의 단구간 에너지를 의미한다. 즉 MCME는 동일한 모듈레이션 주파수에 대하여, 각 MFCC 계수들로 부터 구한 MCMS의 파워에너지를 합한 값이다. 또한 모듈레이션 주파수 분석 구간내의 오디오 신호의 단구간 에너지의 평균으로 나누어 줌으로써, MCME 분석 결과의 변동폭을 줄였다. 우리들은 제안된 MCME를 이용한 음성/음악 판별 실험에서 8 Hz MCME 경우 4 Hz의 ME와 비교하여 71%, cepstral flux와 비교하여 53%의 판별 오류 감소율을 보이는 것을 확인하여 스펙트럼의 포락선의 변화 속도를 반영하는 특징으로 유효함을 보였다[13]. 따라서 본 논문에서는 MCME를 음란 동영상 판별을 위한 특징으로 사용하고자 한다. 다만, 음란 오디오 데이터의 경우 음성/음악 판별과 다소 다른 양상을 보일 수 있으므로 4 Hz 또는 8 Hz의 단일 주파수를 사용하기 보다는 3 Hz ~ 50 Hz 범위의 모듈레이션 주파수에 대하여 MCME를 계산하고 이를 특징 벡터로 이용하고자 한다.

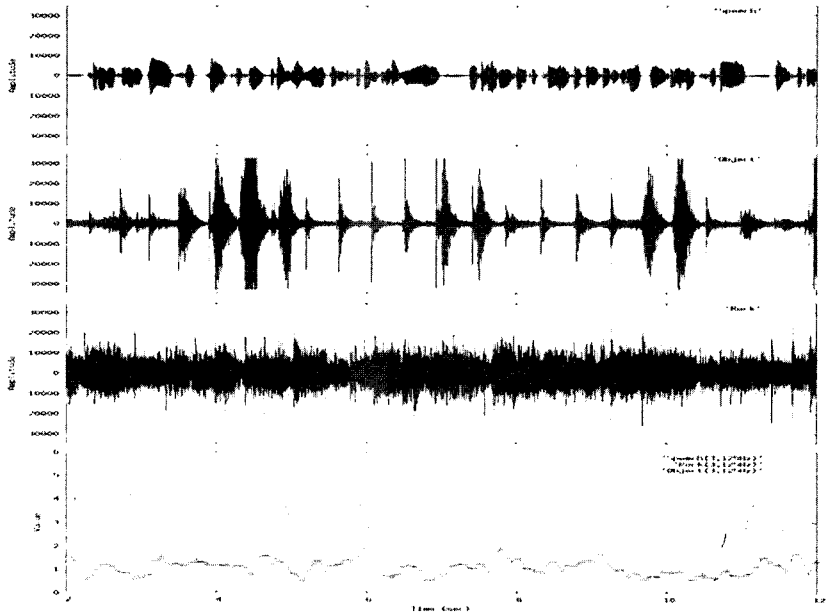
<그림 2>는 MCME 분석의 예를 확인하기 위하여 <그림 1>에 나타낸 각 신호들에 대한 파형과, 각 신호들에서 3.125 Hz의 MCME를 추출한 결과를 나타내었다. 음성의 경우 음악에 비해 빠르게 변화하는 특성으로 인해 높은 에너지 값을 나타내고, 음악의 경우 비교적 빠른 락음악임에도 불구하고 낮은 에너지를 보임을 알 수 있다. 음란성 오디오 신호의 경우 교성 등이 일반적인 음악보다 빠른 주기를 갖고 나타나기는 하지만 음성의 자음, 모음 발성과 같이 빠르게 변화되지는 않으므로 그 중간 정도의 값을 갖는 것을 볼 수 있다.

4. 데이터베이스

4.1 데이터베이스

음란 오디오 모델 및 안티 모델들의 학습 및 평가를 위해 데이터베이스를 구성하였다. 음악 데이터의 경우 다양한 장르의 음악을 반영하기 위하여 RWCP Genre music DB[14]의 내용을 포함하였으며, 자주 들을 수 있는 유명곡들을 포함시키기 위하여 RWCP Popular music DB[14]와 인터넷에서 구한 음악파일을 추가하였다.

