pISSN 2005-8063
eISSN 2586-5854
2018. 09. 30.
Vol.10 No.3
pp. 19-29

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2018.10.3.019

ICA와 DNN을 이용한 방송 드라마 콘텐츠에서 음악구간 검출 성능*

Performance of music section detection in broadcast drama contents using independent component analysis and deep neural networks

허 운 행 · 장 병 용 · 조 현 호 · 김 정 현 · 권 오 욱**

Heo, Woon-Haeng · Jang, Byeong-Yong · Jo, Hyeon-Ho · Kim, Jung-Hyun · Kwon, Oh-Wook

Abstract

We propose to use independent component analysis (ICA) and deep neural network (DNN) to detect music sections in broadcast drama contents. Drama contents mainly comprise silence, noise, speech, music, and mixed (speech+music) sections. The silence section is detected by signal activity detection. To detect the music section, we train noise, speech, music, and mixed models with DNN. In computer experiments, we used the MUSAN corpus for training the acoustic model, and conducted an experiment using 3 hours' worth of Korean drama contents. As the mixed section includes music signals, it was regarded as a music section. The segmentation error rate (SER) of music section detection was observed to be 19.0%. In addition, when stereo mixed signals were separated into music signals using ICA, the SER was reduced to 11.8%.

Keywords: independent component analysis, deep neural network, segmentation error rate

1. 서론

방송 콘텐츠에서 음악은 저작권에 관련하여 민감한 문제를 가진다. 우리나라에서 전체 음원 사용료 징수 금액 중 지상파방송(TV, radio) 징수 금액의 비율은 12.9%로 9개 국가(미국, 일본, 독일, 영국, 프랑스, 이탈리아, 한국, 호주, 스페인) 중에서 가장 낮은 비율을 보인다(Lee, 2015). 호주는 41.0%로 아주 높은 비율을 보이고 있다. 이러한 현상은 음원 사용료가 제대로 징수되고 있지 않다는 것을 나타낸다.

현재 방송 콘텐츠에 쓰이는 음악 제목, 음악 구간 등의 정보는 음원을 사용하는 방송국에서 사람이 직접 입력한다. 사람이 직접 작성하기 때문에 음악 정보가 정확하지 않고, 시간과 노동이 많이 들어간다. 또한, 음원을 사용하는 방송국에서 작성하기 때문에 신뢰도가 낮아지는 문제점이 있다. 이러한 문제점을 보완하기 위하여, 자동 음악 검색을 위한 음악구간 정보를 자동으로 검출하는 방법을 제안한다. 기존의 음악구간 검출은 주로 뉴스에 대하여 실험하였지만(Gallardo-Antolín & Hernández, 2010; Gallardo-Antolín & Montero, 2010), 본 연구에서는 드라마 콘텐츠

* 본 연구는 문화체육관광부 및 한국저작권위원회의 2018년도 저작권 기술개발사업의 연구결과로 수행되었음(2018-micro-9500, 음악 및 동영상 모니터링을 위한 지능형 마이크로 식별 기술 개발).

** 충북대학교, owkwon@cbnu.ac.kr, 교신저자

Received 8 August 2018; Revised 26 September 2018; Accepted 27 September 2018

에 대하여 실험을 하였다. 뉴스 도메인(domain)은 주로 음성 구간이 많이 있다. 반대로, 드라마 도메인은 주로 음악 구간이 많이 있고, 음악과 음성이 섞인 혼합 구간 또한 높은 비율을 차지한다. 그러므로 뉴스 도메인에서는 음성을 잘 판별하는 것이 성능을 크게 좌우하지만, 드라마 도메인에서는 음악을 잘 판별하는 것이 관건이다.

기존 연구에서는 오디오 이벤트를 검출하기 위하여 NMF (non-negative matrix factorization) 기반의 음원 분리 기술을 적용한 연구(Heittola *et al*., 2011)가 존재하지만 모노 신호만을 대상으로 실험되었다. 반면에, 방송 콘텐츠는 일반적으로 스테레오 신호로 구성된다. 이러한 이유로 본 논문에서는 스테레오 환경에 적합한 독립 성분 분석(ICA, independent component analysis) 알고리듬(Hyvärinen & Oja, 2000)을 적용하여 혼합 신호의 음원을 분리한다. ICA는 스테레오 오디오 신호를 음악 신호 채널과 비음악 신호 채널로 분리한다. 두 채널 중 음악 신호를 찾기 위하여 각 채널에서 크로마그램(chromagram)의 엔트로피(entropy)를 비교한다. 오디오 신호에서 MFCC(mel-frequency cepstral coefficient)를 추출하여 가우시안 혼합 모델(GMM, Gaussian mixture model)과 심층신경망(DNN, deep neural network)을 이용한 분류기의 성능을 비교한다.

2. 기존 연구

오디오 신호에서 음악구간을 검출하는 연구는 기존에도 다양한 태스크에서 연구되었다. 보통 묵음, 음성, 음악, 잡음을 분류하기 위한 연구가 대부분이다.

기존 연구는 TV, 라디오 방송 뉴스 도메인에서 음악구간을 검출하거나 음향 샘플에 대하여 음성, 보컬이 있는 음악, 보컬이 없는 음악으로 분류하는 것이다. 성능 개선을 위하여 음향 모델링에 쓰이는 특징의 종류를 MFCC, 크로마그램, 필터뱅크 등과 여러 가지 통계치를 적용하거나 특징의 값을 변환하였다 (Gallardo-Antolín & Hernández, 2010; Gallardo-Antolín & Montero, 2010). 분류기는 은닉 마르코프 모델(HMM, hidden Markov model), GMM, 서포트 벡터 머신(SVM, support vector machine), DNN 등을 이용하였다. 또한, 여러 가지 분류 방법을 제시한 연구들이 있었다. 클래스를 하나씩 순차적으로 분류(Aguilo *et al*., 2009; Gallardo-Antolín & Hernández, 2010)하거나 한 번 분류된 음성 구간에 대하여 특징을 추가하여 더 세분화된 클래스를 분류(Castán *et al*., 2015)하는 방법이 있었다.

Gallardo-Antolín & Hernández(2010)에서는 방송 뉴스 오디오 신호에서 MFCC, PLP(perceptual linear prediction coefficients), 크로마그램을 계산하여 평균, 표준 편차, 왜도, 첨도 통계치를 구하고, 여러 통계치 조합에 따른 인식률을 비교하였다. 분류기는 HMM을 이용하였고, 구간을 분류할 때 서로 다른 특징을 이용하기 위하여 계층(hierarchy) 구조를 가진 분류 시스템을 이용하였다. MFCC, 크로마(chroma), 엔트로피 특징을 이용하여 음악 구간, 음성+음악 구간과 기타 구간을 먼저 분류하고, 기타 구간에서 MFCC 특징을 이용하여 음성, 음성+잡음을 분류하였다.

Gallardo-Antolín & Montero(2010)에서는 오디오 신호에서 얻은 특징에 히스토그램 등화를 이용하여 값을 변환하였다. 태스크는 음성, 보컬이 없는 음악, 보컬이 있는 음악을 분류하는 것이다. MFCC를 계산하였을 때 MFCC 차수의 분포가 다른 것을 보완하기 위하여 PHEQ(polynomial-fit histogram equalization; Wang *et al*., 2014)로 MFCC 값을 변환하였다. 변환 전에 비슷한 위치에 분포를 가지는 각 클래스가 PHEQ 변환 후에는 서로 다른 위치에 분포를 가지게 되어 쉽게 각 클래스를 구분할 수 있다.

