오디오 신호를 이용한 음란 동영상 판별

김 봉완[1], 최 대림[1], 방 만원[2], 이 용주[3]
[1]원광대학교 음성정보기술산업지원센터
[2]목포대학교 전자공학부
[3]원광대학교 전기·전자 및 정보공학부

Classification of Phornographic Videos Using Audio Information

Bong-Wan Kim, Dae-Lim Choi, Man-Won Bang, Yong-Ju Lee
[1]Speech Information Technology & Industry Promotion Center, Wonkwang Univ.
[2]Division of Information Engineering, Mokpo Univ.
[3]Department of Electrical Electronic and Information Engineering, Wonkwang Univ.
E-mail : {bwkim, dlchoi}@sitec.or.kr, bmw@mokpo.ac.kr, yjlee@wonkwang.ac.kr

Abstract

As the Internet is prevalent in our life, harmful contents have been increasing on the Internet, which has become a very serious problem. Among them, pornographic video is harmful as poison to our children. To prevent such an event, there are many filtering systems which are based on the keyword based methods or image based methods.

The main purpose of this paper is to devise a system that classifies the pornographic videos based on the audio information. We use Mel-Cepstrum Modulation Energy (MCME) which is modulation energy calculated on the time trajectory of the Mel-Frequency cepstral coefficients (MFCC) and MFCC as the feature vector and Gaussian Mixture Model (GMM) as the classifier.

With the experiments, the proposed system classified the 97.5% of pornographic data and 99.5% of non-pornographic data. We expect the proposed method can be used as a component of the more accurate classification system which uses video information and audio information simultaneously.

I. 서론

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 손수제작물(User Created Content, UCC) 게시 사이트에 음란 동영상이 게재되어 일반인들에게 공개됨으로써 사회적 파장을 불러 일으킨 사례가 언론을 통해 공개된 경우만 2건이 발생한 바 있다.

따라서 이러한 사례를 방지하기 위한 기술이 지속적으로 개발되어오고 있다. 대표적인 방법들로는 파일이름, 제목 및 본문 내용 등의 키워드를 이용한 검출 방법과 동영상에 포함된 이미지를 분석하여 음란성 여부를 판단하여 검출하는 방법 등이 있다[1, 2, 3]. 그러나 키워드 기반의 필터링 방법의 경우 관련 키워드를 고의로 회피하여 게시할 경우 이를 방지할 수 없다. 또한 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. 오디오 기반 음란성 동영상 판별 방법은 오디오 신호만을 이용한 음란 동영상을 판별 시스템뿐만 아니라, 이미지 기반 판별 시스템과 병렬로 사용됨으로써 판별 성능을 향상시키는데에도 활용될 수 있으리라 사료된다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문에서 오디오 기반 음란 동영상 판별을 위해 사용한 특

징에 대하여 기술하고, 3장에서는 실험을 위해 사용된 데이터베이스에 대하여 기술한다. 4장에서는 실험 결과에 대하여 기술하고 5장에서 결론을 맺는다.

II. 음란 오디오 판별을 위한 특징

음란 동영상에 포함된 오디오의 주요 특징으로는 교성, 신음 소리, 거친 호흡음 및 접촉음 등이 주된 내용을 이루며 일정한 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 컨텐츠가 증가함에따라, 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 일반 모델 및 음악 모델을 사용하여 판별하고자 한다.

오디오 신호를 이용하여 음란 동영상 여부를 판별하기 위해서는 교성, 신음 소리 및 접촉음 등의 주요 요소의 음향적 특징을 반영하기 위해 음성 인식에 자주 사용되는 MFCC외에, 주기성 및 변화의 빠르기에 대한 특성을 반영한 특징이 추가되어야 한다고 생각된다. 이를 위해 본 논문에서는 모듈레이션 분석 결과를 판별을 위한 특징으로 사용한다. 모듈레이션 분석이란 단구간에서 추출한 스펙트럼이 시간에 따라 얼마나 빠르게 변화하는지를 FFT 등의 주파수 분석을 통하여 측정하는 것이다.

