일반논문 (Regular Paper)
방송공학회논문지 제22권 제5호, 2017년 9월 (JBE Vol. 22, No. 5, September 2017)
https://doi.org/10.5909/JBE.2017.22.5.632
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

미디어 오디오에서의 DNN 기반 음성 검출

장 인 선[a], 안 충 현[a], 서 정 일[a], 장 윤 선[b)‡]

DNN based Speech Detection for the Media Audio

Inseon Jang[a], ChungHyun Ahn[a], Jeongil Seo[a], and Younseon Jang[b)‡]

요 약

본 논문에서는 미디어 오디오의 음향 특성 및 문맥 정보를 활용한 DNN 기반 음성 검출 시스템을 제안한다. 미디어 오디오 내에 포함되어 있는 음성과 비음성을 구분하기 위한 음성 검출 기법은 효과적인 음성 처리를 위해 필수적인 전처리 기술이지만 미디어 오디오 신호에는 다양한 형태의 음원이 복합적으로 포함되어 있으므로 기존의 신호처리 기법으로는 높은 성능을 얻기에는 어려움이 있었다. 제안하는 기술은 미디어 오디오의 고조파와 퍼커시브 성분을 분리하고, 오디오 콘텐츠에 포함된 문맥 정보를 반영하여 DNN 입력 벡터를 구성함으로써 음성 검출 성능을 개선할 수 있다. 제안하는 시스템의 성능을 검증하기 위하여 20시간 이상 분량의 드라마를 활용하여 음성 검출용 데이터 세트를 제작하였으며 범용으로 공개된 8시간 분량의 헐리우드 영화 데이터 세트를 추가로 확보하여 실험에 활용하였다. 실험에서는 두 데이터 세트에 대한 교차 검증을 통하여 제안하는 시스템이 기존 방법에 비해 우수한 성능을 보임을 확인하였다.

Abstract

In this paper, we propose a DNN based speech detection system using acoustic characteristics and context information of media audio. The speech detection for discriminating between speech and non-speech included in the media audio is a necessary preprocessing technique for effective speech processing. However, since the media audio signal includes various types of sound sources, it has been difficult to achieve high performance with the conventional signal processing techniques. The proposed method improves the speech detection performance by separating the harmonic and percussive components of the media audio and constructing the DNN input vector reflecting the acoustic characteristics and context information of the media audio. In order to verify the performance of the proposed system, a data set for speech detection was made using more than 20 hours of drama, and an 8-hour Hollywood movie data set, which was publicly available, was further acquired and used for experiments. In the experiment, it is shown that the proposed system provides better performance than the conventional method through the cross validation for two data sets.

Keyword : Speech Detection, Voice Activity Detection

a) 한국전자통신연구원 방송·미디어연구소 미디어연구본부 테라미디어연구그룹(Media Research Division, ETRI)
b) 충남대학교 전자공학과(Dept. of Electronic Engineering, Chungnam National University)
‡ Corresponding Author : 장윤선(Younseon Jang)
E-mail: jangys@cnu.ac.kr
Tel: +82-42-821-6586
ORCID: http://orcid.org/0000-0001-5698-6233
※ 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2015-0-00860, 시청각장애인 방송 접근권 향상을 위한 디지털자막·음성해설 서비스 기술 개발).
· Manuscript received July 11, 2017; Revised July 27, 2017; Accepted July, 27, 2017.

Ⅰ. 서 론

최근 기계학습의 눈부신 발전으로 인해 미디어 서비스 기술이 고도화 되고 있으며 그 분야가 다양화 되고 있다. 특히 음성 인식 및 음성 전사 등 음성 신호 처리 분야에서는 딥 러닝과의 접목을 통해 괄목할 만한 성능 개선을 이루고 있으며 또한, 미디어 서비스로의 확장을 위한 다양한 연구가 활발히 진행되고 있다.

그 중 전통적으로 음성 신호 처리를 위한 전처리 기술로 연구되어온 음성 검출은 미디어 오디오에서의 음성신호 처리를 위한 필수 불가결한 요소 기술로써 그 중요성이 높아지고 있다[1][2]. 미디어 오디오는 음성 및 묵음뿐 만 아니라 음악, 음향 효과 등 다양한 음원을 포함하고 있으며 음원들이 서로 중첩되어 있는 경우가 많다. 또한, 음성 신호 자체도 평이한 감정에 대한 정적인 형태가 아닌 감정적인 음성, 속삭임, 보컬 등 다양한 상황에 대한 발성을 포함하고 있다. 이와 같은 음향적 복잡성으로 인해 미디어 오디오에서의 음성 신호 처리는 음성 검출이 선행되어야 해당 응용 분야에서의 성능을 보장 할 수 있게 된다.

미디어 오디오 중 음성 검출에 대해 가장 먼저 연구가 시작된 분야는 방송 뉴스로써 1990년대 후반에 활발히 연구되었다[3]. 관련 연구로는 MFCC (Mel-Frequency Cepstral Coefficient)를 이용하여 훈련한 HMM (Hidden Markov Model) 기반의 음성 검출 기술이 있었으며 MFCC, ZCR (Zero Crossing Rate) 등 기존 오디오 특징들의 조합을 이용하여 훈련한 SVM (Support Vector Machine) 기반의 접근이 있었다[4][5]. 이외에도 스펙트럴 특성과 하모닉 강화 기반의 오디오 특징 추출 방법에 대한 연구를 통해 뉴스 오디오에서의 음성 검출 성능 개선이 보고되었다[6].

2010년 초 이후 다양한 미디어 분야에 대한 음성 검출 기술이 발표되고 있다. [7]에서는 유튜브와 같은 웹 비디오에 대한 음성 검출 연구가 있었다. 이 연구에서는 기존의 오디오 특징들과 분류기를 조합하여 약 95시간의 유튜브 콘텐츠에 대한 음성 검출 성능을 분석하였다. 유튜브 콘텐츠에 대한 또 다른 연구에서는 MFCC를 이용하여 훈련한 DNN (Deep Neural Network) 기반의 음성 검출 시스템이 제안되었으며, DNN 입력 벡터 구성 시 문맥 윈도우 (context window)를 적용함으로써 음성 검출 성능이 향상됨을 보여주었다[8].

