특집논문 (Special Paper)
방송공학회논문지 제19권 제3호, 2014년 5월 (JBE Vol. 19, No. 3, May 2014)
http://dx.doi.org/10.5909/JBE.2014.19.3.296
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

화면해설방송 저작을 위한 비 대사 구간 검출

장 인 선[a], 안 충 현[a], 장 윤 선[b)‡]

Non-Dialog Section Detection for the Descriptive Video Service Contents Authoring

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

요 약

본 논문에서는 방송 오디오에서로부터 화면해설 삽입을 위한 비 대사 구간 검출 방법을 제시한다. 방송 오디오에서의 대사와 비 대사 구간을 분류하기 위해서는 대사와 배경 음악 등 다양한 종류의 소리가 혼합되어 있는 스테레오 신호로부터 음성 활성 여부의 검출이 우선되어야 한다. 본 논문에서는 방송 오디오 제작과정을 파악함으로써 신호의 채널 특성 분석 결과를 대사 음성 활성 여부 검출에 적용한다. 본 논문에서 제안하는 비 대사 구간 검출 방법은 방송 오디오의 센터채널과 서라운드 성분 간의 에너지 비율을 추가적인 오디오 특징으로 이용하여 센터채널의 음성 활성도와의 결합을 통해 성능 향상을 이루어 낸다. 또한, 실제 화면해설 방송물의 분석을 통해 생성한 규칙 기반의 후처리를 통해 화면해설 삽입이 가능한 비 대사 구간을 검출한다. 이를 실제 방송 컨텐츠를 대상으로 한 실험을 통하여 검증한다.

Abstract

This paper addresses a problem of non-dialog section detection for the DVS authoring, the goal of which is to find meaningful section from the broadcasting audio, where audio description can be inserted. The broadcasting audio involves the presence of various sounds so that it first discriminates between speech and non-speech for each audio frame. Proposed method jointly exploits the inter-channels structure and speech source characteristics of the broadcasting audio whose number of channel is stereo. Also, rule based post-processing is finally applied to detect the non-dialog section whose length is appropriate for audio description. Proposed method provides more accurate detection compared to conventional method. Experimental results on real broadcasting contents show that qualitative superiority of the proposed method.

Keyword : Speech/Non-speech Detection; Descriptive Video Service

a) 한국전자통신연구원 실감방송미디어연구부 감성미디어연구실 (Realistic Broadcasting Media Research Department, 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
※본 연구는 미래창조과학부가 지원한 2014년 정보통신·방송(ICT) 연구개발사업의 연구결과로 수행되었음.
· Manuscript received March 10, 2014 Revised April 30, 2014 Accepted April 30, 2014

Ⅰ. 서 론

미디어 및 방송통신 융합 기술의 발전에 따라 시청자들은 다양한 채널을 통해 많은 컨텐츠를 제공받고 있다. 하지만 시청각 장애인 및 고령자, 다문화 가정과 같이 일반적인 방송 시청에 어려움을 겪는 방송소외계층의 방송 접근성에 대한 보장은 여전히 충분하지 못한 상황이다. 우리나라의 경우, 등록 장애인 수는 2012년 12월 말 기준 251만 1천명으로, 2000년 12월 말 95만 8천명에서 약 162% 증가하였다[1]. 이러한 현상은 선/후천적 장애와 인구 고령화에 기인하는 전 세계적인 추세이며 이들의 방송 접근권을 제고하고 디지털 미디어에 대한 차별 없는 접근성을 제공하는 미디어 복지강화를 위해 다양한 노력이 진행되고 있다[2][3][4]. 우리나라의 경우, 장애인에 대한 안정적이고 체계적인 방송 접근권을 보장하기 위하여 방송통신위원회는 2011년 12월 "장애인방송 편성 및 제공 등 장애인방송 접근권 보장에 관한 고시"를 제정하였다[5]. 고시는 장애인방송 제공 의무대상 사업자, 자막/화면해설/수화통역 방송 등 장애인방송의 편성 비율, 그리고 기술표준 준수 의무화 및 이행시기 등의 세부 이행방법을 포함하고 있다. 이 고시에 따르면 중앙지상파 방송사와 지역지상파 방송사는 각각 2013년과 2015년까지 전체 프로그램 중 자막방송 100%, 수화방송 5%를 편성하여야 하고, 화면해설 방송의 경우에는 각각의 대상사업자별로 2014년과 2015년까지 10%를 편성해야 한다. 보도 및 종합편성채널 사용사업자는 2016년까지 지상파 방송사와 같은 수준의 편성목표를 달성해야 하며 유료방송사업자 중 고시 지정된 사업자는 2016년까지 자막방송 70%, 화면해설 5%, 수화방송 3%에 해당하는 장애인 방송물을 제작/편성해야 한다. 이에 대한 자세한 사항은 표 1과 같다. 이에 따라 장애인 방송을 편성/제공하는 방송 사업자는 2012년 60개 사에서 2013년에서는 153개사로 대폭 늘어나게 되었으며, 시/청각 장애인이 지상파 방송뿐만 아니라 유료방송 채널의 방송 프로그램도 장애인 방송으로 시청할 수 있게 됨으로써 채널 선택권 및 방송 접근권 확대에 크게 기여할 것으로 기대된다.