모듈레이션 정보와 관련하여 음성 및 음악을 구분짓는 중요한 요소 중 하나로 모듈레이션 에너지(Modulation Energy, ME)를 들 수 있다. 음성의 경우 자음과 모음의 연속적 발성으로 인해 약 4 Hz에서 에너지의 피크가 발생한다는 연구 결과[4]에 따라 필터뱅크 출력에서 4 Hz ME를 구하고 이를 특징으로 사용하여 음성과 음악을 판별하는 방법[5]이 제안되어 사용되어 왔다. 음악의 경우 일반적으로 음성에 비하여 변화의 속도가 느리므로 약 1 Hz 이하에서 큰 값을, 음성의 경우 약 2 Hz 이상에서 음성에서 큰 값을 갖는다.

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[6, 7]. Tyagi 등은 MFCC영역에서 구한 MCMS를 MFCC의 다이나믹 특징(dynamic feature)으로 사용할 경우, 차분 파라미터 및 RASTA PLP와 비교하여 가산 잡음 환경에서 음성 인식 시스템의 성능을 향상 시킬 수 있음을 보였다. MCMS의 정의는 다음과 같다.

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

여기에서 n은 프레임 인덱스, l는 MFCC 계수 인덱스, q는 모듈레이션 주파수 인덱스를, P는 주파수 분석을 위한 FFT 포인트수를, 그리고 C는 MFCC 계수를 의미한다.

본 논문에서는 MCMS가 MFCC의 각 계수별 스펙트럼을 모두 별도로 취급함으로써 특징 차수가 커지는 단점(즉, B개의 밴드패스 필터를 통하여 MCMS를 계산할 경우 추출되는 특징 벡터의 차수는 $B \times$ MFCC 차수가 됨)을 피하기 위해 다음과 같이 MCME를 정의하여 이를 특징 벡터로 사용한다.

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

여기에서 E는오디오 신호의 단구간 에너지를 의미한다. 즉 MCME는 동일한 모듈레이션 주파수에 대하여, 각 MFCC 계수들로부터 구한 파워에너지를 합한 값이다. 또한 모듈레이션 주파수 분석 구간내의 오디오 신호의 단구간 에너지의 평균으로 나누어 줌으로써, MCME 분석 결과의 변동폭을 줄였다.

저자는 제안된 MCME를 이용한 별도의 음성/음악 판별 실험에서 8 Hz MCME경우 4 Hz의 ME와 비교하여 71%, Cepstral Flux와 비교하여 53%의 판별 오류 감소율을 보이는 것을 확인한 바 있다. 따라서 본 논문에서는 음란 오디오 데이터의 경우 음성/음악 판별과 다소 다른 양상을 보일 수 있으므로 3 Hz ~ 50 Hz 범위의 모듈레이션 주파수에 대하여 MCME를 계산하고 이를 특징 벡터로 이용하고자 한다.

다음 그림 1은 MCME분석의 예를 확인하기 위하여 10초 분량의 FM 음악방송 중의 아나운서의 멘트에 해당되는 음성 신호와 음란 동영상 데이터에서 추출한 성행위 구간 신호와 록 음악 신호에 대한 파형과, 각각의 신호에 대하여 3.125 Hz 의 MCME를 추출한 결과를 그린 그림이다.

음성의 경우 음악에 비해 빠르게 변화하는 특성으로 인해 에너지 값을 나타내고, 음악의 경우 비교적 빠른

록 음악임에도 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.