또한, Aguilo *et al*.(2009)의 연구에서도 방송 뉴스 오디오 신호에서 계층 구조로 묵음, 음악, 음성 구간을 검출한다. 특징은 16개 필터뱅크의 로그 에너지와 델타, 델타-델타 값을 이용한다. 매 프레임의 에너지를 기반으로 첫 번째 묵음 구간을 찾는다. 첫 번째 묵음 구간이 아닌 구간에서 GMM-HMM 분류기를 이용하여 음악과 비음악 구간을 나누고, 비음악 구간에서 GMM 분류기를 이용하여 두 번째 묵음 구간과 비묵음 구간을 찾는다. 두 번째 비묵음 구간에서 통화 음성을 구분하기 위하여 새로운 스펙트럼 경사(spectral slope) 특징을 추가하여 SVM 분류기에서 음악+통화음성과 음악+통화음성이 아닌 구간을 검출한다. 음악+통화음성이 아닌 구간에서 마지막으로 음악+음성과 음성 구간을 GMM-HMM 분류기를 이용하여 검출한다. 계층 구조로 구간검출을 하면 각 클래스를 검출할 때 서로 다른 특징을 이용할 수 있다.

최근 연구로, Castán *et al*.(2015)은 TV, 라디오의 방송 뉴스 도메인에서 묵음, 음성, 음악, 잡음, 음악+잡음, 음악+음성, 음성+잡음, 음악+음성+잡음을 분류하는 태스크에 대하여 여러 팀들이 참가하여 인식률을 비교함으로써 다양한 접근 방법이 시도되었다. 가장 인식률이 높은 시스템은 2개의 서브시스템의 결과를 통합하는 과정을 가진다. 첫 번째 시스템은 13차 MFCC에 GMM-HMM을 이용하여 위의 8개 구간을 분류한다. 두 번째 시스템은 13차 MFCC 특징과 GMM을 이용하여 묵음, 음악, 잡음, 음성, 음성+잡음, 음성+음악 6개 클래스를 분류하고, GMM 결과의 음성 관련 구간에 대하여 i-vector 특징(Gupta *et al*., 2014)과 MLP(multi-layer perceptron) 분류기로 음성, 음성+잡음, 음성+음악, 음성+잡음+음악 클래스로 재분류한다. 2개의 서브시스템 결과를 결합하여 최종 결과를 얻는다.

3. 제안 방법

제안 방법의 구조도는 그림 1과 같다. 크게 음악/음성 분리 모듈과 음악구간 검출 모듈의 두 가지로 나눌 수 있다.

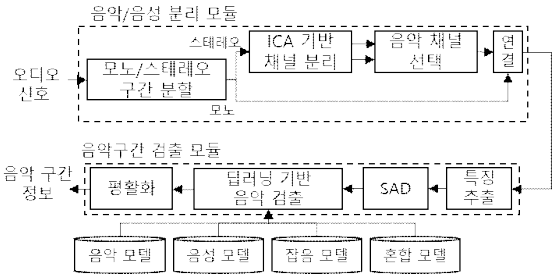

그림 1. 제안 방법 구조도
Figure 1. Block diagram of the proposed method

음악/음성 분리 모듈은 음악과 음성이 혼합된 스테레오 혼합 신호에서 음악 신호를 분리하여 음악 채널을 선택하고, 분리된 음악 신호와 모노 오디오 신호를 연결함으로써 모노 오디오 신호를 출력한다. 음악구간 검출 모듈은 오디오 신호에서 특징을 계산하여 SAD(signal activity detection)에서 묵음 구간을 검출하고, 딥러닝을 이용하여 음악(music), 음성(speech), 잡음(silence), 혼합(mixed) 구간을 검출한다. 프레임 단위 결과를 평활화(smoothing)하고 최종 음악 구간 정보를 얻는다.

3.1. 음악/음성 분리 모듈

3.1.1. 모노/스테레오 구간 분할

ICA는 스테레오 신호에 대하여 신호 분리가 가능하기 때문에 오디오 신호가 입력되면 모노/스테레오 구간을 분할한다. 모노/스테레오 구간은 아래의 식과 같이 오디오 신호에서 채널의 신호 차를 구하여 임계치보다 작으면 모노 구간, 임계치보다 크면 스테레오 구간으로 한다.

$$diff = \mathrm{mean}(\mathrm{abs}(x_l - x_r)) \tag{1}$$

x_l는 스테레오 신호에서 좌측 신호를 나타내고, x_r는 우측 신호를 나타낸다. 신호 차는 좌측과 우측 신호의 차에 절대 값을 취하여 평균을 구한다.

스테레오 구간은 다음 ICA 기반 음악/음성 분리 단계로 넘어가서 음악 신호를 분리하고 모노 구간은 다른 처리를 하지 않고 두 구간을 시간 순에 따라 연결한다.

3.1.2. ICA 기반 채널 분리

음악/음성 채널을 분리하기 위하여 ICA 알고리듬(Hyvärinen, 1999)을 적용하였다. 중심 극한 정리(central limit theorem)에 의하면 둘 이상의 신호가 혼합되면 분포는 정규분포에 가까워진다. ICA는 중심 극한 정리와 반대로 둘 이상의 신호가 혼합된 신호를 분리하기 위하여 입력 신호에 역혼합 행렬(demixing matrix)을 곱한 결과가 비가우시안 분포(non-Gaussian distribution)를 가지도록 역혼합 행렬을 추정하는 과정이다. 이 때, 비정규도(non-Gaussianity)를 측정하는 척도로서 첨도(kurtosis), negentropy, maximum likelihood, infomax, mutual information 등이 있다(Hyvärinen, 1999). 본 실험에서는 negentropy 척도를 이용하여 역혼합 행렬을 추정하였다. 아래의 식은 ICA 알고리듬 관련 식이다.

$$x = As \tag{2}$$

$$y = Wx \tag{3}$$

x는 혼합 신호, A는 혼합 행렬(mixing matrix), s는 소스(source) 신호, y는 추정된 소스 신호, W는 역혼합 행렬이다. 혼합 신호 x는 혼합 행렬 A와 소스 s의 곱에 의하여 생성된다. 추정된 소스 신호 y를 얻기 위하여 역혼합 행렬 W를 혼합 신호 x에 곱하여 얻는다. 이 때, 역혼합 행렬 W는 negentropy 척도를 이용하여 비정규도가 가장 크게 되도록 추정된다. 아래는 negentropy 척도 $J(y)$를 나타내는 식이다.

$$J(y) = H(y_{gauss}) - H(y) \tag{4}$$

H는 엔트로피 함수, y_{gauss}는 y와 동일한 평균 및 분산을 가지는 가우시안 분포를 의미한다. $J(y)$는 항상 0보다 같거나 크며, y가 정규 분포를 따를 때 0이 된다. 반대로 y의 비정규도가 클수록 더욱 큰 값을 가진다. y의 비정규도가 클수록 소스 신호 s를 잘 추정한 것이므로 $J(y)$가 큰 값을 가지도록 역혼합 행렬 W를 추정한다. 별도의 파라미터 없이 빠르게 역혼합 행렬 W를 추정하기 위하여 고정점 알고리듬(fixed point algorithm; Hyvärinen, 1999)을 이용한다. 역혼합 행렬 W를 혼합 신호 x에 곱하여 분리된 채널 신호를 얻는다.