이러한 정책에 발맞춘 장애인 방송 제작비 지원 증액 등에 따라 중앙지상파의 모든 프로그램이 자막 방송으로 제공 되는 등, 국내 복지 방송 프로그램의 편성 비율이 지속적으로 증가하는 추세이다. 하지만 화면해설 방송의 경우 제작 특성상 시간과 비용이 상당히 소요되는 사전제작 과정을 거쳐야하기 때문에 편성 비율은 여전히 부족한 실정이다.

화면해설 서비스(Descriptive Video Service; DVS)란 TV 프로그램, 영화 등 대중 영상 매체에 대하여 영상물의 원래 내용을 침범하지 않는 범위 내에서 때와 장소의 변화, 등장인물의 표정이나 몸짓 등과 같은 상황 변화적 요소는 물론

표 1. 장애인방송 편성비율 목표치
Table 1. Target organization ratio of the broadcasting for the disabled

Provider		Target Provider	Start	Measure	Target of the final organization ratio(%)			Accomplishment
					Subtitles	DVS	Sign language	
Terrestrial	Mandatory	Center	2012.1.	2012.7.	100	10	5	2013.12. (DVS: 2014.12.)
		Local	2012.1.	2012.7.	100	10	5	2015.12
Pay (Platform)	Mandatory	Satellite (Direct operating channel)	2012.1.	2013.1.	70	7	4	2016.12
	Announcement obligation	SO(Local channel)	2012.1.	2013.1.	70	7	4	2016.12
Pay (Program Provider)	Mandatory	News/General service PP	2012.1.	2013.1.	100	10	5	2016.12
	Announcement obligation	General PP IPTV CP	2012.1.	2013.1.	70	5	3	2016.12

이고 자막이나 그래픽과 같은 시각적 요소들을 시각 장애인들이 인지할 수 있도록 별도의 음성 해설을 제공하는 서비스 이다. 이는 영상물의 내용을 시각장애인이 왜곡 없이 이해할 수 있도록 하여 정안인과 동등하게 내용을 파악하고 즐길 수 있도록 도와주는 복지 서비스의 일환이다.

화면해설 방송물을 제작하는 과정은 다음과 같다. 먼저, 대상이 되는 원 영상물을 선정하고 화면해설방송 작가가 영상물을 분석하여 시각장애인에게 영상의 내용을 충분히 전달할 수 있도록 자막이나 그래픽, 배경과 표정 등 중요한 시각적 요소들에 대하여 대화가 이루어지지 않는 시간에 삽입시킬 화면해설용 대본을 작성한다. 이후 전문 성우의 음성 녹음을 통해 화면해설 음성 데이터를 생성하며, 프로듀서가 원본 오디오와 화면해설 음성 데이터를 믹싱 하여 화면해설용 음성 트랙을 생성한다. 화면해설 음성 데이터는 방송물의 본 흐름을 해치지 않도록 원본 오디오 내 음성이 없는 구간에 삽입한다. 완성된 화면해설용 음성 트랙은 별도의 기록매체에 저장되거나, 방송용 마스터 테이프로 제작한다. 이를 도식화 하면 그림 1과 같다.

