https://doi.org/10.5392/JKCA.2020.20.05.001

음향 신호의 양방향적 연관성을 고려한 유해 콘텐츠 검출 기법

Pornographic Content Detection Scheme Using Bi-directional Relationships in Audio Signals

송광호, 김유성
인하대학교 정보통신공학과

KwangHo Song(crossofjc@gmail.com), Yoo-Sung Kim(yskim@inha.ac.kr)

요약

본 논문에서는, 최근 인터넷을 통해 빠르게 확산하고 있는 음향 중심의 음란 콘텐츠를 정확하게 검출하기 위해, 음향의 이웃 신호들 사이에 존재하는 양방향적 연관성을 기반으로 콘텐츠의 유해성을 판단하는 기법을 제안한다. 이웃한 음향 신호들간의 양방향적 연관성을 추출하기 위하여, 양방향 확장-인과 컨벌루션 연산(bi-directional dilated-causal convolution operation)들을 수행하는 확장-인과 컨벌루션 블록을 쌓아 만든 다층구조 양방향 확장-인과 컨벌루션 네트워크를 제안한다. 제안된 유해 콘텐츠 검출 기법의 효용성 검증을 위한 실험에서는 음향 신호의 각 시점으로부터 추출한 단순 특징 벡터를 기계학습 모델로 분류하는 기존 방법, 기존의 확장-인과 컨벌루션 블록을 적용해 음향 시계열 데이터의 순 방향 연관성만을 이용하는 기법, 그리고 본 연구에서 제안한 음향 시계열 데이터의 양방향 연관성까지 이용하여 유해성을 판단하는 기법의 분류 정확성을 비교하였다. 실험 결과에 의하면 본 연구에서 제안한 기법이 최대 84.38%의 인식 정확도를 가지며 이는 기존의 단순 특징 벡터를 이용하는 방법보다 약 25.80% 높고 순 방향 연관성만을 이용하는 기법보다 약 3.10% 높은 것으로 분석되었다.

■ 중심어 : | 음란 콘텐츠 검출 | 음향 신호 | 양방향적 연관성 | 확장 컨벌루션 |

Abstract

In this paper, we propose a new pornographic content detection scheme using bi-directional relationships between neighboring auditory signals in order to accurately detect sound-centered obscene contents that are rapidly spreading via the Internet. To capture the bi-directional relationships between neighboring signals, we design a multilayered bi-directional dilated-causal convolution network by stacking several dilated-causal convolution blocks each of which performs bi-directional dilated-causal convolution operations. To verify the performance of the proposed scheme, we compare its accuracy to those of the previous two schemes each of which uses simple auditory feature vectors with a support vector machine and uses only the forward relationships in audio signals by a previous stack of dilated-causal convolution layers. As the results, the proposed scheme produces an accuracy of up to 84.38% that is superior performance up to 25.80% than other two comparison schemes.

■ keyword : | Pornographic Contents Detection | Audio Signal | Bi-directional Selationship | Dilated Convolution |

접수일자 : 2020년 01월 20일
수정일자 : 2020년 04월 23일
심사완료일 : 2020년 04월 27일
교신저자 : 김유성, e-mail : yskim@inha.ac.kr

I. 서 론

최근 온라인 콘텐츠 유통환경의 급격한 발달[1]과 함께 인터넷에 음란 콘텐츠들이 많이 유통되어 사회적 문제가 되고 있다. 특히 시각적 유해 요소가 배제된 채 성행위를 강하게 연상시키는 유해 음향만으로 구성된, 이른바 '듣는 야동'이 빠르게 확산됨에 따라 음향 중심의 유해 콘텐츠 검출에 관한 연구들이 증가하고 있다[2-6].

기존 연구[2-6]에서는 유해 음향 콘텐츠를 '사람의 교성이나 접촉음 또는 거친 호흡음 등과 같은 유해 음향들로 구성된 콘텐츠'로 정의하고, 이러한 유해 음향들이 콘텐츠 전체에 걸쳐 반복적이고 주기적으로 나타난다고 가정하여, 전체 음향을 주파수 스펙트럼(Frequency Spectrum), MFCC(Mel-Frequency Correlation Coefficient)[7], 자기상관계수(Autocorrelation)[8]와 같은 주파수 도메인의 디지털 신호로 변환하여 유해 음향 신호의 규칙적이고 주기적인 반복 패턴을 유해 콘텐츠 검출에 이용하고자 하였다. 그러나 [4][5]에 따르면 유해 음향은 콘텐츠 전체에 걸쳐 지속적, 반복적으로 나타날 뿐만 아니라 일부 구간에만 집중적으로 나타나기도 하기때문에 기존 기법으로 정확하게 검출되지 않는 경우가 발생할 수 있다. 또한, 온라인 스트리밍 서비스에서 콘텐츠의 유해 여부를 판단하기 위해 유해한 콘텐츠 전체가 재생된 이후에나 검출이 되는 문제점이 있다. [6]에서는 이런 문제를 해결하기 위해 입력된 음향 콘텐츠를 특정 길이로 분할한 음향 세그먼트 단위로 특징 벡터를 추출하고, 이렇게 추출된 음향 세그먼트의 특징벡터들을 Support Vector Machine(SVM) 기계학습 모델로 학습된 인식기로 입력 콘텐츠의 유해성을 판단하기 위해 사용한다. 그러나 이러한 일반적인 기계학습 모델은 입력 특징 벡터를 가상적 다차원 공간에 존재하는 하나의 점으로만 취급하기 때문에, 음향 세그먼트 내에서 진행시간에 따라 출현하는 각 시점의 신호 값 사이의 전후 관계, 출현의 연관성 등을 표현하지 못한다. 일반적으로, 사람의 육성, 교성, 거친 호흡음, 접촉음 등의 다양한 구성 요소들이 동시에 또는 연결되어 나타나는 유해 음향 콘텐츠의 구성 특성으로 인하여 각 시점의 신호와 이웃하는 시점의 신호들 사이에 존재하는 양방향적 연관성 또는 상관관계는 유해 콘텐츠의 검출 과정에서 활용 가능한 중요한 요소가 될 수 있다.