영화 오디오에서의 음성 검출은 장시간의 시간적인 문맥을 모델링 할 수 있는 LSTM (Long Short-Term Memory) 기반의 접근이 있었다[9]. 이 연구에서는 오디오 특징으로 RASTA-PLP (Relative Spectral Analysis-Perceptual Linear Prediction)와 그들의 1 차 미분 값을 사용하였으며 음성 데이터베이스인 Buckeye[10]와 TIMIT[11]을 배블(babble), 도시(city), 백색, 핑크 등 4종류의 잡음 및 음악 신호와 임의로 믹싱한 데이터를 이용하여 실험을 수행하였다. 또한 4편의 영화 오디오에 대한 음성 검출 실험을 하여 실제 미디어 오디오 환경에서의 성능을 검증하였다. [12]에서는 오디오 신호의 시간적 특성과 고조파성을 표현하는 오디오 특징 추출 방법에 대한 연구가 있었다. 이 연구에서는 SVM 기반의 음성 검출 분류기를 기반으로 4시간 길이의 라디오 방송과 8시간의 영화 데이터를 이용하여 제안하는 오디오 특징 추출 기술에 대해 검증하였다. 참고로, 이 연구에서는 [9]에서 사용된 데이터에 대한 검증 및 개정이 이루어졌으며 실험에 사용된 영화 데이터 세트는 이후 범용으로 공개되었다. 한편, 드라마에서의 음성 검출을 위해 방송 오디오의 스테레오 채널 구조를 활용한 접근이 있었다. 이 연구에서는 드라마 오디오에서 배우의 대사 신호의 음상이 스테레오 음상의 중앙에 위치함을 실험적으로 분석하였으며 드라마 오디오의 센터 채널과 서라운드 채널로부터 추출한 STE (Short Time Energy)와 ZCR를 오디오 특징으로 하여 음성을 검출하였다[13]. 또한, 실제 드라마 오디오 60분에 대한 실험을 수행하였다. 하지만, 음성뿐 만 아니라 현장 녹음으로 수음된 음향 효과 및 BGM의 보컬 등 또한 센터 채널 음상을 가지므로 규칙 기반의 분류기 사용으로 인한 성능의 한계가 있었다.

이와 같이, 미디어 오디오에 특화된 음성 검출 기술의 연구는 최근 들어 시작되었으며, 아직까지는 기존 오디오 특징 추출과 분류 방법의 조합을 통한 연구가 그 주를 이루고 있다. 또한, 기존 연구들에서는 임의로 믹싱한 오디오 데이터를 이용하거나 실제 미디어 오디오에 대해 크지 않은 사이즈의 데이터 세트를 이용하여 성능을 검증하였다. 이는 미디어의 저작권 등의 이슈로 인해 공개된 음성 검출용 미디어 오디오 데이터 세트가 거의 없기 때문이다.

본 논문에서는 미디어 오디오의 음향 특성을 고려하여 구성한 DNN 기반의 음성 검출 시스템을 제안한다. 제안하

는 시스템은 미디어 오디오를 고조파 및 퍼커시브(percussive) 성분으로 분리하여 추출한 오디오 특징을 이용하며[14] 오디오의 문맥 정보를 포함하도록 입력 벡터를 구성함으로써[8] 미디어 오디오의 음향적·문맥적 특성을 활용하여 DNN 기반으로 음성 구간 검출을 수행한다. 제안하는 시스템의 우수성을 검증하기 위하여 드라마와 영화, 두 종류의 미디어 오디오 데이터 세트를 사용하였다. 드라마 데이터 세트는 음성 검출을 위해 자체적으로 제작한 총 20시간 정도의 데이터 세트이며 영화 데이터 세트는 오스트리아의 Johannes Kepler 대학의 B. Lehner 등이 공개한 데이터 세트로 4편의 영화에 대한 총 8시간 정도의 음성 검출용 데이터 세트이다[12]. 음성 검출 시스템 내 DNN 구조 설정을 위해 은닉 층의 개수를 점진적으로 증가시키며 성능 검증을 수행하였으며 이후 최적의 은닉 층 개수를 가진 DNN 구조를 이용하여 각 데이터 세트의 교차 실험을 통해 제안하는 음성 검출 기술의 우수성을 확인하였다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 미디어 오디오의 음향적 특성을 설명하고 음성 검출 관점에서 이 특성들을 활용할 수 있는 방안을 자세히 기술한다. Ⅲ장에서는 제안하는 미디어 음성 검출 시스템을 상세히 설명한다. 또한, Ⅳ장에서는 제안하는 미디어 음성 검출 시스템의 성능을 검증하는데 사용한 데이터 세트를 소개하고 Ⅴ장에서는 실험 방법을 기술하고 그 결과를 분석한다. 마지막으로, Ⅵ장에서는 본 논문에 대한 결론을 맺으며 향후 연구 계획을 제시한다.

Ⅱ. 미디어 오디오의 음향적 특성

본 장에서는 음성, 음악, 음향 효과, 잡음 등 미디어 오디오를 구성하는 다양한 음원에 대한 스펙트로그램 상의 음향적 특성을 정리하고 음성 검출 관점에서 이 특성들을 활용할 수 있는 방안을 제안한 [14]의 연구 내용을 자세히 기술한다.

음성 신호는 조음 시 성대의 울림 여부에 따라 유성음과 무성음으로 나눌 수 있다. 파형의 주기성이 있는 유성음이 대부분 고조파 성분으로 이루어진 것과는 대조적으로, 파형의 주기성이 없는 무성음은 고조파 구조를 가지고 있지 않으며 백색 잡음과 유사한 특성을 띈다. 그림 1에서는 16kHz 샘플링 율을 갖는 음성 신호의 파형과 스펙트로그램에서 유성음과 무성음의 비교 예를 도시하였다. 해당 신호는 TIMIT 데이터베이스 중 하나의 여성 음성을 나타낸 것으로 ‘she had your dark suit in greasy wash-water all year’를 발성한 여성의 음성이다[11]. ‘she’ 에 대해 발성된 무성음 /s/와 유성음 /iy/의 스펙트로그램에서 확인할 수 있듯이 유성음은 스펙트로그램 상에서 시간 축을 따라서 에너지가 모여 있고 무성음은 시간 축과 주파수 축에 편향되지 않은 등방성으로 에너지가 분포하고 있다.

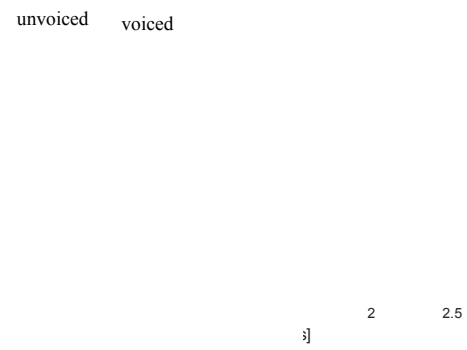

그림 1. 음성 신호의 파형과 스펙트로그램 내 유성음과 무성음 비교 예
Fig. 1. A comparison example of voiced and unvoiced sound in waveform and spectrogram of speech signal

음악 신호의 경우, 바이올린과 같이 현악기 소리는 주로 고조파 성분을 주로 가지고 있으므로 스펙트로그램 상에서 시간축의 방향을 따라 에너지가 모여 있고, 드럼과 같은 타악기 소리는 퍼커시브 성분을 주로 가지고 있으므로 스펙트로그램 상에서 주파수축 방향을 따라 에너지가 집중 되어 있다[15][16]. 음향 효과의 경우에는 그 종류마다 음향 특성이 상이하다. 예를 들어 불꽃놀이 소리, 폭발음, 문 닫는 소리 및 말발굽 소리는 타악기 소리와 유사한 퍼커시브 성분이 많이 포함되어 스펙트로그램 상에서도 쉽게 시간축에 대해 수직적인 구조를 확인할 수 있는 반면, 바람 소리는 음조(tonal) 정보를 전달하므로 고조파 성분으로 인지되지만 스펙트로그램 상에서는 명확한 수평 구조가 없는 경우가 있다[17]. 또한, 백색잡음 신호의 경우에는 스펙트로그램 상에서 시간 축과 주파수 축에 편향되지 않은 등방성으로 에너지가 분포한다.