상기 과정은 화면해설방송 작가가 미리 방송 영상을 보면서 출연자의 대사가 없는 구간에 삽입 가능한 길이의 화면해설을 작성하는 화면해설 대본 작업을 한 이후에 프로듀서가 다시 프로그램을 일일이 확인하며 대사가 없는 구간에 성우가 더빙한 화면해설 오디오를 믹싱 하는 과정을 거쳐야 하므로 다수의 전문가 인력과 많은 시간적 노력이 소요된다. 또한, 미리 방송 영상을 확보해야 하는 화면해설 제작 과정의 특성상, 드라마 등 특정 장르의 경우 사전 제작이 전무한 국내 방송 환경 탓에 본 방송에서의 화면해설은 거의 찾아보기 힘들며 - 한편의 드라마를 화면해설 방송물로 제작하는 데 보통 24시간 이상 소요됨 - 이러한 현실적인 문제가 화면해설 방송물 제작에 큰 제한 점으로 작용하고 있다. 실제로 우리나라에서 화면해설이 제공되는 드라마는 주로 낮이나 주말에 방영하는 재방송으로 방영되고 있으며 본 방송에서 화면해설이 제공되는 경우는 다큐멘터리나 사전 제작된 시사교양프로그램에 국한되어 있다.

이러한 한계를 극복하기 위한 노력의 대표적인 예로, Miranda 사의 Swift ADePT는 화면해설 컨텐츠 제작을 위한 토털 솔루션이 있으며 이외에도 폴란드와 우리나라의 TTS 기반 화면해설 제작에 대한 연구개발 등이 있다[6][7][8].

본 논문에서는 효율적인 화면해설 제작을 위해 방송 오디오로부터의 비 대사 구간 검출 방법을 제안한다. 본 방법은 다양한 음원이 존재하는 방송 오디오의 채널 간 구조 특성과 음성 신호 특성을 활용하여 화면해설을 삽입할 수 있는 후보 구간을 검출한다. 이는 화면해설 저작과정에서 작가의 화면해설 대본 작성과 프로듀서의 화면해설 오디오 삽입에 활용되어 효율적인 화면해설 오디오의 저작을 가능하게 한다.

이와 유사한 연구 분야로써 음성 활성 검출(Speech Activity Detection), 음성/비 음성 분류(Speech/Non-speech Discrimination), 오디오 세그먼테이션(Audio Segmentation) 등이 있다. 기존의 방식은 신호의 에너지 기반으로 분류하거나 GMM(Gaussian Mixture Model), HMM(Hidden Markov Model), SVM(Support Vector Machine), NN (Neural Network) 등의 분류기를 한 개 혹은 다수 개를 결합하여 분류한다[9][10][11][12]. 전자의 방법은 오디오 신호의 에너지 값과 임계값(threshold)의 비교를 통해 음성 여부를 판단하기 때문에 단순한 구조가 장점이다. 하지만, 방송 오디오 신호와 같이 음성/묵음 이외에 배경 음악/음향 효과/배경 잡음 등 다양한 음원이 존재하는 환경에서는 음성 여부