따라서 본 논문에서는 음향 세그먼트의 각 시점에 나타나는 신호와 일정 거리 안에서 이웃하는 다른 시점의 신호들 사이의 연관성 또는 상관관계를 반영하여 음향의 유해성을 결정하는 새로운 음향기반 유해 콘텐츠 검출 기법을 제안한다. 이를 위해 음향 세그먼트 내의 각 신호와 양방향으로 일정 거리 내에 이웃하는 다른 신호들과 갖는 연관성을 추출하여 분류 모델 학습에 반영하도록 인공신경망을 설계함으로써 해당 시점의 입력 음향신호는 물론 이를 둘러싼 전후 시점의 음향 신호들까지도 콘텐츠의 유해 여부를 판단하는 과정에 활용될 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2절에서는 음향 정보를 활용하여 유해 콘텐츠를 검출하는 기존 연구들을 소개하고, 시계열 데이터의 각 신호 값들이 인접한 이전 신호들과 갖는 연관성을 추출하기 위한 확장-인과 컨벌루션(dilated-causal convolution)[9]에 대해 알아본다. 3절에서는 음향 콘텐츠의 유해 여부를 판단하기 위해 음향 콘텐츠의 이웃 음향 신호들 사이에 존재하는 양방향적 연관성을 이용하기 위한 새로운 다층 구조의 양방향 확장-인과 컨벌루션 구조를 제안한다. 4절에서는 제안한 방법의 성능을 검증하기 위한 실험 분석에 대해 기술한다. 마지막으로 본 논문의 결론과 향후 연구에 대해서는 5절에서 기술한다.

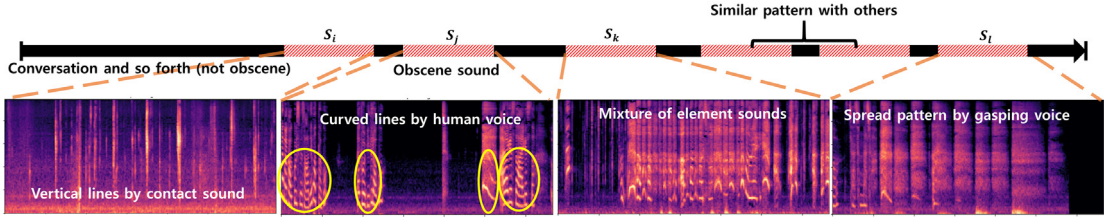


그림 1. 유해 콘텐츠내의 유해 음향요소 구성 예

II. 관련 연구

[그림 1]에 스펙트로그램(spectrogram)을 이용하여 축약적으로 표시한 유해 콘텐츠의 예와 같이, 일반적으로 유해 콘텐츠에는 사람의 신음, 교성, 거친 호흡음, 접촉음 등과 같은 유해행위와 관련된 음향들이 서로 연관되어서 전체 또는 일부분에 반복적으로 나타난다[2-6]. 이에 따라, 콘텐츠의 음향 신호를 식 (1)을 이용하여 사람의 가청 주파수 영역에 적합한 멜-스케일의 주파수(Mel-scale frequency)로 변환하고 MFCC나 MCME (Mel-Cepstrum Modulation Energy)[10] 등을 사용하여 인간의 가청대역과 일치하는 주파수 대역의 에너지 값들을 특징벡터로 표현하였다. 또한, 시계열에 나타나는 에너지 값의 반복적 증감을 수치화 할 수 있는 자기상관계수나 LPCC(Linear Prediction Cepstral Coefficient)[5]와 같은 특징들을 함께 조합함으로써 유해 음향 전체에 걸쳐 나타나는 소리의 반복적 증감 패턴이나 [그림 1]의 S_j 세그먼트에서 타원으로 마킹된 사람의 교성에 의한 강한 소리 신호(스펙트로그램에서 포락선으로 표시)의 반복적 발현 등과 같은 유해음향의 특성을 특징 벡터로 표현하였다.

$$Mel(f) = 2595*\log(1+f/700) \qquad (1)$$

그러나 사람의 육성으로 인한 신호 또는 다른 음향 신호들의 반복적 패턴은 비유해 음향 콘텐츠에서도 나타날 수 있으므로 이러한 음향 신호의 반복적, 주기적 특성이 유해 음향 콘텐츠만의 고유한 특성이라 하기 어렵다. 일반적으로 유해 음향 콘텐츠의 경우에는 앞서 언급한 신호의 반복적 증감이나 육성, 교성 등에 의한 강한 소리 신호의 반복적 발현 이외에도 [그림 1]의 S_i에서 관찰되는 반복적 접촉음에 의한 세로 선들의 군집, S_l에서 관찰되는 거친 호흡음에 의한 번짐 모양, 그리고 S_k에서 관찰되는 것과 같이 여러 음향 요소들이 함께 뒤섞여 나타나는 모양 등과 같은 신호 패턴들이 여러 부분에 함께 혼합되어 출현한다는 특성이 있다. 따라서 시계열 데이터 내의 신호세기의 증감과 같은 반복적 특성만을 강조하는 [2-6]의 특징추출 방법들로는 음향의 각 시점을 위한 특징을 추출함에 있어서 해당 시점과 이웃하는 시점에 출현한 소규모 신호 패턴들의 구성이나 출현의 전후 관계에 따른 양방향적 연관관계 등을 반영해내기 어렵다는 문제가 있다.