분리된 채널 신호는 분리 이전 신호와 신호 크기가 다르기 때문에 신호 볼륨을 매칭하는 과정이 필요하다. 신호의 크기가 다르면 이후에 특징을 추출하였을 때, 에너지에 해당하는 특징에서 특징 값 차이가 크게 난다. 신호 볼륨을 매칭하기 위하여 분리된 신호에서 에너지가 높은 구간과 해당 구간에서 분리 전의 신호와의 정규화 비를 계산한다. 계산된 정규화 비를 분리된 신호에 곱하여 분리 이전 신호와 볼륨 크기를 매칭한다.

3.1.3. 음악 채널 선택

볼륨 매칭이 된 두 채널 신호에서 음악 채널을 선택하여야 한다. 음악 채널을 선택하기 위하여 각 채널에서 음악 신호의 특징을 반영한 값을 얻어야 한다. 음악 신호는 크로마그램의 12음계에서 비음악 신호보다 peakiness가 높게 나타난다. 이러한 특징을 반영하기 위하여 첨도, 엔트로피 등을 이용할 수 있다. 본 연구에서는 크로마그램에 엔트로피 개념을 적용하여 음악 채널을 선택하였다. 크로마그램은 모든 주파수 대역을 12음계의 벡터 길이가 1(unit length)이 되도록 정규화된 에너지로 나타내기 때문에 음악 특성이 잘 반영된다. 크로마그램 계산은 Chroma toolbox(Müller & Ewert, 2011)를 이용하였다. 크로마그램에서 peakiness를 수치화하기 위하여 스펙트럼 엔트로피(spectral entropy; Mirsa *et al*., 2004)를 응용하였다. 스펙트럼 엔트로피는 전력 스펙트럼 밀도를 확률로 간주하여 엔트로피를 정의한 것으로써, 본 논문에서는 크로마그램을 확률값으로 변환하여 크

로마그램 엔트로피를 정의하였다. 12음계 에너지의 peakiness가 높으면 엔트로피가 작고 peakiness가 작으면 엔트로피가 크기 때문에, 엔트로피가 평균적으로 작은 채널을 음악 채널로 선택한다. 음악 채널을 선택하기 위한 크로마그램 엔트로피 $H(p_{n,f})$ 식은 아래와 같다.

$$p_{n,f} = \frac{\sum_{m=1}^{10} x_{m,f}}{\sum_{f=1}^{12}\sum_{m=1}^{10} x_{m,f}} \quad (5)$$

$$H(p_{n,f}) = \sum_{f=1}^{12} p_{n,f} \log \frac{1}{p_{n,f}} \quad (6)$$

f와 m은 크로마그램에서 12음계 스케일과 0.1초 단위 프레임 인덱스를 나타낸다. $x_{m,f}$는 크로마그램에서 m 프레임, f 음계의 12음계 벡터 길이가 1로 정규화된 에너지 값이다. n은 1초 단위 프레임의 크로마그램 인덱스를 나타낸다. $p_{n,f}$은 n 프레임, f 음계의 확률값이다. 식 (5)는 시간 축을 따라 나타나는 횡성분을 고려하기 위하여 10개 프레임을 합하고, 12음계 스케일 에너지를 합이 1이 되도록 하는 확률 질량 함수이다. 식 (6)은 크로마그램 엔트로피를 구하는 식이다. 식 (6)에서 구한 두 채널의 크로마그램 엔트로피 평균을 비교하여 더 작은 평균값을 가지는 채널을 음악 채널로 선택한다.

3.2. 음악구간 검출 모듈

음악구간을 검출하기 위하여 오디오 신호에서 특징을 추출한다. SAD에서 묵음(silence) 구간을 찾고, 비묵음(non-silence) 부분에서 분류기를 이용하여 음악구간을 검출한다.

3.2.1. 특징추출

방송 콘텐츠에는 음성 신호와 음악 신호가 많이 존재하기 때문에 음성인식에서 많이 쓰이는 MFCC 특징을 이용하였다. 음성과 다른 종류의 음향들을 모델링하기 위하여 실험을 통하여 기존 음성인식과 다른 파라미터를 이용하였다. MFCC 특징과 MFCC 특징의 델타와 델타-델타 값을 추가하여 특징으로 이용하였다.

3.2.2. SAD

SAD는 각 프레임의 0 번째 MFCC 계수의 로그 에너지($\log E$)를 기반으로 묵음 구간인지 오디오 신호가 활성화된 구간인지를 탐색하는 부분이다. 신호의 에너지를 계산하고 임계치 이상이면 신호 활성 구간, 임계치보다 작으면 묵음 구간으로 결정한다. 임계치는 다음과 같은 식으로 정의한다.

$$T = 5.5 + 0.5 * \left(\frac{1}{N} \sum_{i=1}^{N} \left(\log x_i^2 \right) \right) \quad (7)$$

x_i는 i번째 프레임의 0 번째 MFCC 계수 값이고, N은 입력된 오디오 신호의 전체 프레임 개수이다. 0.5는 가중치, 5.5는 평균 에너지가 0일 경우를 대비하여 넣어주는 인자이다.

3.2.3. 딥러닝 기반 음악 검출

딥러닝 기반 음향 모델을 이용한 분류기가 GMM보다 성능이 더욱 좋다는 연구가 많이 있었다(Hinton *et al*., 2012; Saon *et al*., 2013). DNN을 이용하여 음악, 음성, 잡음, 혼합 클래스를 모델링하고, 프레임 단위로 클래스를 분류하여 음악 구간을 검출한다.

3.2.4. 평활화

분류기에서 추정된 결과는 프레임 단위이므로 연속된 프레임들 사이에 주변과 다른 결과를 가지는 프레임이 있다. 이러한 가짜 피크(spurious peak)를 없애주기 위하여 메디안 필터(median filter; Justusson, 1981)를 적용한다. 현재 프레임을 중심으로 양쪽으로 N개씩, 총 $2N$+1개의 프레임에 대하여 빈도수가 가장 높은 클래스로 현재 프레임을 재결정한다. 본 연구에서는 N을 50개로 설정하여 실험을 진행하였다.

4. 실험 결과

4.1. 데이터베이스

학습데이터는 공개된 음성/음악 데이터베이스인 MUSAN corpus(Snyder *et al*., 2015)를 사용하였다.