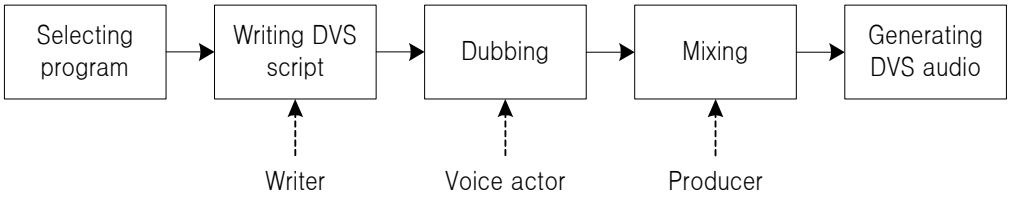

그림 1. 화면해설 제작 과정
Fig. 1. Procedure of the DVS Production

의 판단 율이 저하된다. 후자의 방법은 음성과 비 음성에 대한 DB를 이용하여 오디오 특징을 추출하고 분류기를 미리 학습 시킨 후, 학습된 분류기를 이용하여 입력 오디오 신호의 음성 여부를 판단한다. 이 방법은 전자에 비해 우수한 성능을 제공하지만, 음성/비 음성 여부의 정확한 검출을 위해서는 큰 사이즈의 DB 구축뿐만 아니라 음성/비 음성 신호의 특성을 잘 반영하는 오디오 특징 추출이 전제되어야 한다. 또한, 분류기를 미리 학습시키는 훈련(training) 과정이 선행되어야 한다는 단점이 있다.

본 논문에서 제안하는 방식은 DB 구축과 분류기의 훈련 과정 없이 방송 오디오의 채널 간 구조와 음원 활성 여부를 반영하는 오디오 특징을 이용하여 비 대사 구간 검출을 한다. 이는 분류기 기반의 기존 방식에 비해 간단한 구조를 제공하면서도 에너지 기반의 기존 방식에 비해 높은 성능을 제공한다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 방송 프로그램의 오디오 제작 과정을 설명하고 Ⅲ장에서는 본 논문에서 제안하는 방송 오디오로부터의 비 대사 구간 검출 방법에 대하여 설명한다. Ⅳ장에서는 실제 방송 오디오를 대상으로 한 실험 결과 통해 제안하는 기술이 기존 기술에 비해 비 대사 구간 검출 성능을 향상시킴을 입증하고 그 결과를 분석한다. 마지막으로 Ⅴ장에서는 본 논문에 대한 결론을 맺으며 향후 연구 계획을 제시한다.

Ⅱ. 방송 프로그램 오디오 제작 과정

방송 프로그램은 편성·제작의 편의와 방송국 운영 형편에 따라 여러 가지로 분류되는데 그 중 방송이 실시되는 장소를 기준으로 스튜디오 프로그램과 중계 프로로 나눌 수 있으며, 즉시 방송 여부에 따라 생방송과 녹음/녹화/필름 프로그램으로 구분한다. 방송 프로그램은 프로그램을 기획하고 촬영 준비를 하는 제작-전(Pre-Production) 단계, 실질적인 촬영을 수행하는 제작(Production) 단계, 그리고 편집 및 CG(Computer Graphic) 등 필요한 후속작업을 하는 제작-후(Post-Production) 단계를 거쳐 완성된다. 뉴스와 같은 생방송 프로그램은 제작 단계와 제작-후 단계가 거의 동시에 이루어지나 그 이외의 방송 프로그램은 제작 단계와 제작-후 단계가 순차적으로 이루어진다.