그림 2 는 드라마 ‘불의 여신 정이’ 중 일부의 오디오 파

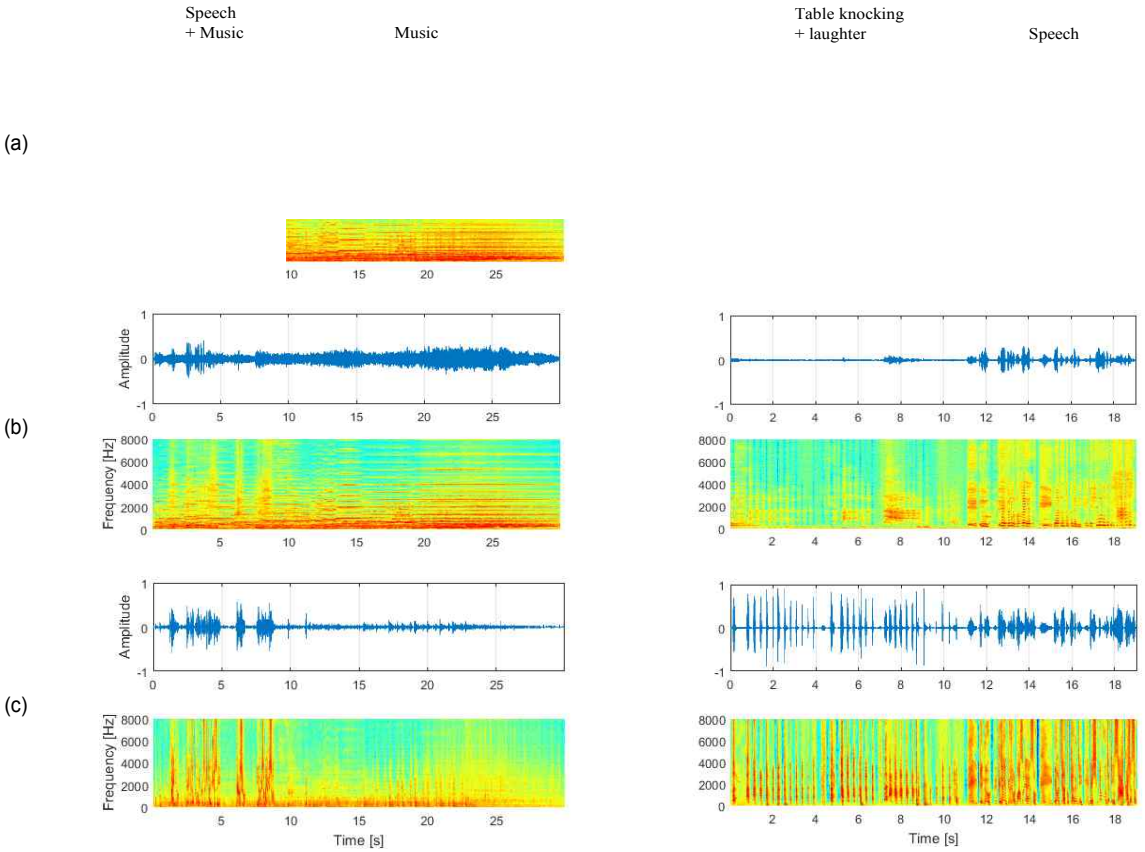

그림 2. 드라마 오디오 신호에 대한 HPSS 결과의 예: (a) 원 미디어 신호 (b) HPSS 적용 후 하모닉 성분 신호 (c) HPSS 적용 후 퍼커시브 성분 신호
Fig. 2 An example of HPSS in drama audio: (a) original media audio; (b) harmonic and (c) percussive components after applying HPSS to media audio shown in (a)

형과 스펙트로그램, 그리고 해당 신호를 고조파 성분과 퍼커시브 성분으로 분해한 예이다. 그림에서 보듯이, 현악기 연주로 이루어진 음악(music) 신호는 주로 고조파 성분으로 구성되어 있으며 테이블 두드리는 소리(table knocking), 배우의 웃음소리(laughter)에 대한 신호는 주로 퍼커시브 성분으로 구성되어 있음을 알 수 있다. 한편, 음성(speech)은 유성음과 무성음의 혼합으로 인해 시간축과 주파수축으로 에너지가 복잡하게 섞여 분포되어 있으며 HPSS (Harmonic-Percussive Source Separation)를 적용한 결과, 고조파와 퍼커시브 성분이 혼재 되어 있음을 알 수 있다. 본 실험은 [15]에 제시된 미디안 필터링 기반 HPSS 알고리즘을 사용하였으며 알고리즘에 대한 자세한 설명은 다음 장에서 기술한다.

이와 같이, 음원의 종류에 따라 고조파 성분과 퍼커시브 성분의 결합 정도가 다르므로 미디어 오디오 신호를 상기 두 성분으로 분해하여 오디오 특징을 추출함으로써 미디어 오디오에 포함되어 있는 음원의 음향 특성을 보다 명확하게 표현할 수 있다. 이를 검증하기 위해 약 20 시간의 드라마 데이터 세트와 8시간의 영화 데이터 세트의 오디오에 HPSS를 적용하여 적용 전과 후 신호의 오디오 특징 값 분포를 살펴보았다. (실험 데이터 세트에 대해서는 Ⅳ 장에서 상세하게 설명한다.) 즉, 모노 오디오 및 그 모노 오디오에 HPSS를 적용하여 얻은 고조파 신호와 퍼커시브 신호 각각으로부터 각각 13-MFCC를 추출하고 정규화 하여 음성 및 비음성에 대한 MFCC값을 비교하였다. 표 1은 MFCC의 차수 별 평균(mean)과 표준 편차(standard deviation; std)이며 확률 밀도 함수 비교의 예는 그림 3에 도시하였다. 표 1과 그림 3에서 보듯이, HPSS를 적용한 후에 음성과 비음성 간

표 1. 드라마 데이터 세트와 영화 데이터 세트 내 음성과 비음성 신호에 대한 MFCC 값의 평균과 표준편차
Table 1. Mean and standard deviation of the MFCC values for speech and non-speech in the drama data set and the movie data set