표 1. 학습에 사용된 MUSAN corpus
Table 1. The MUSAN corpus used in the training

변수	크기(시간)	선택(시간)	수집 방법
음악	42	25	보컬이 없는 음악
음성	60	25	무작위로 선택
혼합	-	25	음성과 음악 합성
잡음	6	6	모든 데이터 사용

표 1과 같이 음악 데이터는 전체 42시간 중에 보컬이 존재하지 않는 25시간이 사용되었고, 장르는 Western art music, jazz, bluegrass, hiphop 등으로 구성되어 있다. 보컬이 존재하지 않는 파일을 찾기 위하여 음악 데이터들의 보컬 존재 유무가 포함된 텍스트 파일을 이용했다. 음성 데이터는 음악 데이터와 동일한 길이를 갖도록 전체 60시간 중 25시간을 무작위로 선택하였다. 음성 데이터는 약 1–2분 정도의 독백으로 녹음된 잡음이 없는 음성 파일로 구성되어 있는데, 음성 파일 목록과 시간 정보가 포함된 텍스트 파일에서 목차를 무작위로 배열한 다음 음성 데이터 총 길이가 25시간이 될 때까지 파일을 선택하였다. MUSAN corpus에 잡음 데이터는 6시간 밖에 없어 6시간 전체를 이용하였고, 기계 소음(다이얼 톤, 사이렌 등)과 주변 소리(자동차, 비, 동물 소리 등)로 구성된다. 혼합 데이터는 음성과 음악의 소리 크기가 다른 다양한 테스트 환경에 대처하기 위하여 음성과 음악의 소리 크기를 일정한 범위에서 무작위로 조절하여 합성한

다. 방송 콘텐츠에서 음성과 음악이 혼합된 신호의 음악 대 음성비(music-to-speech ratio)를 측정해보니 5 dB~−14dB을 갖는 것을 알 수 있었다. 이러한 음악 대 음성 비를 고려하기 위하여 음성은 [−10, −1], 음악은 [−15, −5] 범위에서 무작위로 정수를 선택하고 해당 정수의 dBFS(decibel full scale)를 갖도록 진폭을 조절하여 반복적으로 혼합 데이터를 생성한다. 혼합 데이터는 음성과 음악 데이터 크기와 맞춰주기 위하여 25시간을 생성하여 학습에 사용한다. dBFS는 오디오 신호에서 최대 출력일 때, 0 값을 갖는다. 보통 오디오 신호는 16비트로 녹음되므로 0 dBFS일 때 2^{15} 값을 갖는다. 테스트 데이터는 방송 데이터인 드라마 "가면" 1–3회*(2015년 SBS 방영; SBS, 2015)의 약 3시간 분량이다. 그림 2는 테스트 데이터의 클래스 분포를 나타낸다. 음악 48%, 혼합 19%, 음성 14%, 잡음 10%, 묵음 9%로 구성되어 있다. 테스트 데이터는 Praat 프로그램(Boersma & Weenink, 2001)을 이용하여서 수작업으로 구간 정보를 담은 파일을 생성하였다. 방송 장르 중 드라마 장르는 특히 음악 구간이 많이 존재한다. 혼합과 음악부분을 합치면 약 60% 이상을 차지하는 것을 알 수 있다.

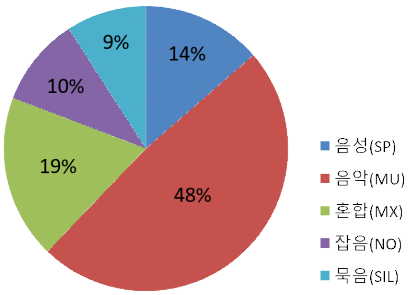

그림 2. 테스트 데이터의 클래스 분포
Figure 2. Class distribution of the test data

테스트 데이터에서는 음성 신호는 모노이고, 음악 신호는 대부분 스테레오 신호이다. 그러므로 ICA를 이용하여 스테레오인 혼합 신호를 음악 신호로 분리한다면, 19%의 혼합 신호는 대부분 음악 신호로 분리된다. 만약 음악 신호와 음성 신호가 모노라면 혼합 신호는 그대로 혼합 신호가 된다. 혼합 신호에서 ICA로 분리된 음악 신호는 음성 신호가 없어지지만 완벽하게 없어지지 않기 때문에 잡음이 남는 단점이 있다.

4.2. 음악/음성 분리 성능

모노/스테레오 구간 분할을 위하여 윈도우 크기 2초, 이동 크기 1초의 프레임에서 평균 차이 값을 계산하였다. 모노, 스테레오 구간을 결정하였을 때, 검출 오류율은 약 2.3% 정도로, 구간 경계에서 1–3초 오류가 발생하였다. 이 오류는 장면이 바뀌어 모노 음성이 나오지만, 앞서 나왔던 스테레오 음악의 잔향이 깔려있을 경우 해당 구간이 잘리게 되면서 생긴다.

음악 채널 선택을 위하여 매 프레임의 스펙트로그램 값(로그 전력)의 첨도를 계산하여 평균값이 큰 채널을 음악 채널로 결정하는 경우에, 음악 채널 선택 정확도는 89.9%였다. 제안한 방법인 크로마그램 엔트로피를 적용하면 음악 채널 선택 정확도는 98.7%로 나타났다.

그림 3은 테스트 데이터 중 "가면" 1회의 556–569초 스테레오 구간을 ICA를 통하여 두 채널로 신호를 분리하고 120 스케일 음계의 스펙트로그램과 12 스케일 음계의 크로마그램으로 나타낸 것이다. 스펙트로그램은 에너지를 로그 스케일(dB)로 변환하여 그래프를 그렸다. 스펙트로그램의 x축은 시간, y축은 120 스케일(10옥타브)의 음계를 나타낸다. 처음 1초 구간은 잡음이고, 561초부터 음악이다. 음악 채널에서 561초부터 60–90 scale 사이에서 높은 에너지를 갖는 것을 볼 수 있다. 음성 채널에서는 에너지가 넓은 범위에서 분포하는 것을 볼 수 있다.

크로마그램의 x축은 시간, y축은 12음계를 나타낸다. 스펙트로그램과 마찬가지로 561초부터 'D' scale에 높은 에너지를 갖는다. 각 채널의 크로마그램 엔트로피를 계산하면 음악 채널과 음성 채널의 차이가 확연히 드러난다.

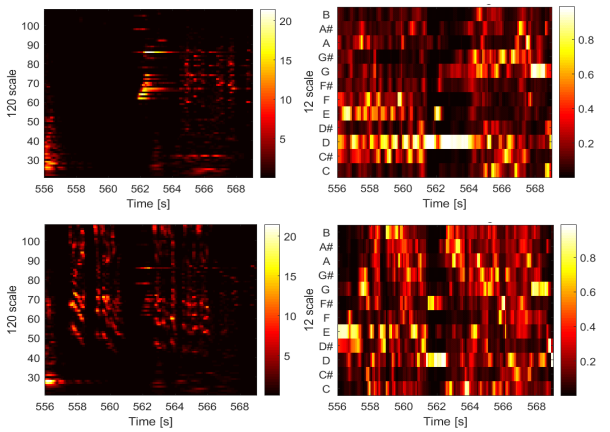

그림 3. 음악 채널의 스펙트로그램(좌측 위)과 크로마그램(우측 위), 음성 채널의 스펙트로그램(좌측 아래)과 크로마그램(우측 아래)
Figure 3. Spectrogram(upper left) and chromagram(upper right) of a music channel, Spectrogram(lower left) and chromagram(lower right) of a speech channel

그림 4는 그림 3의 크로마그램에서 엔트로피를 구하여 그래프로 나타낸 것이다. 처음 시작은 잡음에 의하여 음성 채널의 엔트로피가 높게 나왔지만 평균적으로 실선 그래프인 음악 채널의 엔트로피가 낮게 나오는 것을 알 수 있다. 스테레오 구간에서의 평균 엔트로피를 구하면 음악 채널을 쉽게 구분할 수 있다.

* 방송 콘텐츠를 인터넷으로 구입하여 연구 목적으로 사용하였으므로 저작권을 침해하지 않음.