방송 프로그램의 제작과정에서 방송 오디오는 다음과 같이 만들어 진다. 제작 단계에서는 양질의 오디오를 수음하기 위해 다양한 형태와 지향성을 갖는 마이크를 사용한다[13]. 예로, 드라마와 같이 마이크의 등장을 피하기 위하여 어느 정도 떨어져 수음하게 될 때 사람과 카메라의 움직임, 에어컨, 조명의 울림, 멀리의 작업 소리 등을 무시할 정도로 감소시키기 위하여 초지향성 마이크(super-cardioid microphone)를 사용하며 이를 붐(boom)의 끝 부분에 장착하여, 연기자의 움직임에 따라서 상하좌우로 쉽게 움직이며 길이도 조절할 수 있도록 한다. 또 다른 예로, 라벨리에 마이크는 붐 마이크를 쓰기에 불편한 경우 개개인에게 마이크를 달아주어 수음이 잘 되도록 하는 경우에 사용된다. 이는 컨덴서나 다이나믹 전 방향성 마이크로써 주로 사람의 음성을 잡도록 출연자의 옷깃이나 넥타이 등에 끼워 사용하므로 집게 형(clip-on) 마이크, 핀(pin) 마이크라고도 한다. 크기가 매우 작아 영상에 부담을 주지 않으면서 물리적 충격에서 강하고 고음을 잘 흡입해준다. 마이크를 착용하고 있어 마이크 그림자에 신경 쓰지 않고 조명 사용이 가능하며 연기자의 행동에도 그리 불편하지 않아 뉴스나 인터뷰 등 텔레비전 방송 프로그램 제작에 많이 활용된다. 라벨리에 마이크는 입에서 약 15~20cm 떨어져 고정되면 핸드마이크나 붐 마이크처럼 음원과 마이크와의 거리가 고정되어 마이크 이득(gain)을 일정하게 조정하기 쉽다는 장점이 있다. 이와 같이, 제작 단계에서는 프로그램 제작 환경 특성에 맞는 마이크를 이용하여 - 붐 마이크는 마이크와 마이크의 그림자가 화면에 잡히지 않는 선에서, 라벨리에 마이크는 마이크를 착용한 출연자의 모습이 자연스러운 선에서 - 출연자의 음성을 중심으로 수음한다. 따라서 제작 단계에서의 오디오는 마이크에 따라 모노 혹은 스테레오 형태로 녹음되지만 음성을 좌/우 채널에 거의 동일하게 모노 성분으로 담고 있으며 약간의 주변 음을 포함하고 있다.

방송 오디오는 제작-후 과정에서 영상 편집에 따라 삭제되거나 편집되는 한편, 오디오 신호 자체의 편집 - 오디오 신호 증폭/감쇄, 페이드인/아웃(fade in/out) 등 - 을 거치게 된다. 또한 음향효과 및 배경 음악(Background Music;

BGM) 등을 추가하기도 한다. 이는 방송 프로그램에 대한 현실감과 신뢰성을 주는 동시에 시청자의 주의를 집중시키거나 프로그램의 흐름을 더욱 자연스럽게 하는 등의 풍부한 감성을 전달하고자 하는 목적으로써 오디오 측면에서도 주로 전 방위로 음상이 퍼져있는 스테레오 음원을 사용하여 방송오디오가 풍성한 사운드를 제공하도록 한다.

III. 비 대사 구간 검출 방법

본 장에서는 방송 오디오로부터 비 대사 구간을 검출하는 방법을 설명한다. 앞 장에서 설명한 것처럼 방송 오디오에는 음성과 묵음뿐만 아니라 배경 음악, 음향 효과, 배경 잡음 등 다양한 음원이 존재하며 제작/편집 단계의 각 특성상 방송 오디오의 대사 음성의 음상은 주로 센터채널에, 그 이외의 음향은 스테레오 음상 전역에 걸쳐 각각 존재하게 된다. 제안하는 방법은 방송 오디오 신호의 채널 간 구조 특성과 음성 신호의 오디오 특성을 활용하여 방송 오디오로부터 대사 구간의 시작과 끝 시간을 추정하며 이를 기반으로 화면해설 삽입이 가능한 비 대사 구간을 검출한다. 그림 2는 제안하는 비 대사 구간 검출 방법의 구조이다.

1. 전처리

입력신호를 이용하여 센터채널 추출과 다운믹스를 수행한다. 이는 분류하고자 하는 오디오의 특성을 효과적으로 반영하는 특징을 추출하기 위한 전처리 과정이다. 센터채널 추출 모듈은 방송 오디오 신호로부터 센터채널 성분과 서라운드 성분을 분리하여 각각의 신호를 생성한다. 이는 방송 오디오가 스테레오 음상 전반에 걸친 음상을 가지는 배경음 및 음향 효과 음원에 비해 대사 음원은 가운데 음상에 몰려서 위치하는 경향성을 이용해서 주로 대사 음원이 담겨 있는 센터채널 신호와 주로 배경음이 담겨 있는 서라운드 신호를 분리 생성하려는 목적이다.