Order of MFCC	Original				Harmonic				Percussive			
	Speech		Non-speech		Speech		Non-speech		Speech		Non-speech	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	0.5046	1.0825	-0.6766	1.4037	-0.2819	0.9676	-1.2474	1.3514	-0.1259	1.1470	-1.4027	1.3390
1	-0.9155	1.0733	-0.7856	0.9431	-0.6899	0.8591	-0.6164	0.9322	-0.9203	1.1770	-0.8220	0.8618
2	-0.0003	1.0276	-0.0673	0.7742	0.1285	0.8679	-0.0554	0.7754	-0.0379	1.1397	0.0629	0.7294
3	-0.0325	1.0757	0.0347	0.7381	0.1388	0.9397	0.0972	0.7422	-0.0758	1.1627	0.1020	0.6788
4	-0.1964	1.0988	0.0538	0.7904	-0.0888	0.9880	0.1011	0.8164	-0.1880	1.1663	0.1213	0.6918
5	-0.3760	1.1606	0.0475	0.8312	-0.3190	1.0331	0.1119	0.8629	-0.3139	1.2456	0.0735	0.7179
6	-0.0468	1.1458	0.1522	0.8386	0.0172	1.0362	0.2235	0.8706	0.0280	1.1968	0.1502	0.7151
7	-0.2525	1.1064	0.1399	0.8334	-0.2204	1.0387	0.2024	0.8674	-0.1217	1.1238	0.1415	0.7093
8	-0.0985	1.1053	0.2759	0.8356	-0.0983	1.0539	0.3281	0.8752	0.0614	1.1032	0.2842	0.7008
9	-0.0237	1.0509	0.2151	0.8265	-0.0443	1.0138	0.2443	0.8728	0.1634	1.0336	0.2613	0.6806
10	0.1803	0.9822	0.3111	0.7992	0.1675	0.9605	0.3338	0.8507	0.3645	0.9466	0.3576	0.6454
11	0.0861	0.9831	0.3634	0.7666	0.0644	0.9764	0.3797	0.8230	0.2835	0.9137	0.4144	0.6065
12	0.2304	0.8685	0.3963	0.7272	0.2114	0.8651	0.3940	0.7807	0.4188	0.8213	0.4827	0.5666

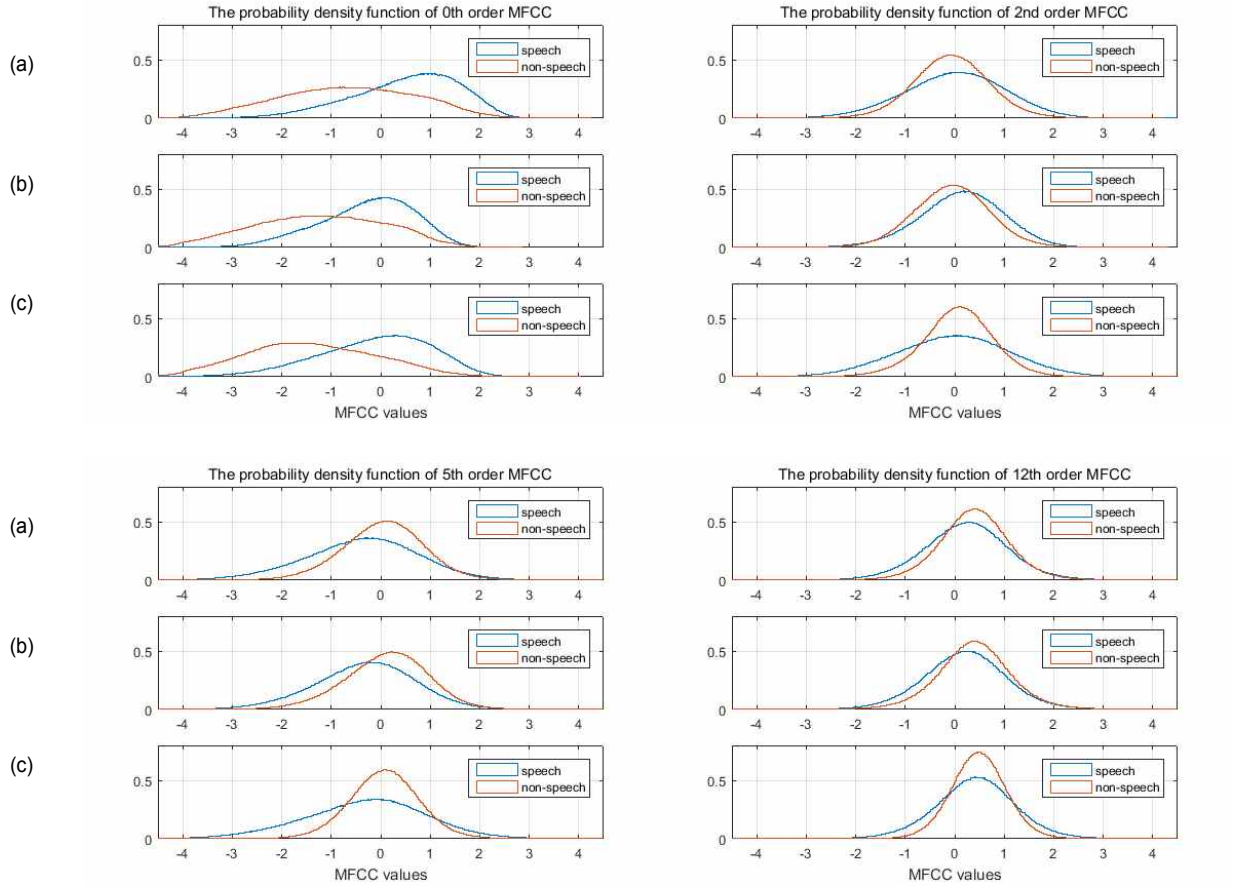

그림 3. 드라마 데이터 세트와 영화 데이터 세트 내 음성과 비음성 신호의 MFCC 값에 대한 PDF 비교의 예: (a) 원 미디어 신호 (b) HPSS 적용 후 하모닉 성분 신호 (c) HPSS 적용 후 퍼커시브 성분 신호
Fig. 3. Examples of the PDF comparison of MFCCs for speech and non-speech in the drama data set and the movie data set: extracted from (a) the original media audio; (b) harmonic and (c) percussive components after applying HPSS

의 확률 밀도 함수 사이의 구별이 명확해짐을 알 수 있으며 특히, 퍼커시브 성분의 경우 그 추이가 더욱 명확해짐을 확인할 수 있다.

Ⅲ. 미디어 음성 검출 시스템

본 장에서는 제안하는 DNN 기반 미디어 음성 검출 시스템을 자세히 설명한다. 전술한 바와 같이 미디어 오디오는 다양한 음원을 포함하며 해당 음원의 종류에 따라 고조파 및 퍼커시브 성분의 결합 정도가 다르다. 본 논문에서 제안하는 방법은 미디어 오디오의 고조파 및 퍼커시브 성분 특성을 활용하기 위하여 각 성분으로 분해하여 오디오 특징을 추출하고 오디오의 문맥 정보를 포함하도록 입력 벡터를 구성하여 DNN 기반의 음성 검출을 수행한다. 그림 4는 제안하는 음성 검출 방법의 구조도이다.