일반 동영상 자료로는 2005년 10월 1일부터 2005년 10월 25일까지 25일분의



<그림 2> 음성(첫 번째 창), 음란 오디오(두 번째 창), 록 음악(세 번째 창)에 대한 파형 및 이들로부터 추출한 3.125 Hz MCME(네 번째 창 - 첫 번째 선이 음성, 두 번째 선이 음란 오디오, 세 번째 진한 선이 록 음악으로 부터 추출한 MCME)

KBS 뉴스 동영상과; 인터넷에서 구한 26개의 다큐멘터리, 인터넷에서 구한 33개의 영화 동영상이 포함되었다. 영화 동영상의 경우 영화 1편이 2개의 CD 분량으로 나누어져 있는 경우 그 중 1개의 CD 분량만 포함하도록 하였다. 스포츠 데이터의 경우 음악 및 일반 데이터와 그 음향적 특성이 현저히 다르리라고 예상되어 별도의 유형으로 모델링되어야 된다고 판단되지만, 자료 입수의 한계로 인하여 일반 동영상 데이터에 포함되었다. 음란 동영상의 경우 인터넷에서 149개의 파일을 입수하여 구성하였다.

4.2 학습 세트, 평가 세트의 구성 및 전처리

음악 모델의 경우 다양한 장르의 음악 유형을 반영하기 위해 Genre DB의 33개의 서브 장르에서 임의로 1곡씩을 고르고, 인터넷 음악 데이터에서 54개의 파일을 임의로 골라 학습 데이터로 사용하고 나머지는 평가 데이터로 사용하였다.

나머지 데이터의 경우 각 유형별로 임의로 20%의 데이터를 선정하고 이를 학습 데이터로, 나머지 80%를 평가 데이터로 사용하였다. 학습 데이터의 총량은 135 파일(31시간 분량)이며 평가 데이터의 총량은 546 파일(142시간 분량)이다. 이와 같은 실험용 데이터베이스의 내용, 학습 및 평가용 세트의 구성에 대한 사항은

<표 1>에 정리되어 있다.

오디오 기반 판별을 위한 전처리 과정으로서 모든 동영상 및 오디오 데이터로부터 11 kHz, 16 비트 linear PCM 포맷으로 오디오 신호를 추출하였다.

<표 1> 실험 데이터베이스, 학습 및 평가 세트의 구성

구분	유 형	학습용		평가용		계	
		파일 수	시간	파일 수	시간	파일 수	시간
음악	RWCP Genre 음악 DB	33	2.4	69	4.6	102	7.0
	RWCP Popular 음악 DB	-	-	100	6.8	100	6.8
	인터넷에서 구한 음악	54	4.6	181	15.3	235	19.9
	계	87	7.0	350	26.6	437	33.7
일반	KBS 뉴스 (25일분)	5	4.2	20	17.8	25	22.0
	다큐멘터리	5	3.8	21	16.6	26	20.4
	영화	6	7.1	27	37.9	33	45.0
	스포츠	2	0.8	9	5.3	11	6.1
	계	18	15.9	77	77.6	95	93.5
음란	인터넷에서 구한 음란 동영상	30	7.9	119	38.4	149	46.3
총 계		135	30.8	546	142.7	681	173.5

5. 판별 실험 및 결과

5.1 실험 환경 및 구성

학습 및 평가를 위해 음란 오디오 데이터에서 음란 부분만을 세그멘테이션하는 작업은 전혀 수행하지 않았으며 전체 파일을 사용하였다. 특징 추출을 위하여 25 ms의 해밍윈도우를 사용하여 10 ms 단위로 프레임을 이동하면서 12차의 MFCC 결과와 단구간 에너지를 추출하였으며, MFCC 추출 시 채널의 영향을 최소화하기 위해 CMN(Cepstral Mean Normalization)이 적용되었다. 따라서 MCME의 추출을 위한 샘플링 주파수는 100 Hz이다. 1차적으로 추출된 MFCC 결과에 대해 32 포인트의 FFT를 프레임 단위로 이동하면서 3.125 Hz ~ 46.88 Hz 범위의 15차의 MCME를 추출하였다.