따라서 유해 콘텐츠 검출을 위한 기계학습 모델을 훈련하는 과정에서 이러한 특성들을 반영하여 검출에 활용해야 하나, 기존 연구들에서는 대표 유해 표본들과의 평균 코사인 유사도 비교[2]나 상관관계 비교[3]와 같은 단순 비교로 콘텐츠의 유해 여부를 판단하기 때문에 음향 신호의 양방향적 연관관계를 이용하지 못하였다. 또한 이전 연구들에서 사용한 GMM (Gaussian Mixture Model)[4], ANN(Artificial Neural Network)[5], SVM[6] 등과 같은 일반적인 기계학습 알고리즘 역시 학습과정에서 특징 벡터를 가상적 다차원 공간에 존재하는 하나의 점으로서 취급할 뿐 입력 음향 신호의 출현 순서나 상호 연관관계 등을 고려하여 특징 벡터를 추출하지 못한다. 따라서 유해 음향 콘텐츠 검출 방법에서 각 시점에 발현되는 음향 신호와 이전 시점에 나타났거나 또는 이후 시점에 나타난 음향 신호들 사이에 존재할 수 있는 양방향성 연관관계를 반영하기 위해서는 각 시점의 음향 신호로부터 추출한 특징벡터를 단순히 가상적 다차원 공간에 존재하는 하나의 점으로서 취급하는 것이 아니라 해당 시점의 특징벡터가 그와 이웃한 시점의 특징 벡터들과 가질 수 있는 연관관계를 추출하여 모델의 훈련에 활용하도록 해야 한다.

[9]에서는 음향 콘텐츠에서 각 시점에서 나타나는 음향 신호들은 주로 이전의 시점에 나타났던 신호들에 큰 영향을 받는다고 보고, 이러한 관계는 특정 시점에 일어난 사건에 대해 그 이전에 일어난 사건들이 미치는 인과적 효과(causal effect)[11]로 추론할 수 있다고 보았다. 인과적 효과란 매 시점에 일어나는 사건들이 서로 독립이라 가정할 때, 특정 시점에 일어난 사건에 대해 그 사건이 일어나기 이전에 일어났던 사건들이 미치는 인과적인 영향을 의미한다. 이를 추론하기 위해서는 조건부 결합 확률(conditional joint probability)의 곱을 이용하는데, T 길이의 입력 $X=\{x_{t-T+1},...,x_t\}$에 대하여 이전 시점의 입력들에 의해 t 시점에서 받는 인과적 영향 $p_c(x_t)$는 식 (2)와 같이 그 이전까지 일어난 사건들의 조건부 확률의 곱으로 나타낸다.

$$p_c(x_t) = \prod_{i=1}^{T-1} p(x_t|x_{t-1},\ldots,x_{t-i}) \ (t \geq T) \quad (2)$$

[9]에 따르면 이러한 인과적 효과의 추론을 위한 조건부 확률의 곱을 신경망을 이용해 추정하기 위해서는 인과 컨벌루션(causal convolution) 연산을 수행하는 계층을 포함하는 인과 컨벌루션 블록(causal convolution block)들을 여러 겹으로 쌓아 만든 다층구조의 인공신경망을 사용해야 한다. 그러나 이 방법은 학습 및 활용과정에서 과도한 연산 비용을 필요로 하므로, [9]은 이 비용을 줄이고자 [그림 2]와 같이 인과 컨벌루션 블록내의 인과 컨벌루션 계층을 확장-인과 컨벌루션(dilated-causal convolution)으로 대체하고 그 층이 쌓일 때마다 확장-인과 컨벌루션 연산의 확장률(Dilated Rate: DR)을 2의 배수로 늘리도록 하여 수용영역(Receptive field)내 신호들의 영향이 입력의 각 시점에 대하여 빠짐없이 반영되도록 하였다. 여기서 확장률이란 다층구조 확장-인과 컨벌루션의 각 레벨의 한 뉴런이 인과적 영향을 계산하기 위해서 참조해야 하는 하위 레벨의 다른 참조 뉴런과의 거리를 의미하며, 수용영역이란 확장률에 따라 각 계층의 뉴런이 인과적 영향을 계산하기 위해 참조해야 하는 입력 시계열 데이터에서의 이전 시점의 개수를 의미한다. 따라서 확장-인과 컨벌루션 블록들은 크기가 2인 커널을 사용하고 확장률을 쌓이는 블록의 수(n)에 따라 2배씩 증가시키므로, 수용영역의 크기는 2^n으로 늘어나게 된다. 따라서 [그림 2]에서 보는 바와 같이 입력 시퀀스에 대해 수용영역 T를 8로 확보하기 위해서는 $2^n = 8$ 을 만족하도록 n을 3으로 하여 3층의 블록을 쌓아 구성한다. 만일 일반적인 인과 컨벌루션 계층을 사용한 블록들을 통해 동일한 T 크기 수용영역을 확보하는 경우에 쌓아야 하는 블록의 수는 T-1이 되고 자연히 학습해야 하는 가중치의 수와 비용은 늘어나게 된다. 따라서 확장-인과 컨벌루션 블록을 사용하는 방법이 적은 수의 블록만으로도 상위계층에서 입력으로 수용하고자 하는 길이만큼의 수용영역을 확보할 수 있으므로 학습에 필요한 비용을 절감할 수 있다[9].

확장-인과 컨벌루션 블록은 [그림 3]과 같이 먼저 이전 계층의 출력을 입력으로 받아 확장-인과 컨벌루션 계층에 전달한다. 이어서 정류 활성화(Gated Activation)를 위해 그 출력에 서로 다른 두 개의 활성화 함수 tanh와 sigmoid를 적용한 후, 그 결과를 곱하여 relu 활성화 함수를 사용하는 1*1 컨벌루션 계층에 전달하며 이 컨벌루션 계층의 출력을 해당 블록의 출력으로 사용하게 되는데 이것이 [그림 2]의 각 블록에서 출력되는 조건부 확률이 된다. 다만 인과 컨벌루션이 아닌 확장-인과 컨벌루션을 사용했기 때문에 식 (2)에서 필요로 하는 모든 i에 대한 조건부 확률 중 일부만을 추론하게 되며 그에 따라 정확한 $p_c(x_t)$가 아닌 근사치를 추정하게 된다. 이어서 각 블록의 출력들을 모아 도약 연결(skip-connection)함으로써 각 층의 블록이 추론하는 수용영역에 대한 조건부 확률들을 앞서 식 (2)와 같이 그들의 곱으로 나타내도록 하며, 이러한 구조를 통하여 전체 모델은 음향 신호들 사이의 인과적 효과를 추론할 수 있게 된다. 또한 각 블록의 학습속도 증진을 위해 각 블록의 출력을 다음 블록의 계층을 위한 입력으로 넘겨줄 때 해당 블록의 입력과 더해주는 잔차(Residual) 학습을 적용하였다.