우선, 입력된 미디어 오디오 신호를 HPSS 알고리즘을 이용하여 고조파 및 퍼커시브 성분으로 분해한다. HPSS는 오디오 혼합신호를 고조파 및 퍼커시브 성분으로 분리하는 기술로써, 음악 프로세싱 영역에서 리듬/고조파 악기 전사 및 코드 감지는 물론 리믹스 용도의 전처리 기술로 개발되어왔다. 최근 자동 가사 인식, 자동 가수 판별 및 자동 자막 정렬 등의 분야에서 보컬음성 분리 성능을 향상시키는 요소 기술로 활발히 연구되고 있다[16][18]. 본 논문에서는 미디안 필터 기반의 HPSS 기술을 이용하였다. 이 알고리즘은 고조파 성분과 퍼커시브 성분이 스펙트로그램에서 각각 수평(시간축) 및 수직(주파수축)으로 에너지 성분을 가지는 것에 착안하여 수평 및 수직 방향으로 중간 값 필터링을 이용함으로써 하모닉 및 퍼커시브 성분을 계산하는 방식으로 간단한 구조로 우수한 분리 성능을 제공하는 대표적인 HPSS 알고리즘이다.

제안하는 시스템에서는 16kHz로 다운샘플링한 미디어 오디오 신호에 대해 분석 윈도우 길이는 1024, 홉(hop) 사이즈는 256을 적용한 STFT (Short-Time Fourier Transform)를 수행하여 스펙트로그램을 생성하였다. HPSS를 위한 중간 값 필터의 길이는 31로 설정하였으며 소프트 마스킹을 적용하였다. 그 후 고조파와 퍼커시브 성분 각각으로부터 13차 MFCC 계수를 추출하였다. 음성 검출을 위한 오디오 특징으로 MFCC를 사용한 이유는 계산량을 줄이기 위해서 이다. 즉, 오디오 신호의 시간-주파수 정보를 모두 담고 있는 스펙트로그램 자체를 오디오 특징 벡터로 사용하는 것이 음성 검출 성능 측면에서는 더 바람직하다고 할 수 있으나 DNN 입력벡터 구성 시 문맥 정보를 반영하기 위하여 해당 프레임의 전과 후의 오디오 특징 벡터를 결합하게 되면 입력벡터의 차원 수가 매우 커지므로 DNN의 구조 선정 문제와 계산량을 고려하여 스펙트로그램보다 작은 차수의 MFCC를 오디오 특징으로 사용하였다. 이후 추출된 오디오 특징은 파일 단위로 평균이 0, 분산이 1이 되도록 정규화 하였으며, 미디어 오디오 내 음성 존재에 대한 문맥 정보[8]를 반영하기 위하여 해당 프레임의 직전과 직후 각 5개의 프레임의 오디오 특징 벡터들을 결합하여 DNN 입력 벡터를 구성하였다.

한편, 음성 및 비음성 분류를 위해 DNN을 이용하였으며 본 논문에서 사용한 DNN 구조 및 파라미터는 표 2과 같다. 입력 층과 출력 층의 노드 수는 각각 입력 벡터의 차원 수인 286개와 음성 및 비음성 구분을 위한 출력 벡터의 차원 수인 2개이다. 은닉 층의 노드 수는 입력 층의 노드 수와 동일하게 하였으며 은닉 층의 개수는 총 4개로 구성하였다.

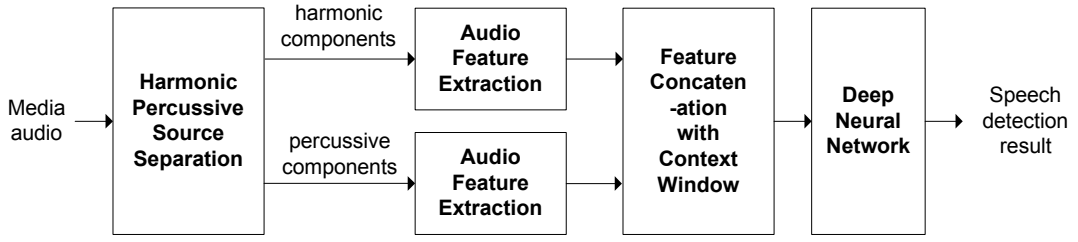

그림 4. 제안하는 음성 검출 시스템 구조도
Fig. 4. Diagram of the proposed speech detection system

DNN 구조 내 은닉 층 개수 선정과 관련해서는 Ⅳ장에서 자세하게 설명한다. 은닉 층과 출력 층의 비선형 활성화 함수는 각각 sigmoid와 softmax를 사용하였다. 미니 배치(minibatch)의 크기는 100, 학습율은 0.01, 모멘텀을 0.9로 하였으며 확률 경사 하강법(stochastic gradient descent)을 이용하여 역전파 알고리즘(back-propagation algorithm)에 의해 DNN을 최대 200 epoch 동안 훈련하였다. 또한, DNN의 과적합(over-adaptation)을 막기 위하여 5 epoch 동안 손실 값의 개선이 이루어지지 않으면 DNN 훈련을 멈추도록 하였다. 이때, 검증 데이터에 대한 목표 값과 DNN이 예측한 값 사이의 손실 값을 평가 기준으로 사용하였다.

표 2. 제안하는 DNN 구조의 파라미터와 셋팅
Table 2. Parameters and settings for the proposed DNN architecture

Input layer setting	286th order of context feature
Output layer setting	2nd order of binary output
Hidden layer setting	[286 286 286 286]
Activation function	sigmoid
Output nonlinear function	softmax
Loss function	cross-entropy
Learning rate	0.01
Momentum	0.9
Optimization	stochastic gradient descent

Ⅳ. 음성 검출용 미디어 오디오 데이터 세트

미디어 오디오에서의 음성 검출용 데이터베이스 중 공개된 경우는 저작권 등의 이슈로 인해 그 수가 매우 적다. 따라서, 본 연구를 위해 자체적으로 드라마 오디오에 대한 음성 검출용 데이터 세트를 구축하였으며 또한 오스트리아 Johannes Kepler 대학의 영화 오디오 데이터 세트를 확보하여 실험에 활용하였다[12]. 본 장에서는 두 미디어 오디오 데이터 세트에 대한 세부 내용을 기술한다.