센터 채널 추출을 위해서는 양 채널간의 음량 차이와 위상 차이 정보를 이용하는 방식이 가장 일반적이다. 먼저 단구간 푸리에 변환(Short-Time Fourier Transform) 등의 시간-주파수 변환 과정을 통해 변환된 신호 중 특정 프레임 t와 특정 주파수 ω에서의 신호 성분을 $X(t,\omega)$라고 할 때, 좌 채널과 우 채널의 차이를 이용하여 양 채널이 가지는 음량의 차이를 제한할 수 있다.

$$|\log_{10}(|X_L(t,\omega)|)-\log_{10}(|X_R(t,\omega)|)|<\lambda \quad (1)$$

여기에서, $|X_L(t,\omega)|$과 $|X_R(t,\omega)|$은 각각 왼쪽 채널과 오른쪽 채널의 시간-주파수 영역 절대 값을 나타내며, λ는 센터 채널 여부를 판별하기 위한 임계값이다. 식 (1)의 부등식이 성립하는 경우의 특정 프레임 t와 주파수 ω의 성분은 센터 채널에 위치한다고 간주한다.

또한 음량 정보 이외에도 위상의 차이를 이용해서, 유사

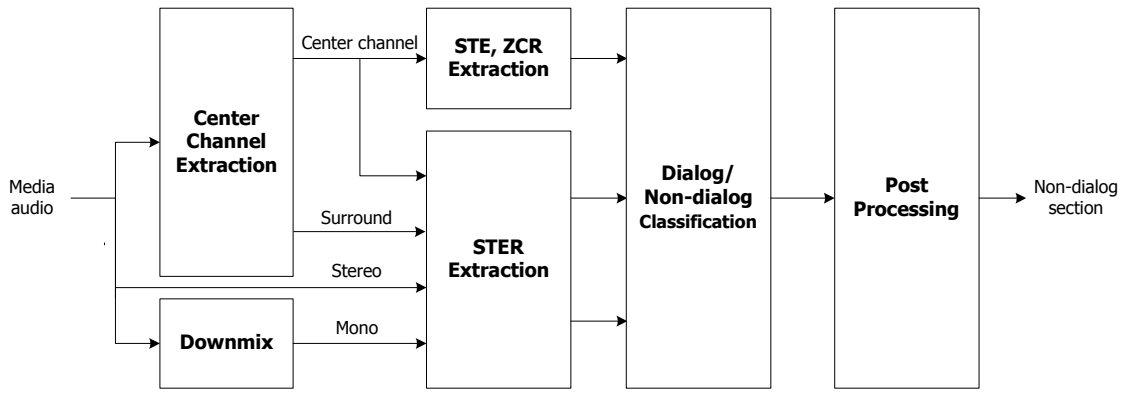

그림 2. 제안하는 비 대사 구간 검출 방법 구조도
Fig 2. Diagram of the proposed non-dialog section detection

한 방식의 판별을 할 수 있다.

$$|\Phi(X_L(t,\omega))-\Phi(X_R(t,\omega))|<\phi \quad (2)$$

여기에서, $\Phi(X)$는 복소수 X의 위상 값을 의미하며, ϕ는 식 (1)의 λ와 마찬가지의 임계값이다.

상기와 같은 시간-주파수 영역 신호의 판별을 통하여 센터채널 추출을 수행하고 이 외의 성분은 서라운드 성분으로써 분리한다.

다운 믹스 모듈은 입력받은 스테레오 방송 오디오 신호의 채널 간 평균값을 계산함으로써 모노 신호를 출력한다.