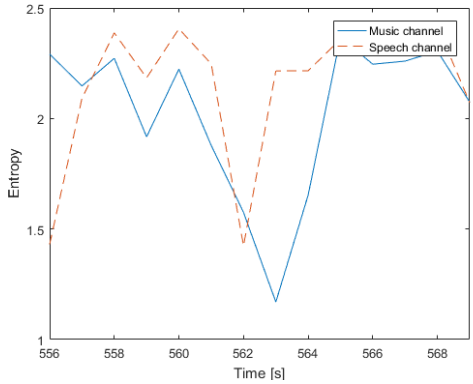

그림 4. 크로마그램 엔트로피 예시
(음악 채널: 실선, 음성 채널: 점선)
Figure 4. Example of chromagram entropy
(Music channel: solid line, Speech channel: dotted line)

4.3. 음악구간 검출 모듈 학습

MFCC 특징을 계산하기 위하여 멜 필터(mel filter) 차수는 65, 윈도우 크기는 125 ms, 이동 크기는 50 ms으로 설정하였다. 보통 음성 인식에 쓰이는 MFCC는 윈도우 크기 25 ms, 이동 크기가 10 ms이다. 윈도우 크기를 크게 하면 MFCC 고주파 성분을 더욱 잘 반영할 수 있다. 음악은 보통 음성보다 더 높은 주파수 범위를 가지므로 기존의 음성 인식에서 사용하던 윈도우 크기보다 더 큰 윈도우 크기를 가지도록 특징에 하여야 된다. 25 ms 윈도우 크기와 125 ms 윈도우 크기에서 얻은 13차 MFCC 특징을 이용하여 묵음, 잡음, 음성, 음악, 혼합 5개 클래스를 검출하는 사전 실험을 하였다. 25 ms 윈도우일 때 57.1% 분할 오류율 결과를 얻었고, 125 ms 윈도우일 때 48.3% 분할 오류율 결과를 얻을 수 있었다. 125 ms일 때 구간 검출 성능이 더욱 높은 것을 알 수 있다. 마찬가지 이유로 멜 필터 차수와 MFCC 차수도 크게 하였다. 실험을 통하여 멜 필터 차수와 MFCC 차수를 결정하였고, 여기에 델타, 델타-델타를 추가하여 최종 특징 벡터로 사용하였다. MFCC 특징 계산은 Kaldi toolkit(Povey *et al*., 2011)을 이용하였다.

딥러닝 기반 분류기를 구성하기 전에 베이스라인으로 GMM 기반 분류기(Metallinou *et al*., 2008)를 구축하였다. GMM 분류기는 평균 벡터와 완전 공분산(full covariance) 행렬을 갖는 64개 가우시안 분포를 사용하여 음향 클래스를 모델링하였고, DNN은 1,024, 1,024, 1,024, 1,024, 512, 512, 128개의 노드를 가지는 7개의 은닉층과 4개의 노드를 가지는 1개의 출력층으로 구성되어 있다. 은닉층의 활성 함수는 ReLU(rectified linear unit), 출력층은 softmax, 학습률은 0.0001, 40% 드랍아웃(dropout; Srivastava *et al*., 2014) 확률을 이용하였다.

4.4. 분할 오류율(SER, segmentation error rate)

분류 시스템의 성능은 분할 오류율(Galibert, 2013)을 이용하여 평가하였다. 시스템 성능을 수치화하기 위하여 아래와 같은 식을 이용한다.

$$SER = \frac{\sum_{n=1}^{N} E(n)}{\sum_{n=1}^{N} T(n) N_{ref}(n)} \tag{8}$$

$$E(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)] \tag{9}$$

N은 정답 또는 시스템 결과에서 클래스 구간이 바뀌는 부분을 분할하였을 때 생기는 세그먼트 개수이다. $T(n)$은 n번째의 세그먼트의 시간이다. $N_{ref}(n)$은 n번째 세그먼트에서 정답 클래스 개수를 의미하고, $N_{sys}(n)$은 n번째 세그먼트에서 시스템 결과 클래스 개수를 의미한다. $N_{correct}(n)$은 n번째 세그먼트에서 정답과 시스템 결과가 일치하는 클래스 개수를 의미한다. 본 시스템에서는 한 세그먼트에 한 가지 클래스만 존재하므로 $N_{ref}(n)$, $N_{sys}(n)$은 모두 1이다. $E(n)$은 본 시스템에서 정답과 출력 결과가 같으면 0이 되고, 정답과 출력 결과가 다르면 $T(n)$이 된다. SER은 전체 구간 길이에 대한 오류 구간 길이의 비로 결정된다.

그림 5를 보면, 정답과 시스템 결과의 경계를 기준으로 분할하여 여러 세그먼트로 나눈다. ①처럼 정답과 시스템 결과가 음성(SP)로 같으면 오류가 아니다. ②처럼 정답과 시스템 결과가 음성(SP)과 음악(MU)으로 다른 클래스라면 SE(segment error)로 오류가 된다. ③처럼 정답이 SIL(silence)이고 시스템 결과에서 음악(MU) 클래스가 나오면 FA(false alarm)이 된다. ④는 정답과 시스템 결과가 모두 묵음(SIL)이고 오류가 아니다. ⑤처럼 정답에 혼합(MX) 클래스가 존재하고 시스템 결과는 묵음(SIL)이면 MS(missed) 오류이다. ⑥은 ①과 같은 상황이므로 오류가 아니다. SER은 SE, FA, MS 구간 길이의 합에 오디오 전체 길이를 나눠서 구한다.

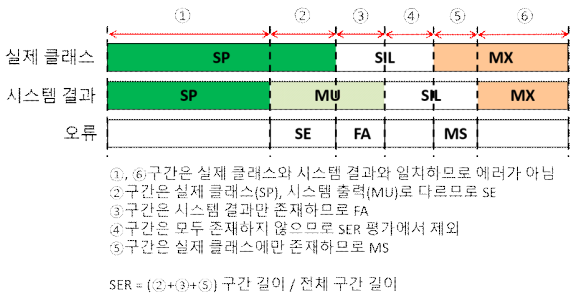

그림 5. SER 예시
Figure 5. Example of SER

4.5. 음악구간 검출 성능

혼합구간에도 음악 신호가 존재하므로 음악과 혼합 구간을 음악구간으로 정의하고 SER 결과를 계산한다. 실험 결과는 원본 신호와 ICA 출력 신호에 대하여 GMM과 DNN의 분류기를 이용한다. 여기에서 원본 신호는 ICA를 거치지 않고 입력된 스테레오 신호를 모노 신호로 변환한 것을 말한다.

표 2와 표 3은 음악구간 검출에 적절한 특징 파라미터를 얻기

위한 실험결과이다. 표 2는 멜 필터 차수와 MFCC 차수를 증가시키면서 ICA 출력 신호에 대하여 SER 결과를 나타낸 것이다. 학습과 테스트에 사용되는 특징 벡터는 MFCC 차수에 델타와 델타-델타가 추가된다. 표 2의 결과에서 멜 필터 차수 65, MFCC 차수 65 일 때, GMM 및 DNN에서 전체적으로 좋은 성능을 보인다. 표 3은 멜 필터 차수를 65로 고정하고, MFCC 차수를 증가시키면서 ICA 출력 신호에 대하여 SER 결과를 나타낸 것이다. 이 결과로부터 최대한 많은 MFCC 계수를 이용하는 것이 가장 좋은 성능을 보이는 것을 알 수 있다.