드라마 오디오에 대한 데이터 세트를 구축하기 위하여 우선, 3 개의 지상파 방송사로부터 MPEG-2 TS로 인코딩된 드라마 방송스트림 파일을 캡처하고 TS 파일로부터 AC3로 인코딩된 (48kHz 샘플링율, 16비트, 스테레오) 오디오 비트스트림의 추출 및 디코딩을 통해 PCM 데이터를 확보하였다. 드라마 프로그램의 특성상 장르에 따라 등장인물의 발성 뿐 만 아니라 사운드 효과, 배경음 등 음원의 유형이 다르므로 사극(역사 드라마)과 현대극으로 장르를 다양화하여 데이터 세트를 구성하였다. 총 30 종류의 드라마 오디오를 확보하였으며 직접 청취를 통해 각 세그먼트에 해당하는 음성/비음성 라벨을 주석으로 제작하였다. [12]에서 언급되었듯이, 라벨링 결과는 주석자가 음성과 비음성을 결정하는 것에 달려 있으므로 라벨을 달기 전에 일관된 규칙을 설정하는 것이 중요하다. 우리는 Lehner의 영화 데이터 세트와 같이 ASR (Automatic Speech Recognition) 시스템을 염두에 두고 드라마 오디오에 주석을 달았으며 숨소리, 비명소리, 웃음소리, 신음소리 및 노래 소리는 비음성으로 라벨링하였다. 결과적으로, 총 20시간 23분 정도의 드라마 데이터 세트를 구성하였으며 이들 중 약 7시간 20분 (35.97 %)이 음성구간이다. 표 3는 드라마 데이터 세트에 대한 자세한 통계를 나타낸다.

표 3. 드라마 데이터 세트의 상세 통계
Table 3. Statistics of the Drama dataset

Type	Number	Length	Speech [%]
Modern	18	10:50:20	35.62
Historical	12	9:32:23	36.36
Total	30	20:22:43	35.97

한편, 앞서 기술한 바와 같이 미디어 오디오에서의 음성 검출용 데이터베이스 중 공개된 경우는 저작권 등의 이슈로 인해 그 수가 매우 적다. 그 중, 본 논문에서 사용한 영화 데이터 세트는 오스트리아 Johannes Kepler 대학의 B. Lehner 등이 Interspeech 2015에서 공개한 음성 검출용 데이터 세트로써 4개의 헐리우드 영화의 음성과 비음성 세그먼트에 대한 라벨 정보이다[12]. 이 데이터 세트는 30분 길이의 청크로 나누어져 있으며 각 청크는 각각 10ms, 20ms, 200ms 단위의 음성 또는 비음성 라벨 데이터로 구성되어 있다. 또한, 이 데이터 세트에는 각 청크 내 음성 및 비음성 세그먼트의 경계 시간과 라벨 정보가 포함되어 있다. 이 데이터 세트에는 오디오 신호 데이터가 포함되어 있지 않으므로 해당 영화에서 오디오 트랙을 추출한 후 주석의 음성 세그먼트 경계를 비교 정렬하여 실험에 사용하였다. 영화

데이터 세트는 총 8시간 이상으로 구성되어 있으며 이들 중 음성은 25.16%의 비율로 존재한다. 표 4는 영화 데이터 세트에 대한 자세한 통계이다.

표 4. 영화 데이터 세트에 대한 상세 통계
Table 4. Statistics of the Movie dataset

Title	Length	Speech [%]
Bourne Identity	1:58:24	26.75
I Am Legend	1:40:22	18.35
Kill Bill 1	1:46:08	19.15
Saving Private Ryan	2:42:27	32.12
Total	8:07:21	25.16

V. 실험 및 결과

본 실험에서는 앞서 설명한 두 종류의 데이터 세트를 이용하여 실제 미디어 오디오에 대한 음성 검출 성능을 검증하였으며 총 네 가지의 실험을 수행하였다. 첫 번째, 제안한 음성 검출 시스템의 DNN 구조를 선정하기 위하여 은닉 층의 개수를 점진적으로 증가시키면서 음성 검출 성능을 측정하였다. 두 번째, 드라마 데이터 세트를 이용하여 5겹 교차 검증(5-fold cross validation)을 수행하여 큰 사이즈의 미디어 오디오에 대해 제안한 음성 검출 방법의 성능을 검증하였다. 세 번째, 드라마 데이터 세트 전체로 훈련한 DNN 기반 음성 검출 시스템을 이용하여 영화 데이터 세트에 대한 음성 검출을 수행함으로써 훈련과 검증에 서로 다른 데이터 세트를 사용한 경우에도 일관된 성능 개선을 보이는지 확인하였다. 마지막으로, DNN 기반의 분류기는 훈련 데이터의 특성에 의존하므로 공개되어 있는 영화 데이터 세트에 대해 LOOCV (Leave-One-Out Cross Validation)을 추가로 수행하였다.

상기 실험들에서는 모노 다운믹스한 미디어 오디오로부터 13-MFCC를 오디오 특징으로 추출하여 이를 입력 값으로 사용한 DNN 기반의 음성검출 방법을 비교군으로 성능을 검증하였다. 이 비교군은 모노 신호로부터, 제안한 방법은 고조파 신호와 퍼커시브 신호로부터 오디오 특징을 추출하여 DNN 입력 벡터를 구성하였으므로 입력 벡터의 차원 수가 각각 143과 286 이다. 이 점을 제외한 나머지의 파라미터와 DNN 구조는 동일하도록 표 2와 같이 설정하여 실험하였다.

한편, DNN을 이용하는데 있어서 가장 큰 어려움은 주어진 작업에 대해 DNN의 세부 구조를 선험적으로 정의할 수 있는 이론적 증거가 없다는 것이다[19]. 이에 본 실험에서는 가장 먼저, 음성 검출 시스템의 은닉 층의 개수를 점진적으로 증가시키면서 DNN의 모델링 성능을 측정함으로써 최적의 은닉 층 개수를 찾는 실험을 수행하였다. 본 실험에서는 드라마 데이터 세트를 이용하여 실험하였으며 이들 중 90%를 훈련 세트로, 10%를 검증 세트로 랜덤하게 각각 구성하여 총 3번의 실험을 수행하였다. DNN의 모델링 성능을 측정하기 위하여 검증 데이터 세트에 대한 손실 값을 측정하였으며 그 결과는 그림 5와 같다. 그림에서 보는 바와 같이, 제안하는 방법과 비교군 모두 은닉 층의 개수가 많아질수록 검증 손실 값이 점점 작아지며, 제안하는 방법의 검증 손실 값이 비교군의 검증 손실 값에 비해 평균 약 73.36% 정도임을 알 수 있었다. 은닉 층 개수가 4개 이상이 되면서 손실 값의 차이가 거의 없으므로 본 연구에서는 성능 및 계산량 등을 고려하여 4개의 은닉 층을 갖는 DNN 구조를 이용하여 이후의 실험들을 진행하였다.