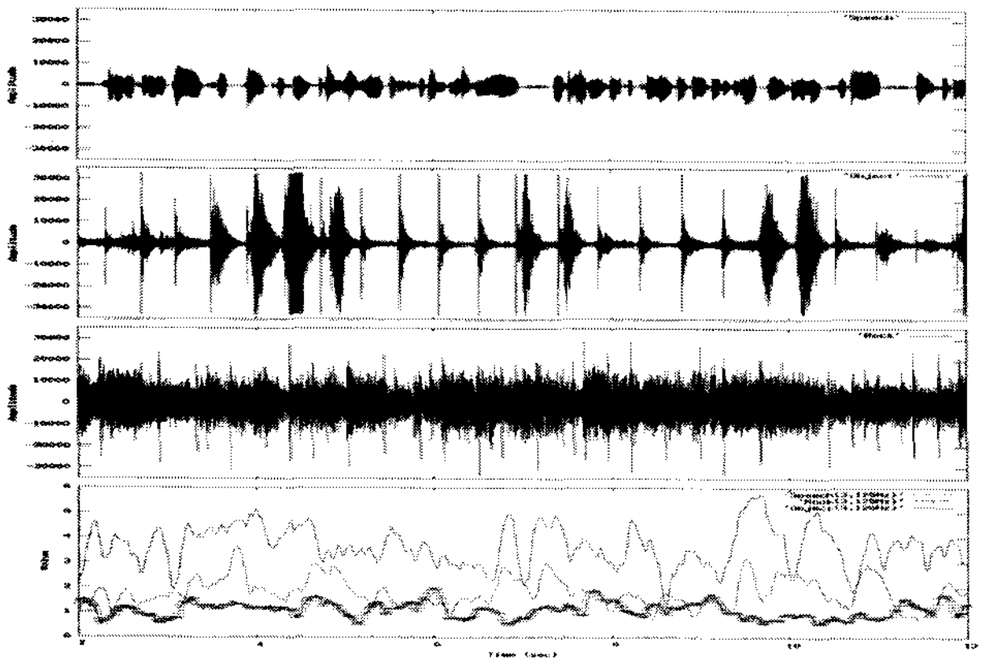

그림 1. 음성(첫 번째 창), 음란 오디오(두 번째 창), 록 음악(세 번째 창)에 대한 파형 및 그로부터 추출한 3.125 Hz MCME(네 번째 창 - 첫 번째 라인이 음성, 두 번째 라인이 음란 오디오, 세 번째 진한 라인이 록 음악으로부터 추출한 결과)

III. 데이터베이스

3.1 데이터베이스

음란 오디오 모델 및 안티 모델의 학습 및 평가를 위해 표 1과 같이 데이터베이스를 구성하였다. 음악 데이터의 경우 다양한 장르의 음악을 반영하기 위하여 RWCP Genre music DB[8]의 내용을 포함하였으며, 유명곡들을 반영하기 위해 RWCP Popular music DB [8]와 인터넷에서 구한 음악파일을 추가하였다.

일반 동영상 자료로는 2005년 10월 1일 ~ 2005년 10월 25일까지 25일분의 KBS 뉴스 동영상과, 인터넷에서 구한 26개의 다큐멘터리, 인터넷에서 구한 33개의 영화 동영상이 포함되었다. 영화 동영상의 경우 영화 1편이 여러개의 CD 분량으로 나누어져 있는 경우 그 중 1 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.

3.2 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.

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

표 1. 데이터베이스의 구성

구분	유형	파일수	분량 (시간)
음악	RWCP Genre music DB	102	7
	RWCP Popular music DB	100	7
	인터넷 음악	235	20
일반	KBS News	25	22
	Documentary	26	20
	Movie	33	45
	Sports	11	6
음란	-	149	46
계		681	173

IV. 판별 실험 및 결과

4.1 실험 환경 및 구성

특징 추출을 위하여 25ms의 해밍윈도우를 사용하여 10ms 단위로 프레임을 이동하면서 12차의 MFCC 결과와 에너지를 추출하였다. MFCC 결과에 대해 32포인트의 FFT를 프레임 단위로 이동 수행하면서 3.125 Hz ~ 46.88 Hz 범위의 15차의 MCME 를 추출하였다.

성능 비교를 위해 12차의 MFCC만 사용한 경우(MFCC), MFCC와 그 차분 및 차차분 파라미터를 사용한 경우(MFCC+D+A), MCME만 사용한 경우(MCME), MCME와 12차의 MFCC를 함께 사용한 경우(MFCC+MCME)로 구분하여 특징을 추출하였다.

실험에 사용된 판별기는 GMM을 사용하였으며 학습 데이터로부터 음란 모델과 일반 모델 및 음악 모델의 안티 모델을 학습하였다. 판별을 위한 테스트에서는 세 모델 중 음란 모델의 확률값이 가장 높으면 음란으로 판별하고, 일반 또는 음악의 확률값이 높으면 비음란으로 판별하였다.

4.2. 실험 결과

판별기로 사용된 GMM의 혼합의 수를 1부터 64까지 점진적으로 증가시키면서 판별 성능을 검증하였으며, 그 결과를 표 2에 정리하였다. 지면의 절약을 위하여 1개의 혼합을 사용한 경우와 각 특징별로 최고의 성능을 보인 경우를 요약하여 표에 정리하였다. 속도는 Pentium 4 3.0 GHz, 2 GByte의 RAM을 갖는 Winodws XP 환경에서, 특징 추출과 판별에 걸린 시간을 테스트 데이터의 시간으로 나눈 것이다(동영상 데이터에서 오디오 데이터를 추출하는 시간은 포함되지 않았음에 주의할 것).