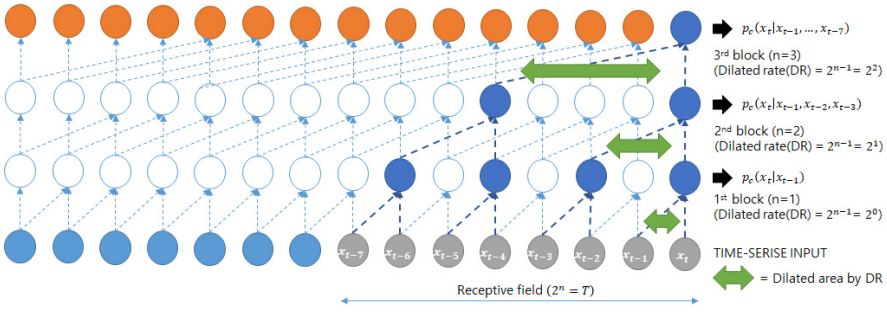


그림 2. 다층구조의 단방향 확장-인과 컨벌루션 예

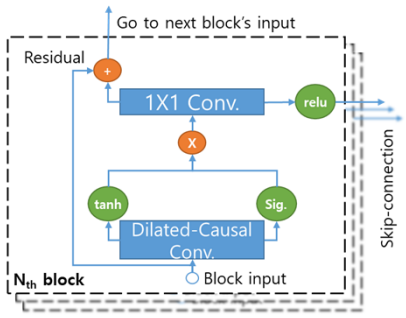


그림 3. 확장-인과 컨벌루션 블록 구조

그러나 다양한 음향 요소들이 다양한 패턴으로 조합되어 함께 나타나는 유해 음향 콘텐츠의 특성상 특정 시점 t의 신호와 이웃하는 신호들 사이의 연관 관계를 추론할 때, t 시점 이후에 나타나는 음향 신호들 역시 그 이전에 나타났던 신호들만큼이나 중요하게 고려되어야 한다. 따라서 연관관계를 추론할 때 양방향의 신호들을 모두 고려할 수 있어야 하나, 단방향의 신호만을 취급하는 확장-인과 컨벌루션으로는 이를 충족하기 어렵다. 따라서 본 연구에서는 음향 콘텐츠의 유해 여부를 판단함에 있어서 특정 시점에 나타난 음향 신호와 양방향으로 이웃하는 시점에 나타난 음향 신호들과 갖는 연관관계 또는 그로 인한 인과 영향을 추론하여 활용할 수 있도록 양방향 확장-인과 컨벌루션(bi-directional dilated-causal convolution)으로 개선하고 이를 다층구조로 구성한 다층구조의 양방향 확장-인과 컨벌루션 네트워크를 활용한 새로운 유해 음향 콘텐츠 검출 기법을 제안한다.

Ⅲ. 다층구조 양방향 확장-인과 컨벌루션 기반 유해 콘텐츠 검출 기법

본 절에서는 음향 콘텐츠의 유해 여부를 판단하는 모델이 특정 시점에 나타난 음향 신호가 양방향으로 일정 거리 내에 이웃하는 다른 시점의 음향 신호들로부터 받을 수 있는 영향을 추론하여 활용할 수 있도록 수정된 양방향 확장-인과 컨벌루션 계층을 쌓아 만든 다층구조의 양방향 확장-인과 컨벌루션 네트워크를 적용한 새로운 유해 콘텐츠 검출 기법에 관하여 기술한다.

본 논문에서 제안하는 음향 기반 유해 콘텐츠 검출 과정은 [그림 4]와 같이 입력 음향을 단위 세그먼트로 분리하고 이를 멜-스케일 주파수 기반의 신호 특성인 멜-스케일 스펙트로그램(Mel-scaled spectrogram)으로 변환하는 특징 추출(Feature extraction) 과정과 다층구조의 양방향 확장-인과 컨벌루션 네트워크와 다수의 1*1 컨벌루션 및 완전연결 계층들을 활용한 유해 여부 판단 과정으로 이루어진다.

먼저 특징 추출 과정에서는 [그림 4]의 ①과 같이 입력 콘텐츠로부터 음향을 분리하고, 전체 음향을 단위 시간의 세그먼트로 분절한 음향 세그먼트를 유해 여부 판단의 대상으로 한다. 이때, 음향 세그먼트의 길이는 기존 연구와의 비교를 위해서 [6]에서 단위 콘텐츠 길