표 2. 멜 필터 차수/MFCC 차수와 ICA 출력 신호를 이용한 음악구간 검출 실험의 SER(%)
Table 2. SER(%) of music section detection experiment using ICA output signals with different mel filter order and MFCC order

차수(멜 필터/MFCC) / 분류기	40/40	50/50	65/65	80/80	95/95
GMM	17.0	15.8	**15.7**	16.6	16.3
DNN	14.3	12.9	**11.7**	11.8	11.5

표 3. 멜 필터 차수 65/MFCC 차수와 ICA 출력 신호를 이용한 음악구간 검출 실험의 SER(%)
Table 3. SER(%) of music section detection experiment using ICA output signals with 65 mel filter order and varying MFCC order

차수(멜 필터/MFCC) / 분류기	65/20	65/35	65/50	65/65
GMM	24.5	17.3	17.7	**15.7**
DNN	16.8	14.5	13.1	**11.8**

그림 6과 그림 7은 t-SNE(van der Maaten & Hinton, 2008)를 이용하여 원본 신호와 ICA 출력 신호의 65차 MFCC 특징을 2차원 평면에 시각화한 것이다. 그림 6을 보면, 원본 신호에서 음성(SP) 클래스와 음악(MU) 클래스가 좌우로 분포하고 있고, 혼합(MX) 클래스가 전체적으로 분포하고 있다. 이 신호를 ICA 출력 신호에 시각화 한 그림 7을 보면, 음성 클래스가 원본 신호에서 보다 군집화 되어 잘 구분되는 것을 알 수 있다. 결론적으로 원본 신호에서 얻은 MFCC 특징은 혼합 클래스가 음성과 음악 클래스에 전반적으로 나타나 음악 클래스 검출이 어렵지만, ICA 출력 신호에서 얻은 MFCC 특징은 음성 클래스가 군집화가 잘 되어 음악 클래스 검출이 더욱 쉬운 것을 나타낸다.

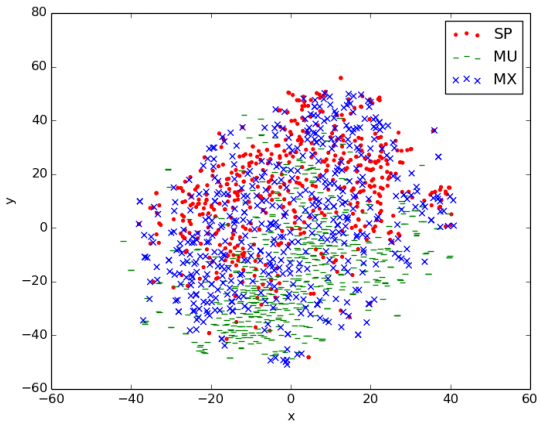

그림 6. t-SNE를 이용한 원본 신호의 MFCC 특징 시각화
Figure 6. Visualization of MFCC features for original signal using t-SNE

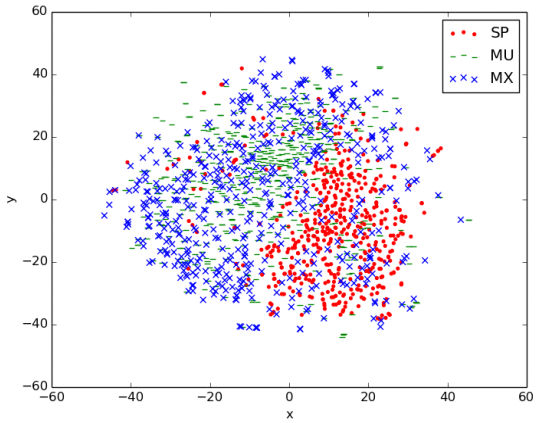

그림 7. t-SNE를 이용한 ICA 출력 신호의 MFCC 특징 시각화
Figure 7. Visualization of MFCC features for ICA output signal using t-SNE

표 4는 음악구간 검출 실험의 SER 결과이다. 원본 신호에 GMM을 이용하여 음악구간을 검출한 SER 결과는 20.4%이다. 원본 신호에 DNN 분류기를 이용하여 음악 구간을 검출한 SER 결과는 19.0%이다. ICA 출력 신호에 GMM을 이용하여 음악구간을 검출한 SER 결과는 15.7%이다. 마찬가지로 DNN 분류기와 ICA를 이용하여 음악구간을 검출한 SER 결과는 11.8%이다.

실험 결과를 보면, GMM보다 DNN이 더 좋은 성능을 보였고, ICA를 통하여 음악 신호를 분리하는 것 또한 원본 신호보다 더 좋은 성능을 보이는 것을 알 수 있다. ICA와 DNN을 동시에 적용한 음악구간 검출 결과가 가장 좋은 성능을 보였다. 원본 신호에 GMM을 이용한 베이스라인 실험 결과보다 ICA와 DNN을 이용한 제안 방법의 성능이 8.6% 절대적 성능 개선이 있었고, 42.2% 상대적 성능 개선이 있었다.

표 4. 음악구간 검출 SER(%)
Table 4. SER(%) of music section detection

입력 / 분류기	원본 신호	ICA 출력 신호
GMM	20.4	15.7
DNN	19.0	**11.8**

표 5–8은 4가지 결과의 묵음, 잡음, 음성, 음악, 혼합 클래스에 대한 혼동 행렬을 나타낸 것이다. 혼동 행렬은 평활화 전의 프레임 단위 분류기 결과에서 얻었다. 그림 8은 테스트 데이터의 분포를 고려하여 성능을 나타내기 위하여 표 5–8의 결과를 그래프로 나타낸 것이다. 그림 8의 (a)에서 경계를 기준으로 왼쪽은 테스트 데이터의 음악 구간 비율, 오른쪽은 테스트 데이터의 비음악 구간 비율을 나타낸다. (b)와 (c)에서 (a)의 음악 구간에 나타난 비음악 블록은 오류를 나타낸다. 마찬가지로 (a)의 비음악 구간에 나타난 음악 블록은 오류를 나타낸다.

원본 신호와 ICA 출력 신호에 대한 성능을 비교하기 위하여 표 5, 표 7과 표 6, 표 8을 비교하여 보면, 원본 신호에 대한 혼동 행렬인 표 5, 표 7에서 실제 혼합 클래스인 부분이 음성 클래스로 오인식되는 에러가 ICA에 대한 혼동 행렬인 표 6, 표 8에서는 현저하게 줄어든 것을 볼 수 있다. ICA로 혼합 신호를 음악 신호로 분리하여 줌으로써 원본 신호에서 실제 혼합 클래스인 부분이 음성 클래스로 오인식되는 오류를 줄여준다.

GMM과 DNN의 성능을 비교하기 위하여 그림 8의 (b), (c)의 아래 그래프를 보면 GMM과 DNN의 비음악 클래스의 성능은 같지만 음악 클래스의 성능은 DNN에서 더 높게 나온다. 생성 모델인 GMM보다 판별 모델인 DNN이 클래스 간의 차이를 학습함으로써 분류 태스크에서는 더 좋은 성능을 보이는 것을 확인할 수 있다.