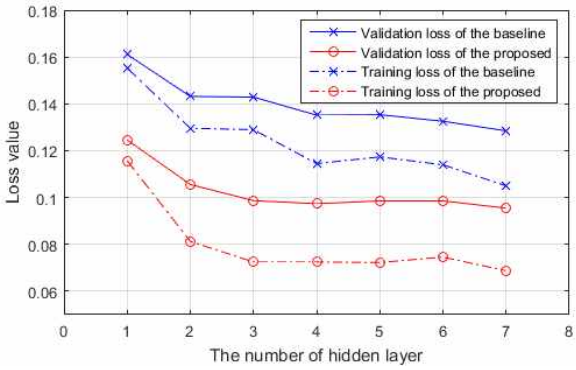

그림 5. DNN 은닉층 개수에 따른 손실 값
Fig. 5. Loss value according to the number of hidden layers in DNN

두 번째 실험으로는 제안하는 음성 검출 시스템을 드라마 데이터 세트에 대해 5겹 교차 검증을 수행하였으며 그 결과는 표 5 같다. 표에서 PREC, REC, F1, ACC는 각각 정밀도, 재현율, F1-점수, 정확도를 의미하며 FPR과 FNR은 각각 거짓 긍정율(false positive rate)과 거짓 부정율

(false negative rate)을 의미한다. 또한, PROP와 BASE는 제안하는 시스템과 비교군으로 사용되는 시스템을 뜻한다. 표에서 보듯이, 제안하는 방법은 베이스라인의 정밀도, 재현율, F1-점수 및 정확도 대비 각각 1.24%, 2.69%, 1.98%, 1.37%의 성능 개선을 이루었으며, 제안하는 방법의 FNR, FPR은 베이스라인 대비 2.69%, 0.63%의 성능 개선이 있었다. 특히, FNR 값의 개선 폭이 컸으며 이는 제안하는 음성 검출 방법이 미디어 오디오에 포함되어 있는 음향 효과, 음악 등의 비음성 요소들을 검출하는데 우수한 성능을 제공함을 의미한다.

표 5. 드라마 데이터 세트에 대한 5겹 교차 검증 결과
Table 5. 5-fold CV results on drama dataset

	PREC	REC	F1	ACC	FPR	FNR
BASE	0.9367	0.9078	0.9220	0.9448	0.0345	0.0922
PROP	0.9491	0.9347	0.9418	0.9585	0.0282	0.0653

드라마 데이터 세트 전체를 이용하여 훈련한 음성 검출 시스템을 이용하여 영화 데이터 세트에 대한 음성 검출 성능을 측정한 결과는 표 6와 같다. 표 내 약어는 표 5와 동일하며 W.Average는 각 영화 오디오의 길이에 대해 가중치를 주어 평균을 산출한 가중치 평균을 의미한다. 표에서 보듯이, 제안하는 방법은 비교군의 정밀도, 재현율, F1-점수 및 정확도 대비 각각 4.87%, 4.90%, 5.53%, 2.00%의 성능 개선을 이루었으며, 제안하는 방법의 FNR, FPR은 베이스라인 대비 4.90%, 1.07%의 성능 개선이 있었다. 이 결과는 표 5와 마찬가지로 제안하는 방법이 더 우수한 음성 검출 성능을 제공함을 보여준다.

한편, 표 6의 결과는 표 5의 드라마 데이터 세트에 대한 교차 검증결과에 비해 그 성능이 저하되었음을 알 수 있는데, 이는 기존의 다른 연구에서도 동일한 현상을 지적했듯이 다른 종류 및 크기의 훈련데이터를 사용함에 의한 것으로 생각된다[9][12]. 하지만, 표 6에서 보듯이 비교군 대비 제안한 방법의 정확도 개선 정도는 2.00%로써 앞의 실험과 유사한 정도로 개선되었음을 확인할 수 있으며 이로써 훈련과 검증에 다른 장르의 미디어 데이터 세트를 사용한 경우에도 일관되게 성능이 개선됨을 알 수 있다.

전술했듯이 DNN 기반의 분류기는 훈련 데이터의 특성에 의존하므로 공개되어 있는 영화 데이터 세트에 대해 LOOCV을 추가로 수행하여 성능 개선 정도를 측정하였다.

표 6. 드라마 데이터 세트로 훈련한 DNN 기반 음성 검출기의 영화 데이터 세트에 대한 음성 검출 결과
Table 6. Results on movie dataset using DNN based speech detection trained with complete drama dataset

Title	PREC		REC		F1		ACC		FPR		FNR	
	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP
Bourne Id.	.8280	.8699	.5891	.6139	.6884	.7199	.8573	.8722	.0447	.0335	.4109	.3861
I Am Leg.	.8263	.8561	.4425	.5109	.5763	.6399	.8806	.8945	.0209	.0193	.5575	.4891
Kill Bill 1	.6033	.7042	.6014	.6627	.6023	.6828	.8479	.8821	.0936	.0659	.3986	.3373
Saving P.	.8902	.9214	.3734	.4201	.5261	.5771	.7839	.8022	.0218	.0170	.6266	.5799
W.Average	.7994	.8481	.4897	.5387	.5925	.6477	.8356	.8556	.0428	.0321	.5103	.4613

표 7. 영화 데이터 세트에 대한 LOOCV 결과
Table 7. LOOCV results on movie dataset

Title	PREC		REC		F1		ACC		FPR		FNR	
	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP	BASE	PROP
Bourne Id.	.6696	.7013	.7088	.7327	.6886	.7166	.8285	.8450	.1277	.1140	.2912	.2673
I Am Leg.	.6490	.6902	.6301	.6688	.6394	.6794	.8696	.8841	.0766	.0675	.3699	.3312
Kill Bill 1	.4728	.5772	.7353	.7650	.5756	.6580	.7923	.8477	.1942	.1327	.2647	.2350
Saving P.	.8458	.8459	.5484	.6058	.6654	.7060	.8228	.8379	.0473	.0522	.4516	.3942
W.Average	.6812	.7202	.6449	.6843	.6461	.6926	.8272	.8513	.1049	.0879	.3551	.3157

그 결과는 표 7과 같다. 표에서 보듯이, 제안하는 방법은 비교군 시스템의 정밀도, 재현율, F1-점수 및 정확도 대비 각각 3.90%, 3.94%, 4.65%, 2.41%의 성능 개선을 이루었으며, 제안하는 방법의 FNR, FPR은 비교군 대비 3.94%, 1.70%의 성능 개선이 있었다. 이 결과 또한 앞의 실험 결과와 같이 일관된 성능 개선을 보여준다.