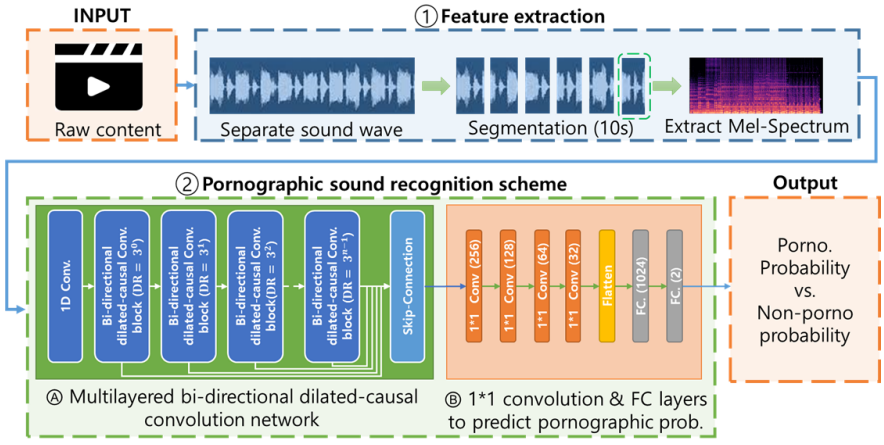


그림 4. 유해 음향 콘텐츠 검출 과정

이로 사용한 10초를 동일하게 사용하였다. 이어서 각 음향 세그먼트로부터 추출한 음향 신호의 주파수 데이터를 기존의 [2-6]에서와 같이 인간의 가청대역 및 비선형적 청음 특성에 가까운 멜-스케일의 주파수로 변환한 후, 사람이 발성 가능한 영역의 주파수 대역의 신호 값을 특징에 부각시킬 수 있는 [12]의 스펙트로그램의 형태로 변환한다. 이는 0 Hz ~ 8,192 Hz 사이에서 특정 시점에 나타나는 주파수 신호를 128개의 필터 뱅크(filter bank)를 이용하여 구간화하고 각 구간의 대표 에너지 값을 구하여 이를 시계열로 나타낸다. 따라서 스펙트로그램을 사용하면 최대한 원본 음향에 가까우면서도 미세 잡음에 의해 발생 가능한 오류를 보완함은 물론 사람이 발성 가능한 영역의 주파수 대역의 소리신호 값들을 부각시킬 수 있다. 따라서 각 주파수 영역의 신호 세기 값만을 알 수 있는 스펙트럼이나 그 값의 변화만 알 수 있는 MFCC, LPCC 등과는 달리 스펙트로그램은 시간의 흐름에 따라 나타나는 각 주파수 대역에서의 신호 에너지의 변화 추이까지 알 수 있으므로 더 많은 정보를 표시할 수 있다.

특징 추출 과정을 통해 만들어진 스펙트로그램들은 [그림 4]의 ②와 같이 인공신경망 기반의 유해 탐지 모델에 전달되어 원본 음향 세그먼트의 유해 여부를 판단한다. 본 논문에서 사용한 인공신경망 기반 유해 탐지 모델은 다시 크게 두 부분으로 나눌 수 있는데, 먼저 전반부Ⓐ에서는 입력으로 받은 스펙트로그램이 가진 각 시점에서의 특징 벡터들과 일정 거리 내에 이웃하는 다른 시점들의 특징 벡터들이 이루는 양방향 연관관계를 추출하기 위하여 양방향 확장-인과 컨벌루션 블록을 다층구조로 쌓아 구성하였다. 이어지는 Ⓑ에서는 Ⓐ로부터 전달되는 결과를 연속되는 1*1 컨벌루션 계층 및 완전연결 계층들에 전달하여 입력 음향의 유해여부를 두 개의 확률값으로 출력하도록 설계하였다.

먼저 Ⓐ의 첫 단계로 특정 시점 t에 나타난 음향 신호 x_t와 일정 거리 T내에 이웃하는 시점에 나타나는 신호들 사이의 양방향(bi-directional) 연관 관계 $p_{bd}(x_t)$를 추론하기 위해 x_t의 전후 양방향에 이웃한 신호들의 영향을 모두 고려해야 하므로 식 (3)과 같이 x_t를 포함하는 이전의 신호들$(x_{t-T+1},\ldots,x_{t-1},x_t)$은 물론 x_t를 포함하는 이후의 신호들$(x_t,x_{t+1},\ldots,x_{t+T-1})$도 조건부 확률에 포함하도록 한다.

$$p_{bd}(x_t)=\prod_{i=1}^{T-1}p(x_t|x_{t-i},\ldots,x_{t-1},x_{t+1},\ldots,x_{t+i})\ (t \geq T) \quad (3)$$

따라서 이를 구현하기 위한 확장-인과 컨벌루션 블록은 [그림 5]와 같이 양방향 확장-인과 컨벌루션(bi-directional dilated-causal convolution)으로 개선하여 양방향의 영향을 모두 고려할 수 있도록 확장하고 이들을 다층구조의 네트워크으로 구성한다. 이에 따라 최상위 계층의 출력에 해당하는 중심 시점 t를 기준으로 T 크기의 정방향 수용영역(forward receptive field) 내에 나타난 이전 시점의 입력들$(x_{t-T+1},\ldots,x_{t-1},x_t)$에 의한 영향과 같은 크기의 역방향 수용영역(backward receptive field) 내에 나타난 이후 시점의 입력들$(x_t,x_{t+1},\ldots,x_{t+T-1})$에 의한 영향을 고루 반영할 수 있도록 크기가 3인 커널을 사용하고 대칭적 확장-인과 컨벌루션 연산을 수행하는 양방향 확장-인과 컨벌루션 계층을 사용하였다. 따라서 컨벌루션 연산의 확장률(DR)을 다층 구조의 블록 수 n에 따라 3배씩