표 5. 원본 신호에 대한 GMM 결과의 혼동 행렬(%/프레임)
Table 5. Confusion matrix(%/frame) of GMM results for original signal

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	**55.3/ 57267**	28.8/ 29857	1.6/ 1674	13.5/ 13973	0.8/ 12987
혼합	12.0/ 4777	**58.0/ 23061**	23.0/ 9145	6.9/ 2726	0.1/ 22
음성	3.4/ 925	14.4/ 4822	**62.6/ 18382**	3.2/ 4230	16.4/ 1007
잡음	12.1/ 2637	5.9/ 1284	7.4/ 1606	**61.0/ 13281**	13.6/ 2954
묵음	4.1/ 771	2.2/ 422	5.3/ 1007	20.0/ 3800	**68.4/ 12987**

표 6. ICA와 GMM을 이용한 시스템의 혼동 행렬(%/프레임)
Table 6. Confusion matrix(%/frame) of system using ICA and GMM

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	**53.9/ 55776**	37.5/ 38881	3.3/ 3410	4.8/ 4977	0.5/ 519
혼합	28.4/ 11297	**58.5/ 23240**	11.1/ 4416	1.1/ 423	0.9/ 355
음성	7.3/ 1322	11.0/ 6051	**56.6/ 16633**	4.5/ 3222	20.6/ 2138
잡음	15.9/ 3453	15.6/ 3394	10.2/ 2221	**40.3/ 8770**	18.0/ 3924
묵음	4.6/ 864	5.8/ 1095	4.6/ 880	19.5/ 3705	**65.5/ 12443**

표 7. 원본 신호에 대한 DNN 결과의 혼동 행렬(%/프레임)
Table 7. Confusion matrix(%/frame) of DNN results for original signal

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	**58.0/ 60042**	26.4/ 27349	1.2/ 1277	13.6/ 14103	0.8/ 792
혼합	13.1/ 5186	**57.9/ 22994**	21.0/ 8361	8.0/ 3168	0.0/ 22
음성	3.4/ 764	18.7/ 3725	**62.6/ 18368**	2.6/ 5502	12.7/ 1007
잡음	12.3/ 2678	3.3/ 720	8.9/ 1943	**61.9/ 13467**	13.6/ 2954
묵음	4.1/ 781	1.9/ 366	5.8/ 1103	19.8/ 3750	**68.4/ 12987**

표 8. ICA와 DNN을 이용한 시스템의 혼동 행렬(%/프레임)
Table 8. Confusion matrix(%/frame) of system using ICA and DNN

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	**66.6/ 68952**	28.4/ 29441	0.3/ 326	4.2/ 4325	0.5/ 519
혼합	43.1/ 17134	**52.5/ 20861**	1.5/ 591	2.0/ 790	0.9/ 355
음성	7.3/ 2131	14.3/ 5390	**52.8/ 15506**	7.2/ 4201	18.4/ 2138
잡음	21.5/ 4677	9.7/ 2103	7.5/ 1633	**43.3/ 9425**	18.0/ 3924
묵음	6.1/ 1154	5.4/ 1025	4.8/ 916	18.2/ 3449	**65.5/ 12443**

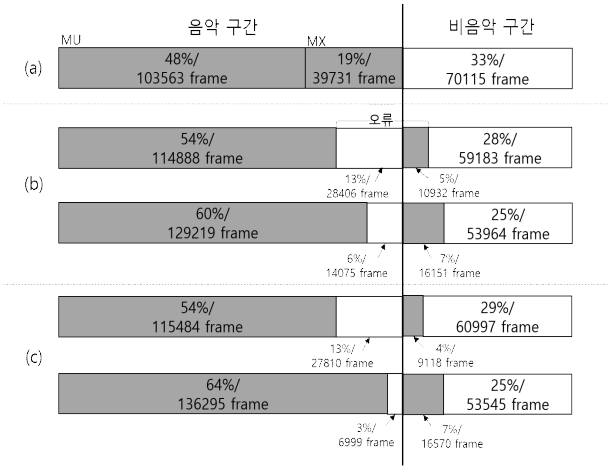

그림 8. 음악구간 검출 실험의 프레임 단위 결과
(a): 테스트 데이터의 음악/비음악 구간 분포, (b): 원본 신호에 대한 GMM 결과(위), ICA와 GMM을 이용한 시스템 결과(아래)/ (c): 원본 신호에 대한 DNN 결과(위), ICA와 DNN을 이용한 시스템 결과(아래)
Figure 8. Frame-based result of music section detection experiment.
(a): Music/non-music section distribution of the test data, (b): GMM results for original signal(upper), system using ICA and GMM(lower), (c): DNN results for original signal(upper), system using ICA and DNN(lower)

MUSAN corpus의 95%는 학습에 사용하고, 약 4시간 분량의 5%는 테스트 데이터로 이용하여 GMM과 DNN의 성능 차이를 알아보았다. MUSAN corpus는 모노 신호이므로 제안 방법에서 음악/음성 분리 모듈이 적용되지 않았다.

표 9와 표 10은 학습에 이용하지 않은 MUSAN corpus를 이용한 GMM, DNN 시스템 결과의 혼동 행렬이다. 그림 9는 분포를 고려한 결과를 나타내기 위하여 그래프로 나타낸 것이다. 그림 9의 그래프는 GMM과 DNN의 성능 차이를 보기 힘들다. 표 9와 표 10을 살펴보면 혼동 행렬에서 GMM보다 DNN에서 음악과 혼합 클래스를 잘 구분하는 것을 볼 수 있다. 드라마 실험 결과에서와 마찬가지로 특성이 유사한 혼합과 음악 클래스를 DNN이 GMM보다 더욱 잘 모델링하는 것을 알 수 있다.

표 9. MUSAN corpus에서 GMM을 이용한 시스템 결과의 혼동 행렬(%/프레임)
Table 9. Confusion matrix(%/frame) of GMM-based system in the MUSAN corpus

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	39.1/ 37092	**58.7/ 55637**	0.0/ 36	0.7/ 623	1.5/ 1424
혼합	10.3/ 10604	**88.0/ 90311**	1.5/ 1519	0.0/ 0	0.2/ 204
음성	0.0/ 0	0.0/ 0	**91.2/ 49515**	0.5/ 275	8.3/ 4491
잡음	4.8/ 1190	3.1/ 765	3.6/ 886	**84.4/ 20891**	4.1/ 1027
묵음	-	-	-	-	-

표 10. MUSAN corpus에서 DNN을 이용한 시스템 결과의 혼동 행렬(%/프레임)
Table 10. Confusion matrix(%/frame) of DNN-based system in the MUSAN corpus

출력 / 실제	음악	혼합	음성	잡음	묵음
음악	**52.2/ 49517**	45.7/ 43291	0.0/ 0	0.6/ 615	1.5/ 1389
혼합	4.6/ 4679	**94.9/ 97381**	0.3/ 357	0.0/ 0	0.2/ 221
음성	0.0/ 0	0.0/ 28	**91.9/ 49876**	0.3/ 144	7.8/ 4233
잡음	5.6/ 1391	1.2/ 290	0.2/ 49	**88.9/ 22001**	4.1/ 1028
묵음	-	-	-	-	-