성능 비교를 위해 <표 2>와 같이 4가지의 경우로 구분하여 특징을 추출하였다. 실험에 사용된 판별기는 GMM을 사용하였으며 학습 데이터로부터 음란 모델과 일반 모델 및 음악 모델의 안티 모델을 학습하였다. 판별을 위한 테스트에서는 파일별로 추출된 특징벡터 열을 이용하여 각 모델별 출현 확률을 구하고, 세 모델 중 음란 모델의 확률값이 가장 높으면 음란으로 판별하고, 일반 또는 음악의 확률값이 높으면 비음란으로 판별하였다.

<표 2> 실험에 사용된 특징

특징	설 명
MFCC	- 12차의 MFCC
MFCC+D+A	- 12차의 MFCC와 차분 및 차차분 파라메터 - 총 36차의 특징 벡터
MCME	- 12차의 MFCC에 대하여 32포인트 FFT를 적용하여 구한 15차의 MCME
MCME+MFCC	- 12차의 MFCC + 15차의 MCME - 총 27차의 특징 벡터

5.2 실험 결과

판별기로 사용된 GMM의 혼합(mixture)의 수를 1부터 64까지 점진적으로 증가시키면서 판별 성능을 검증하였으며, 그 결과를 <표 3>에 정리하였다. 음란 판별 오류율은 비음란으로 판별된 음란 데이터의 개수를 음란 데이터의 개수로 나눈 것이며, 비음란 판별 오류율은 음란 데이터로 판별된 비음란 데이터의 개수를 비음란 데이터의 개수로 나눈 것이다. 전체 판별 오류율은 판별 오류가 발생한 데이터의 개수를 전체 테스트 데이터의 개수로 나눈 것이다.

속도는 Pentium 4 3.0 GHz, 2 GByte의 RAM을 갖는 Windows XP 환경에서, 특징 추출과 판별에 걸린 시간을 테스트 데이터의 전체 시간으로 나눈 것이다. 동영상 데이터에서 오디오 데이터를 추출하는 시간 및 11 kHz로의 다운샘플링 시간은 계산에 포함되지 않았다.

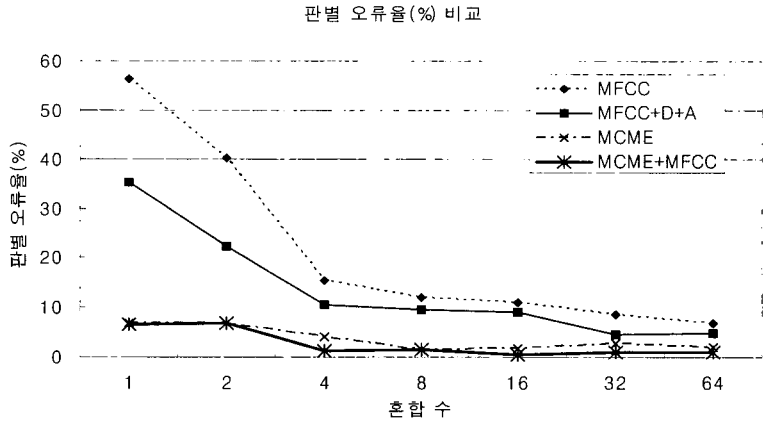
<표 3>에 나타난 바와 같이 MFCC 기반 방법의 경우 다이나믹 특징을 포함한 MFCC+D+A가 좋은 성능을 보이는 것을 볼 수 있다. 또한 MCME를 이용한 경우 더 적은 혼합에서 MFCC 기반의 방법보다 좋은 성능을 보임을 알 수 있다. 16개의 혼합을 사용하는 MCME+MFCC의 경우 32개의 혼합을 사용하는 MFCC+D+A에 비해 판별 오류 감소율 약 86%를 보이며, 특징 차수가 작음으로 인해 속도가 더 빠른 것을 볼 수 있다. 약 142시간 분량의 테스트 데이터에 대하여 특징 추출 및 판별에 1시간 41분이 소요되었다.