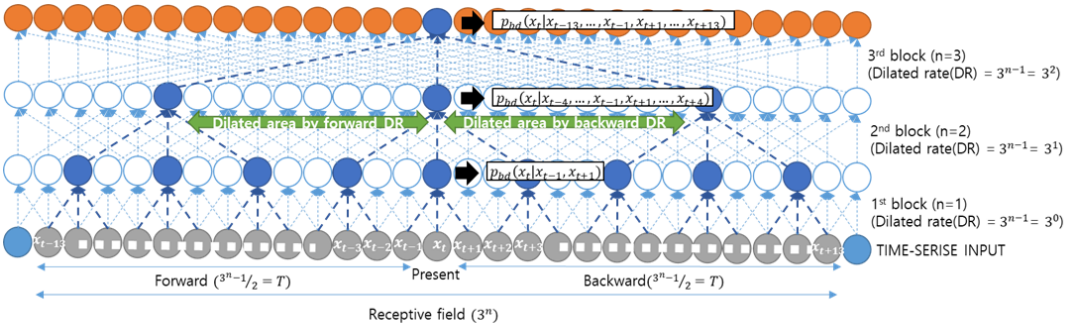


그림 5. 다층구조의 양방향 확장-인과 컨벌루션 예

증가하도록 하여 수용영역이 입력의 각 시점에 대하여 빠짐없이 미칠 수 있도록 하였다. 따라서 수용영역의 크기는 3^n으로 늘어나게 되며 [그림 5]에서와 같이 입력 시퀀스에 대해 기준 시점 t를 중심으로 정방향 수용영역의 크기 T와 역방향 수용영역의 크기 T를 함께 포함하는 2T-1 크기의 전체 수용영역을 27로 만들기 위해서는 $3^n = 27$ 을 만족하도록 n을 3으로 하여 3층의 블록을 구성하게 된다. 양방향 확장-인과 컨벌루션 블록의 나머지 부분은 [그림 3]의 구조와 동일하게 구성하였으며 이에 따라 각 블록의 출력들은 함께 모여 도약 연결(skip-connection)하게 된다. 이러한 구조를 통해 앞서 식 (3)에서 보인 바와 같이 각 층의 블록이 추론하는 수용영역에 대한 조건부 확률들을 그들의 곱으로 나타내도록 하며, 이를 통해 전체 모델이 음향 신호들 사이에 존재하는 양방향적 연관관계를 추론할 수 있도록 한다. 다만 [9]와 마찬가지로 밀집된 커널을 사용하는 일반 컨벌루션이 아닌 확장 컨벌루션을 사용했기 때문에 식 (3)에서 필요로 하는 모든 i에 대한 조건부 확률 중 일부만을 추론하게 되며 그에 따라 정확한 $p_{bd}(x_t)$가 아닌 근사치를 추정하게 된다.

마지막으로 유해 탐지 모델의 후반부에서는 [그림 4]의 ②의 Ⓑ에 표시된 바와 같이 전반부 Ⓐ에서 출력되는 행렬들을 4개의 1*1 컨벌루션 계층과 2개의 완전 연결 계층으로 구성된 신경망에 전달하여 이진 분류되도록 함으로써 입력 음향의 유해 여부를 예측하도록 한다. 이에 따라 모델은 그 입력에 대한 결과로서 해당 입력 세그먼트의 유해 정도를 표시하는 확률 값과 비유해 가능성을 표시하는 확률 값으로 출력한다.

Ⅳ. 실험 및 분석

본 절에서는 본 논문에서 제안하는 다층구조의 양방향 확장-인과 네트워크를 이용하여 유해 콘텐츠 검출 기법의 성능을 평가하기 위해 사용한 데이터를 소개하고, 실험 및 결과 분석을 기술한다. 먼저 모델의 학습 및 평가에 사용하기 위한 데이터 세트는 기본적으로 유해 콘텐츠 검출 기법을 개발하기 위해 널리 사용되는 'Pornography-2k'[13]을 활용하였다. 이 데이터 세트에는 1000개의 유해 영상과 1000개의 일반 비유해 영상으로 구성되어 있고, 비유해 영상으로는 레슬링이나 젖 먹이는 영상과 같이 유해 영상에서 나타날 수 있는 호흡음이나 접촉음 등을 담은 영상들이 다수 포함되어 있다. 추가로 최근 인터넷에서 문제가 되는 음향 중심의 유해 영상들의 특성을 모델에 반영하기 위해 약 140분 분량의 해당 영상들을 인터넷 환경으로부터 수집하였다. 'Pornography-2k'에 포함된 영상들 및 수집된 각 영상들의 재생 시간이나 속도 등과 같은 조건들이 서로 다르므로 이들을 균일하게 만들기 위해 각 영상을 44 KHz의 샘플 추출율(sampling rate)를 사용하여 10초 길이의 CD 수준의 음질이 보존된 음향 세그먼트들로 분리하였다. 이 과정을 통해 유해 음향과 비유해 음향이 각각 3300개씩 포함된 전체 6600개의 음향 데이터 세트를 구성하였으며, 유해 음향의 경우에는 'Pornography-2k'로부터 3000개 그리고 인터넷에서 자체 수집한 음향 데이터로부터 300개가 포함되었다. 모델의 학습 및 평가에는 10겹 교차 검증(10-fold cross validation) 기법을 사용하였으며, 이를 위해 각 회차 마다 데이터 세트로부터 6000개의 샘플을 무작위로 추출하여 학습에 활용하고 나머지 600개를 모델의 평가에 사용하는 과정을 10회 반복하고 각 반복 실험에서 산출된 결과들로부터 평균과 표준편차의 값을 계산하여 최종 결과로 나타냈다. 이어서 음향 세그먼트들을 멜-스케일의 스펙트로그램으로 변환하기 위하여 'librosa'[14] 라이브러리에서 제공하는 변환함수를 사용하였으며, 스펙트로그램 변환을 위한 샘플링 파라미터는 음향 세그먼트의 샘플 추출율인 44 KHz에 따라 함수 내에서 결정된다. 이에 따라 10초 길이의 음향 세그먼트들은 128차원 특징 벡터 431개로 이루어진 (431, 128) 크기의 스펙트로그램으로 변환되며 이를 학습 및 평가에 활용하였다.

첫 번째 실험에서는 [그림 6]에 제시된 바와 같이, 본 논문에서 제안한 다층구조의 양방향 확장-인과 컨벌루션 네트워크를 이용하여 음향 신호간의 양방향 연관관계를 반영하여 유해성 검사를 하는 기법과, [9]의 다층구조의 단방향 확장-인과 컨벌루션을 이용하여 음향 신호간의 전방향의 연관관계만을 이용하는 기존 기법, 그리고 [6]에서처럼 음향 신호를 독립적인 단순 특징 벡

터로 표현하고 SVM으로 분류하는 기법의 유해 음향 분류 정확도를 비교하였다. 이때 [9]의 단방향 확장-인과 컨벌루션 블록을 사용한 인공신경망의 입력에 대한 수용영역이 431개의 입력 특징벡터들을 모두 포함하려면 블록의 레벨 수가 총 9개이어야 하므로 동일한 구조의 인공신경망을 유지하기 위해 본 연구에서 제안하는 양방향 확장-인과 컨벌루션 블록을 사용한 인공신경망의 블록 수 또한 9개로 설정하고 부족한 입력은 제로패딩(zero padding)하였다. 이때 컨벌루션 블록 내 컨벌루션 계층이 갖는 출력의 차원 수는 입력되는 스펙트로그램의 차원 수와 동일하게 128로 설정하고 실험하였다.