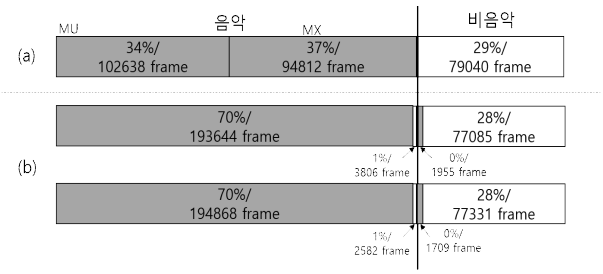

그림 9. MUSAN corpus에서 프레임 단위 시스템의 음악구간 검출 결과
(a): 테스트 데이터의 음악/비음악 분포, (b): GMM 시스템 결과(위)와 DNN 시스템 결과(아래)
Figure 9. Music detection results of frame-based system for the MUSAN corpus.
(a): Music/non-music distribution of the test data, (b): Result of GMM-based system(upper) and result of DNN-based system(lower)

5. 결론

본 논문에서는 드라마 콘텐츠에서 음악구간 검출을 위한 방법을 제시하였다. ICA에서 스테레오 혼합 신호에서 음악 신호를 분리하여 혼합 모델에서 음성으로 오인식되는 오류를 줄이고, DNN을 이용하여 GMM보다 음향 모델 성능을 개선하였다.

실험 결과를 통하여 MFCC 특징의 적절한 파라미터를 결정하였다. 보통 음성인식에 쓰이는 MFCC 특징보다 더 많은 멜 필터와 MFCC 특징 개수가 쓰였다. 음성보다 음성 신호가 나타나는 주파수 영역이 더 넓기 때문인 것으로 보인다. t-SNE를 통하여 원본 신호와 ICA 출력 신호의 MFCC 특징을 시각화하였을 때, ICA 출력 신호에서 얻은 MFCC 특징이 음악구간 검출에 더욱 적합하다는 것을 볼 수 있었다. 원본 신호에서 음악구간을 검출하는 것보다 스테레오 혼합 신호를 ICA를 통하여 음악 신호를 분리하여 음악구간을 검출하는 것이 성능이 더욱 좋다는 것을 알 수 있다. ICA와 DNN을 적용한 실험이 가장 좋은 성능을 보였다.

스테레오 구간의 혼합 신호를 음악 신호로 분리하여 줌으로써 실제 혼합 클래스인 부분이 음성 클래스로 오인식되는 오류를 줄이고, DNN을 이용함으로써 음향 모델 성능을 개선하였다.

앞으로, DNN보다 음향 모델 성능이 우수한 다른 딥러닝 관련 모델을 적용하여 음악구간 검출 성능을 개선할 계획이다.

참고문헌

Aguilo, M., Butko, T., Temko, A., & Nadeu, C. (2009). A hierarchical architecture for audio segmentation in a broadcast news task. *Proceedings of I Iberian SLTech* (pp. 17-20).

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10), 341-345.

Castán, D., Tavarez, D., Lopez-Otero, P., Franco-Pedroso, J., Delgado, H., Navas, E., Docio-Fernández, L., Ramos, D., Serrano, J., Ortega, A., & Lleida, E. (2015). Albayzín-2014 evaluation: Audio segmentation and classification in broadcast news domains. *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, 2015(1), 33.

Galibert, O. (2013). Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech. *Proceedings of INTERSPEECH-2013* (pp. 1131-1134).

Gallardo-Antolín, A., & Hernández, R. S. S. (2010). UPM-UC3M system for music and speech segmentation. *Proceedings of VI Jornadas en Tecnología del Habla II Iberian SLTech Workshop (FALA)* (pp. 421-424).

Gallardo-Antolín, A., & Montero, J. M. (2010). Histogram equalization-based features for speech, music and song discrimination. *IEEE Signal Processing Letters*, 17(7), 659-662.

Gupta, V., Kenny, P., Ouellet, P., & Stafylakis, T. (2014). I-vector based speaker adaptation of deep neural networks for French broadcast audio transcription. *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing (ICASSP)* (pp. 6334-6338).

Heittola, T., Mesaros, A., Virtanen, T., & Eronen, A. (2011). Sound event detection in multisource environments using source separation. *Proceedings of Workshop Machine Listening in Multisource Environments* (pp. 36-40).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6), 82-97.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks*, 10(3), 626-634.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5), 411-430.

Justusson, B. I. (1981). Median filtering: Statistical properties. In T. S. Huang (Ed.), *Two-Dimensional Digital Signal Processing II* (pp. 161-196). Berlin: Springer.

Lee, G. H. (2015). A study on the appropriateness of music broadcasting fee of terrestrial broadcasters. *Music Content and Law*, 203-250. (이규호 (2015). 지상파 방송사의 음악저작물 방송사용료의 적정성에 대한 연구. *음악콘텐츠와 법*, 203-250.)

Metallinou, A., Lee, S., & Narayanan, S. (2008). Audio-visual emotion recognition using Gaussian mixture models for face and voice. *Proceedings of International Symposium on Multimedia (ISM)* (pp. 250-257).

Mirsa, H., Ikbal, S., Bourlard, H., & Hermansky, H. (2004). Spectral entropy based feature for robust ASR. *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing (ICASSP)* (pp. 193-196).

Müller, M., & Ewert, S. (2011). Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features. *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)* (pp. 215-220).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Saon, G., Soltau, H., Nahamoo, D., & Picheny, M., (2013). Speaker adaptation of neural network acoustic models using i-vectors. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 55-59).

SBS (2015). SBS drama special: Mask. Retrieved from http://programs.sbs.co.kr/drama/2015mask on September 27, 2018.

Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. Retrieved from https://arxiv.org/abs/1510.08484 v1 on September 27, 2018.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9(1), 2579-2605.

Wang, S. S., Lin, P., Lyu, D. C., Tsao, Y., Hwang, H. T., & Su, B. (2014). Acoustic feature conversion using a polynomial based feature transferring algorithm. *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 454-458).

- **허운행 (Heo, Woon-Haeng)**
  충북대학교 제어로봇공학전공 대학원생
  충북 청주시 서원구 충대로 1(개신동)
  Email: whheo@cbnu.ac.kr
  관심분야: 음성인식, 감정인식
  2017-현재 제어로봇공학전공 박사과정 재학 중

• **장병용 (Jang, Byeong-Yong)**
충북대학교 제어로봇공학전공 대학원생
충북 청주시 서원구 충대로 1(개신동)
Email: byjang@cbnu.ac.kr
관심분야: 음성인식, 유창성
2015-현재 제어로봇공학전공 박사과정 재학 중

• **조현호 (Jo, Hyeon-Ho)**
충북대학교 제어로봇공학전공 대학원생
충북 청주시 서원구 충대로 1(개신동)
Email: jhh9601@cbnu.ac.kr
관심분야: 오디오 분할, 음성 신호처리, 음성인식
2017-현재 제어로봇공학전공 석사과정 재학 중

• **김정현 (Kim, Jung-Hyun)**
한국전자통신연구원 책임연구원
대전광역시 유성구 가정로 218
Email: bonobono@etri.re.kr
관심분야: 콘텐츠 식별, 음원 분리, 유사음악 검색

• **권오욱 (Kwon, Oh-Wook)** 교신저자
충북대학교 전자공학부 교수
충북 청주시 서원구 충대로 1(개신동)
Tel: 043-261-3374
Email: owkwon@cbnu.ac.kr
관심분야: 음성인식, 감정인식, 음성신호처리