<그림 3>은 각 특징별 혼합 수에 따른 전체 판별 오류율 변화를 그림으로 나타내어 본 것이다. <그림 3>에서 볼 수 있는 것처럼 MCME를 이용한 경우 MFCC 정보만을 이용한 경우에 비해 매우 안정적인 판별 성능을 보임을 알 수 있다. 혼합의 수를 증가시킴에 따라 MFCC 기반 특징들의 판별 성능이 향상되고 있으나 MCME를 이용한 특징들에는 미치지 못함을 알 수 있다.

16개의 혼합을 사용하는 MCME+MFCC의 경우 총 546개의 테스트 데이터 중, 단 3개의 데이터에 대해서 판별 오류가 발생하였으며 이 중 음란 데이터를 비음란으로 판별한 경우가 2개, 비음란 데이터 중 음란으로 판별한 경우가 스포츠 데이터 1개이다. 이에 대한 자세한 판별 혼동표를 <표 4>에 나타내었다.

<표 3> 특징별 혼합 수에 따른 판별 오류율 (%) 및 속도 (xRT)

특징	판별오류율 (%) 및 속도 (xRT)	혼합 수						
		1	2	4	8	16	32	64
MFCC	음란 (%)	45.4	9.2	22.7	17.6	16	16	15.1
	비음란 (%)	59.3	48.9	13.3	10.3	9.4	6.3	4.2
	전체 (%)	56.2	40.3	15.4	11.9	10.8	8.4	6.6
	속도 (xRT)	0.007	0.007	0.007	0.007	0.008	0.01	0.013
MFCC +D+A	음란 (%)	9.2	8.4	17.6	8.4	5.9	5	4.2
	비음란 (%)	42.6	26	8.4	9.6	9.8	4.2	4.7
	전체 (%)	35.4	22.2	10.4	9.3	9	4.4	4.6
	속도 (xRT)	0.007	0.007	0.008	0.008	0.01	0.013	0.019
MCME	음란 (%)	15.1	6.7	5	4.2	2.5	2.5	4.2
	비음란 (%)	4.4	6.6	3.5	0.7	1.4	2.8	1.4
	전체 (%)	6.8	6.6	3.9	1.5	1.7	2.8	2.0
	속도 (xRT)	0.007	0.007	0.007	0.008	0.008	0.01	0.013
MCME +MFCC	음란 (%)	10.9	2.5	4.2	4.2	1.7	2.5	4.2
	비음란 (%)	5.2	8	0.5	0.7	0.2	0.5	0.2
	전체 (%)	6.4	6.8	1.3	1.5	0.6	0.9	1.1
	속도 (xRT)	0.007	0.007	0.008	0.008	0.009	0.012	0.016



<그림 3> 각 특징별 혼합 수에 따른 전체 판별 오류율(%) 비교

음악이 일반으로 판별된 경우는 대부분 삼바, 펑크 등 빠르고 흥겨운 노래가 일반으로 판별되었으며, 일반 데이터에서 음악과 음란으로 판별된 경우는 모두 스포츠 영역에서 발생하였다. 스포츠 데이터의 경우 선수들의 호흡음, 충격음 등이 자주 발생하여 그 음향적 특성이 음성이 주도하는 일반 데이터와 다소 다른점을 감안하면, 별도의 유형으로 모델링할 경우 성능을 향상시킬 수 있으리라 기대한다. 음란 데이터가 음악으로 판별된 2개의 경우는 강한 음악배경을 갖고 있으며

그 중 1개는 압도적인 배경 음악 및 잡음으로 음란성 있는 음향이 거의 드러나지 않는 데이터이다. <표 4>에 나타난 바와 같이 제안된 시스템의 음란/일반/음악의 3 유형 판별율은 97.8%에 달하여 3유형 판별에서도 효과가 있음을 알 수 있다.