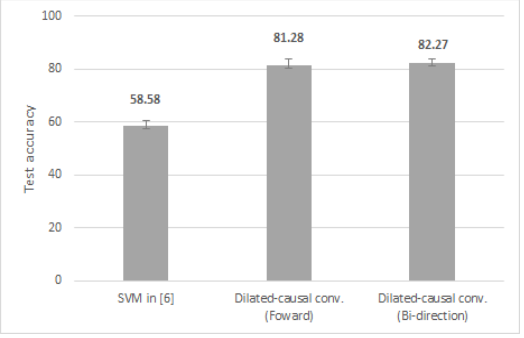


그림 6. 제안 기법과 이전 기법들의 검출 정확도 비교

실험 결과에 따르면 본 논문에서 제안한 다층구조의 양방향 확장-인과 컨벌루션 네트워크를 이용하여 유해 음향 콘텐츠를 검출하는 기법이 검출 정확도 측면에서 평균 82.27%(표준편차 1.34)로 가장 우수한 결과를 보였다. 이는 [6]에서와 같이 특징 벡터 사이의 상호 연관관계를 고려하지 않고 단순 특징 벡터로 사용하여 기계학습 방법인 SVM으로 유해성을 판정하는 기존 방법보다 평균적으로 23.69% 우수한 결과이며, 같은 특징벡터 시계열 데이터에서 전방향 인과 관계만을 추출하는 다층구조의 단방향 확장-인과 컨벌루션 네트워크를 이용하는 기법보다 평균적으로 0.99% 더 우수한 결과이다. 이에 따라 입력 음향의 유해 여부를 판단함에 있어 각 시점에 출현하는 음향 신호의 순서나 상호 연관관계 등을 고려할 때 해당시점 전후 양방향의 영향을 모두 고려하는 것이 이웃 신호들간의 상호 영향을 전혀 고려하지 않는 [6]보다 우수한 검출 성능을 가질 뿐만 아니라 [9]에서 사용한 인공 신경망으로 한쪽 방향의 영향만을 고려하는 방법보다 더욱 효과적인 방법으로 평가된다.

두 번째 실험에서는 본 논문에서 제안한 다층구조의 양방향 확장-인과 컨벌루션 네트워크의 최적 구조를 결정하기 위해서, [그림 7]에서와 같이 각 층의 컨벌루션 블록에서 컨벌루션 계층이 갖는 출력의 차원 수를 입력 특징 벡터의 크기인 128 차원의 1/8배 크기(16차원), 1/4배 크기(32차원), 1/2배 크기(64차원), 동일 크기(128차원), 2배 크기(256차원)로 바꾸어가면서 실험하여 컨벌루션 계층이 갖는 출력의 차원 수가 분류 결과에 어떠한 영향을 미치는지 분석하였다.

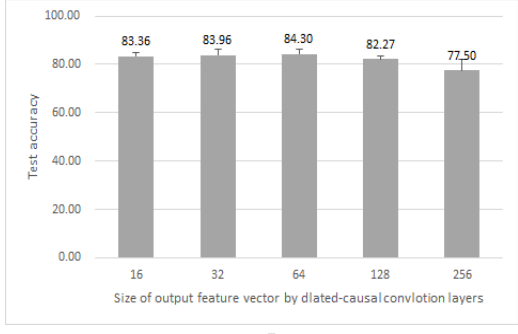


그림 7. 확장 컨벌루션의 출력크기에 따른 성능비교

[그림 7]과 같이 컨벌루션 계층의 출력 크기가 입력 스펙트로그램의 1/2배 크기인 64 차원일 때 정확도 측면에서 평균 84.30% (표준편차 2.04)로서 가장 우수한 성능을 보였으며, 출력의 크기가 커지는 경우와 작아지는 경우에 모두 평균 성능이 최소 0.34% 에서 최대 6.80% 씩 감소하는 모습을 보였다.

세 번째 실험에서는 다층구조의 양방향 확장-인과 컨벌루션 네트워크의 최적의 계층 수를 결정하기 위해서 양방향 확장-인과 컨벌루션 블록이 쌓인 계층 수에 따라 달라지는 수용영역의 크기의 변화에 따른 유해 음향 콘텐츠 검출 정확도를 비교 분석하였다. 이때 각 층의 컨벌루션 블록에서 컨벌루션 계층이 갖는 출력의 크기는 [그림 7]에서 가장 우수한 성능을 보인 64 차원으로 고정하였으며 확장 컨벌루션 블록의 수는 그 수용영역의 크기가 입력 특징벡터의 개수인 431을 넘지 않도록 1~6개까지 바꾸어가며 실험하였다.

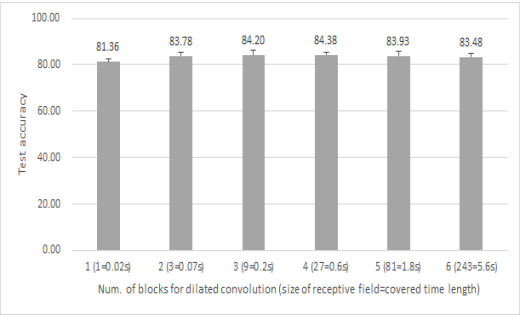


그림 8. 다층구조 양방향 확장-인과 컨벌루션 네트워크의 계층 수에 따른 성능 비교

[그림 8]에서와 같이 블록의 수가 4일 때 평균 84.38%(표준편차 1.13)의 정확도로 가장 좋은 성능을 보였으며 이는 다층구조의 양방향 확장-인과 컨벌루션 네트워크에서 양방향 확장-인과 컨벌루션 블록들에 의한 출력이 입력층에 대하여 해당 시점을 전후로 약 0.6초 범위의 수용영역을 가질 때, 즉 특정 시점의 입력에 대하여 0.6초 전후의 이웃한 입력들이 미치는 양방향의 영향을 반영하여 유해 음향 콘텐츠를 검출하는 것이 최대의 검출 정확도를 가지는 것으로 분석되었다. 따라서 실험 결과에 따르면 제안된 다층구조의 양방향 확장-인과 컨벌루션 네트워크는 확장 컨벌루션의 출력 크기를 64차원으로 하고 4개의 확장 컨벌루션 블록을 쌓아 다층 구조로 만드는 것이 최적의 구조라고 평가할 수 있다.