표 2에 나타난 바와 같이 MFCC기반 방법의 경우 다이나믹 특징을 포함한 MFCC+D+A가 좋은 성능을 보이는 것을 볼 수 있다. 또한 MCME를 이용한 경우 더 적은 혼합에서 MFCC기반의 방법보다 좋은 성능을 보임을 알 수 있다. 18개의 혼합을 사용하는 MFCC+MCME의 경우 48개의 혼합을 사용하는 MFCC+D+A에 비해 평균 판별 오류 감소율 63%를 보이며, 속도에 있어서도 27%의 속도 향상을 볼 수 있다.

18개의 혼합을 사용하는 MFCC+MCME의 경우 총 546개의 테스트 데이터 중, 단 5개의 데이터에 대해서 판별 오류가 발생하였으며 이 중 음란 데이터를 비음란으로 판별한 경우가 3개, 비음란 데이터 중 음란으로 판별한 경우가 스포츠 데이터 2개이다.

표 2. GMM의 혼합 수에 따른 판별 성능 및 속도

특징 (차수)	구분	GMM의 혼합 수				
		1	12	18	32	48
MFCC (12)	음란	40.3	17.6	13.4	11.8	14.3
	비음란	55.0	7.5	7.0	4.9	3.5
	평균	47.7	12.6	10.2	8.3	8.9
	속도	0.008	0.009	0.009	0.010	0.012
MFCC+ D+A (39)	음란	16.8	5.0	5.0	5.0	5.9
	비음란	39.3	5.9	3.3	6.3	2.1
	평균	28.1	5.5	4.2	5.7	4.0
	속도	0.008	0.010	0.011	0.014	0.017
MCME (15)	음란	17.6	2.5	1.7	0.8	23.5
	비음란	5.9	1.9	10.3	20.4	3.3
	평균	11.8	2.2	6.0	10.6	13.4
	속도	0.010	0.011	0.011	0.013	0.015
MFCC+ MCME (27)	음란	16.0	3.4	2.5	1.7	5.0
	비음란	5.9	0.5	0.5	9.1	0.5
	평균	11.0	2.0	1.5	5.4	2.8
	속도	0.010	0.011	0.012	0.014	0.018

V. 결론

본 논문에서는 동영상에서 추출된 오디오 신호를 이용하여 음란 동영상을 검출하는 방법을 제안하였다. 제안된 MCME와 MFCC를 함께 이용할 경우 음란 데이터 판별율 97.5%, 비 음란 데이터 판별율 99.5%를 보임으로써 제안된 방법의 유효함을 알 수 있었다. 미리 추출된 142시간 분량의 테스트 데이터에 대하여 특징 추출 및 판별에 소요된 시간은 1시간 45분이다. 제안된 오디오 기반 음란 동영상 검출 방법은 그 단순성으로 인해 이미지 기반의 검출 방법보다 속도의 측면에서 장점이 있으며 이미지 기반 검출 시스템의 취약부분을 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 검증하고자 한다. 또한 간단한 오디오 신호만으로 판별이 어려운 데이터에 대한 판별 성능을 향상시키기 위한 방법을 연구할 필요가 있다고 판단된다.

참고문헌

[1] M. Hammani, Y. Chahir and L. Chen, "WebGuard : A Web Filtering Engine Combining Textural, Structural, and Visual Content-Based Analysis," IEEE Trans. on Knowledge and Data Engineering, Vol 18, No. 2, pp. 272-284. 2006

[2] H. Lee, S. Lee and T. Nam, "Implementation of High Performance Objectionable Video Classification System," Proc. ICACT2006, pp. 959-962, 2006

[3] W. Kim, H. Lee, J. Park and K. Yoon, "Multi Class Adult Image Classification Using Neural Networks," LNAI 3501, pp. 222-226, 2005

[4] T. Houtgast, H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acoustica, 28:66-73, 1973

[5] E. Scheirer, M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," Proc. ICASSP-97, Vol. 2, pp. 1331-1334, 1997

[6] V. Tyagi, I. McCowan, et al, "On Factorizing Spectral Dynamics for Robust Speech Recognition," Proc. Eurospeech-2003, pp. 981-984, 2003

[7] V. Tyagi, I. McCowan, et al., "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," Proc. ASRU '03, pp. 399-404, 2003

[8] RWC Music Database, http://staff.aist.go.jp/m.goto/RWC-MDB/