<표 4> 판별 혼동표 (confusion matrix)

구분	테스트 데이터		판별 결과		
	유형	파일수	음악	일반	음란
음악	RWCP Genre 음악 DB	69	67	2	-
	RWCP Popular 음악 DB	100	98	2	-
	인터넷에서 구한 음악	181	179	2	-
일반	KBS 뉴스	20	-	20	-
	다큐멘터리	21	-	21	-
	영화	27	-	27	-
	스포츠	9	2	6	1
음란	인터넷에서 구한 음란 동영상	119	2	0	117

6. 결 론

본 논문에서는 동영상에서 추출된 오디오 신호를 이용하여 음란 동영상을 검출하는 방법을 제안하였다. MCME와 MFCC를 함께 이용할 경우 음란 데이터 판별율 98.3%, 비음란 데이터 판별율 99.8%를 보임으로써 제안된 방법의 유효함을 알 수 있었다. 제안된 오디오 기반 음란 동영상 검출 방법은 그 단순성으로 인해 이미지 기반의 검출 방법보다 속도의 측면에서 장점이 있으며 이미지 기반 검출 시스템의 취약 부분을 보완하기 위한 수단으로 사용될 수 있으리라 기대한다.

향후 연구 방향으로는 스포츠 영역의 데이터를 별도로 모델링함으로써 판별 성능을 높이고, UCC 등 대량의 비정형화된 데이터를 추가하여 그 성능을 검증하고, 특징 추출 속도를 향상시키기 위한 방안에 대한 연구를 진행하고자 한다. 또한 오디오 신호만으로 판별이 어려운 데이터에 대한 보완 방법을 연구할 필요가 있다고 판단된다.

참 고 문 헌

[1] 연합뉴스(2007. 3. 19), http://www.hani.co.kr/arti/society/society_general/197252.html.

[2] 연합뉴스(2007. 3. 21), http://www.hani.co.kr/arti/society/society_general/197780.html.
[3] M. Hammani, Y. Chahir, L. Chen, "WebGuard: a web filtering engine combining textural, structural, and visual content-based analysis", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 2, pp. 272-284. 2006.
[4] H. Lee, S. Lee, T. Nam, "Implementation of high performance objectionable video classification system", *Proc. ICACT*, pp. 959-962, 2006.
[5] W. Kim, H. Lee, J. Park, K. Yoon, "Multi class adult image classification using neural networks", *Lecture Notes in Artificial Intelligence,* Vol. 3501, pp. 222-226, 2005.
[6] 조동욱, 김지영, "음란 유해사이트 차단을 위한 음향 신호 처리 및 분석", *한국콘텐츠학회 논문지*, 제4권 제2호, pp. 1-6, 2004.
[7] T. Houtgast, H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility", *Acoustica*, Vol. 28, pp. 66-73, 1973.
[8] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", *Proc. ICASSP,* Vol. 2, pp. 1331-1334, 1997.
[9] J. Pinquier, J.-L. Rouas, "A fusion study in speech/music classification", *Proc. ICME,* Vol. 1, pp. 409-412, 2003.
[10] S. Karnebäck, "Discrimination between speech and music based on a low frequency modulation feature", *Proc. Eurospeech*, pp. 1891-1894, 2001.
[11] V. Tyagi, I. McCowan, H. Bourlard, H. Misra, "On factorizing spectral dynamics for robust speech recognition", *Proc. Eurospeech*, pp. 981-984, 2003.
[12] V. Tyagi, I. McCowan, H. Misra, H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR", *Proc. ASRU*, pp. 399-404, 2003.
[13] B. Kim, D. Choi, Y. Lee, "Speech/music discrimination using mel-cepstrum modulation energy", *Lecture Notes in Artificial Intelligence,* Vol. 4629, pp. 406-414, 2007.
[14] M. Goto, "Development of the RWC music database", *Proc. ICA*, Vol. 1, pp. 553-556, 2004.

접수일자: 2007년 8월 14일
게재결정: 2007년 9월 14일

▶ 김봉완(Bong-Wan Kim)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: bwkim@sitec.or.kr

▶ 최대림(Dae-Lim Choi)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: dlchoi@sitec.or.kr

▶ 이용주(Yong-Ju Lee) : 교신저자
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 원광대학교 전기 전자 및 정보공학부, 음성정보기술산업지원센터
전화: 063) 850-7451
E-mail: yjlee@wku.ac.kr