Ⅴ. 결론

본 연구에서는 유해 음향 콘텐츠를 검출하기 위해서 이웃한 음향 신호들간의 양방향 연관성을 추출하기 위한 다차원 양방향성 확장-인과 네트워크를 제안하여 유해성을 판단하는 기법을 제안하였다. 제안된 다층 구조의 양방향 확장-인과 컨벌루션 네트워크는 기존의 단방향 확장-인과 컨벌루션(forward dilated-causal convolution)을 확장하여 변형시킨 양방향 확장-인과 컨벌루션(bi-directional dilated-causal convolution)을 다층 구조로 쌓아 만든 네트워크 구조로서, 해당 출력의 각 시점을 중심으로 수용영역내의 입력 신호들이 미치는 양방향적 영향을 추출할 수 있기에 이를 이용하여 음향 콘텐츠의 유해 여부를 정확하게 판단할 수 있게 된다. 또한 실험을 통해 제안한 방법이 이전 시점으로부터의 단방향의 영향만을 고려하는 방법이나 이러한 영향을 전혀 고려하지 않는 방법들보다 더 우수한 84.38%의 정확도를 나타냄으로써 양방향의 영향을 고려하는 방법이 음향 신호를 기반으로 하는 유해 콘텐츠 탐지에 보다 적합한 방법임을 보였다. 향후에는 음향 정보뿐만 아니라 콘텐츠의 시각정보도 함께 활용함으로써 더 정확하게 콘텐츠의 유해 여부를 판별하는 연구를 진행할 예정이다.

참 고 문 헌

[1] 김현영, 김재웅, "'문화예술 콘텐츠 제작 및 유통에서의 빅데이터 활용 연구," 한국콘텐츠학회논문지, Vol.19, No.7, pp.384-392, 2019.

[2] Zhiyi Qu, Jing Yu, and Qiang Niu, "Pornographic Audio Detection Using MFCC Feature and Vector Quantization," International Conference on Computational and Information Sciences, pp.924-927, 2010.

[3] 조동욱, 김지영, "음란 유해사이트 차단을 위한 음향 신호 처리 및 분석," 한국콘텐츠학회논문지, Vol.4, No.2, pp.1-6, 2004.

[4] 김용운, 김봉완, 최대림, 김태권, 고락환, 이용주, "안드로이드 OS기반 음향정보를 이용한 음란동영상 검출 서비스 구현," 한국멀티미디어학회 학술발표논문집, pp.416-419, 2010.

[5] Ziqiang Shi, Tieran Zheng, Jiqing Han, and Boyang Gao, "Erotic Audio Recognition Using Heterogeneouos Ensemble Classifiers," International Journal of Computer and Electrical Engineering, Vol.4, No.5, pp.666-669, 2012.

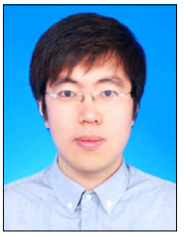
[6] KwangHo Song and Yoo-Sung Kim, "Pornographic Video Detection Scheme Using Multimodal Feature," Journal of Engineering and Applied Sciences, Vol.13, No.5, pp.1174-1182, 2018.

[7] M. Slaney, *A critique of pure audition*, Computational Auditory Scene Analysis, 1997.

[8] I. Mierswa and K. Moric, “Automatic feature extraction for classifying audio data,” Machine Learning Journal, Vol.58, pp.127-149, 2005.
[9] Google Deepmind, “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO,” arXiv:1609.03499v2 [cs.SD], 2016.
[10] B. Kim, D. Choi, and Y. Lee, “Speech/music discrimination using mel-cepstrum modulation energy,” Lecture Notes in Artificial Intelligence, Vol.4629, pp.106-414, 2007.
[11] J. Pearl, “Causal inference in statistics: An overview,” Statistics surveys, Vol.3, pp.96-146, 2009.
[12] X. Zhang, J. Yao, and Q. He, "Research of STRAIGHT Spectrogram and Difference Subspace Algorithm for Speech Recognition," 2nd International Congress on Image and Signal Processing, pp.1-4, 2009.
[13] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha, “Pornography Classification: The Hidden Clues in Video Space-Time,” Forensic Science International, Vol.268, pp.46-61, 2016.
[14] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of The 14th Python in Science Conference, pp.18-25, 2015.

저 자 소 개

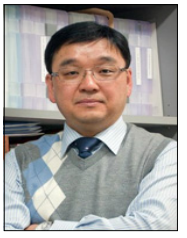
송 광 호(KwangHo Song) 정회원



- 2015년 2월 : 인하대학교 정보통신공학과(공학사)
- 2017년 2월 : 인하대학교 정보통신공학과(공학석사)
- 2017년 3월 ~ 현재 : 인하대학교 정보통신공학과 박사과정

〈관심분야〉 : 데이터 마이닝, 빅데이터

김 유 성(Yoo-Sung Kim) 정회원



- 1986년 2월 : 인하대학교 전자계산학과(이학사)
- 1988년 2월 : 한국과학기술원 전산학과(공학석사)
- 1992년 8월 : 한국과학기술원 전산학과(공학박사)
- 1992년 9월 ~ 현재 : 인하대학교 정보통신공학과 교수

〈관심분야〉 : 멀티미디어 마이닝, 빅데이터, 지능형 비디오 감시 시스템