V. 결 론

본 논문에서는 미디어 오디오에 포함되어 있는 다양한 음원의 음향적 특성을 조사 분석하고 이를 오디오 특징 추출 방법에 반영한 DNN 기반의 음성 검출 시스템을 제안하였다. 이 시스템은 미디어 오디오 신호의 고조파와 퍼커시브 성분으로부터 오디오 특징을 각각 추출하고 해당 프레임과 그의 전후 프레임의 오디오 특징 벡터들을 결합을 통해 입력 벡터를 구성하여 음성 검출을 위한 DNN 모델링을 수행함으로써 미디어 오디오의 음향적·문맥적 특성을 활용한 음성 검출을 수행한다. 제안하는 방법의 성능을 검증하기 위하여 실제 미디어 오디오에 대해 제작된 음성 검출용 데이터 세트를 활용하였다. 교차 검증 결과, 제안하는 방법은 드라마 데이터 세트에 대해 약 95% 이상의 음성 검출율을 보였다. 또한, 영화 데이터 세트에 대한 교차 검증 결과, 기존 방식 대비 약 2.41%의 성능 개선을 보임으로써 그 우수성을 입증하였다.

본 연구는 추후 시각장애인을 위한 화면해설방송 저작 및 분석 등 미디어 서비스 기술에 적용되어 미디어 서비스 기술의 보편화에 일조할 것으로 기대하며, 해당 작업의 목적에 맞는 세그멘테이션을 포함하기 위하여 딥러닝 기반의 모델 구조 확장을 수행할 예정이다.

참 고 문 헌 (References)

[1] D. Lee, S. Kim, and Y. Kay, “A speech recognition system based on a new endpoint estimation method jointly using audio/video informations,” *Journal of Broadcast Engineering*, Vol. 8, No.2, pp.198-203, 2003.

[2] G. Kim, J. Ryu, and N. Cho, “Voice activity detection using motion and variation of intensity in the mouth region," *Journal of Broadcast Engineering*, Vol. 17, No.3, pp.519-528, 2012.

[3] DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[4] T. Hain, P. C. Woodland, “Segmentation and classification of broadcast news audio,” *Proceeding of International Conference on Spoken Language Processing (ICSLP)*, pp. 2727‐2730, 1998.

[5] L. Lu, H. J. Zhang, and S. Z. Li, “Content-based audio classification and segmentation by using support vector machines,” *Multimedia Systems*, Vol. 8, No. 6, pp. 482-492, 2003.

[6] T. L. Nwe and H. Li, “Broadcast news segmentation by audio type analysis,” *Proceeding of 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[7] A. Misra, “Speech/nonspeech segmentation in web video,” *Proceeding of 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012)*, September 9-13, Portland, Oregon, USA, pp. 1977-1980, 2012.

[8] N. Ryant, M. Libeman, J. Yuan, “Speech activity detection on YouTube using deep neural network,” *Proceeding of 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, August 25-29, Lyon, France, pp. 728-731, 2013.

[9] F. Eyben, F. Weninger, S. Squartini and B. Schuller, “Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies,” *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 483-487, 2013.

[10] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, Buckeye Corpus of Conversational Speech (2nd release), Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA, 2007, www.buckeyecorpus. osu.edu (accessed Aug. 18, 2017).

[11] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgrena, and V. Zue, “TIMIT acoustic-phonetic continuous speech corpus,” 1993, https://catalog.ldc.upenn.edu/ldc93s1 (accessed Aug. 18, 2017).

[12] B. Lehner, G. Widmer and R. Sonnleitner, “Improving voice activity detection in movies,” *Proceeding of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, September 6-10, Dresden, Germany, pp. 2942-2946, 2015.

[13] I. Jang, C. Ahn, Y. Jang, “Non-dialog section detection for the descriptive video service contents authoring,” *Journal of Broadcast Engineering*, Vol. 19, No. 3, pp. 296-306, 2014.

[14] I. Jang, C. Ahn, J. Seo, Y. Jang, “Enhanced feature extraction for speech detection in media audio,” *Proceeding of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, August 20-24, Stockholm, Sweden, pp. 479-483, 2017.

[15] D. FitzGerald, “Harmonic/percussive separation using median filtering,” *Proceeding of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.

[16] Chao-ling Hsu , Deliang Wang , Jyh-shing Roger Jang , Ke Hu , “A tandem algorithm for singing pitch extraction and voice separation from music accompaniment,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 5, pp. 1482-1491, 2012.

[17] R. Füg, A. Niedermeier, J. Driedger, S. Disch, M. Müller "Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," *Proceeding of Acoustics, Speech and Signal Processing (ICASSP)*, 2016.
[18] D. FitzGerald and M. Gainza, “Single channel vocal separation using median filtering and factorisation techniques,” *ISAST Transactions on Electronic and Signal Processing*, Vol. 4, No. 1, pp. 62-73, 2010.
[19] S. Leglaive, R. Hennequin, R. Badeau. "Singing voice detection with deep recurrent neural networks," *Proceeding of Acoustics, Speech and Signal Processing (ICASSP)*, pp.121-125, 2015.

저 자 소 개

장 인 선

- 2001년 2월 : 충북대학교 전기전자공학부 정보통신공학 학사
- 2004년 2월 : 포항공과대학교 컴퓨터공학과 석사
- 2004년 8월 ~ 현재 : 한국전자통신연구원 선임연구원
- ORCID : http://orcid.org/0000-0003-2237-2668
- 주관심분야 : 음성/오디오 신호처리, 객체기반 오디오, 복지방송 오디오

안 충 현

- 1985년 2월 : 인하대학교 해양학과 학사
- 1989년 8월 : 인하대학교 해양학과 석사
- 1986년 ~ 1991년 : 한국해양연구소 연구원
- 1995년 3월 : 일본 치바대학교 환경원격탐사센터 박사
- 1995년 3월 ~ 12월 : 일본 치바대학교 정보공학과 연구조수
- 1996년 ~ 현재 : 한국전자통신연구원 책임연구원
- 주관심분야 : 디지털방송 서비스, 실감방송, 감성미디어, 장애인방송, GIS/RS/LBS

서 정 일

- 1994년 2월 : 경북대학교 전자공학과 학사
- 1996년 2월 : 경북대학교 전자공학과 석사
- 2005년 8월 : 경북대학교 전자공학과 박사
- 1998년 3월 ~ 2000년10월 : LG반도체주임연구원
- 2000년 11월 ~ 현재 : 한국전자통신연구원 책임연구원, 테라미디어연구 그룹장
- 2010년 8월 ~ 2011년 7월 : 영국 사우스햄튼대학 방문연구원
- ORCID : http://orcid.org/0000-0001-5131-0939
- 주관심분야 : 실감방송, 영상 신호처리, UWV, 360 비디오, 오디오 신호처리, 멀티모달 인터페이스

장 윤 선

- 1992년 2월 : 경북대학교 전자공학과 학사
- 1994년 2월 : KAIST 전기및전자공학과 석사
- 1999년 2월 : KAIST 전기및전자공학과 박사
- 1999년 2월 ~ 2006년 2월 : 한국전자통신연구원 선임연구원
- 2006년 3월 ~ 현재 : 충남대학교 전자공학과 정교수
- ORCID : http://orcid.org/0000-0001-5698-6233
- 주관심분야 : 음성/오디오 신호처리, 유무선 통신