2. 특징 추출 및 대사/비 대사 분류

높은 분류 성능을 얻기 위해서는 분류의 목적을 잘 반영하는 특징 추출이 우선 되어야 한다. 방송 오디오에는 대사를 위한 음성 신호뿐만 아니라 음향 효과, 배경 음악, 주변 잡음 및 묵음 등 다양한 종류의 신호가 혼재되어 있다. 본 논문에서 제안하는 방법에서는 대사 음성 신호의 음상이 주로 중심에 위치하는 방송 오디오 신호의 채널 특성과 음성 음원 특성을 이용하여 오디오 특징을 추출한다. 즉, 음성의 활성 여부를 판단하기 위해 대사 음원이 주를 이루는 센터 채널 신호로부터 기존의 에너지(Short-Time Energy; STE)과 영점 교차율(Zero Crossing Rate; ZCR) 그리고 센터채널/서라운드 분리 전과 후의 에너지 비율(Short-Time Energy Ratio; STER)을 특징으로 하여 더욱 정교한 비 대사 구간을 검출한다. 센터채널 신호에는 서라운드 신호나 배경 잡음 성분이 적으므로 프레임 간의 스무딩 없이 각각의 프레임에서 상기의 오디오 특징 값을 추출하여 활용한다.

STE는 각 프레임에 대한 신호의 에너지 값을 나타내며 식 (3)과 같이 계산된다. 여기에서 $x(i)$는 이산 오디오 신호이며 N 은 프레임 길이, i는 시간 인덱스(time index)이다. 에너지는 묵음 구간을 추출할 때 사용되는 대표적인 특징으로서, 배경 잡음이 크지 않는 경우에는 보편적으로 음성 구간의 에너지가 묵음 구간보다 크게 추출된다.

$$STE=\sum_{i=0}^{N-1}x^2(i) \quad (3)$$

ZCR은 신호의 부호가 바뀌는 양을 나타내며 식 (4)와 같이 계산된다. 여기에서 $x(i)$는 이산 오디오 신호이며 N 은 프레임 길이, i는 시간 인덱스(time index)이다. 이 값은 파열음(plosive)을 분류하는데 주로 사용되는데 유성음인 경우 ZCR이 낮게 되고, 무성음의 경우 ZCR이 높아진다. 이러한 성질을 이용하여 VAD(Voice Activity Detection)에 활용된다.

$$ZCR=\frac{1}{2(N-1)}\sum_{i=1}^{N-1}|sign[x(i+1)]-sign[x(i)]| \quad (4)$$

전처리에서 생성된 센터채널 신호로부터 각 프레임 별로 추출된 STE와 ZCR 값을 임계값과 비교하여, 두 특징 값 중 하나라도 임계값에 비해 크면 대사에 해당하는 프레임으로 판단하고 그렇지 않은 경우에는 비 대사에 해당하는 프레임으로 판단한다. 여기에서 STE와 ZCR 각각의 임계값은 이산 오디오 신호 전체의 STE, ZCR의 평균값을 이용하였다.

앞서 언급한대로, 거의 대부분의 대사 음원의 음상은 방송 오디오 신호의 중심에 위치한다. 하지만 방송 오디오의 센터채널 신호에는 그 이외에도 음향 효과나 배경 음악 등 다른 신호의 센터채널 성분이 존재한다. 따라서 신호의 활성도를 나타내는 기존의 상기 오디오 특징으로는 여러 신호가 혼재되어 있는 상황에서 대사와 비 대사 프레임을 분류하는데 한계가 있다. 이를 개선하기 위해 본 논문에서 제안하는 방법에서는 방송 오디오 신호의 채널 구조를 나타내는 STER을 오디오 특징으로 활용하였다.

STER은 각 프레임 별로 두 신호의 에너지 값의 비율을 나타내며 다음과 같이 계산된다.

$$STER_{xy}=\frac{\sum_{i=0}^{N-1}x^2(i)}{\sum_{i=0}^{N-1}y^2(i)} \quad (5)$$

여기에서 $x(i)$와 $y(i)$는 각각 센터채널 신호와 (좌 혹은 우) 서라운드 신호이며 N 은 프레임 길이, i는 시간 인덱스

(time index)이다.

방송 오디오 제작 특성 상, 대사 신호에 해당되는 프레임에서는 센터채널 신호의 에너지가 서라운드 신호의 에너지에 비해 크며, 음향 효과나 배경 음악과 같이 편집 작업을 통해 추가된 신호에 해당되는 즉, 비 대사 프레임에서는 서라운드 신호의 에너지가 센터채널 에너지에 비해 크다. 따라서 각 프레임 별로 추출된 센터채널 신호와 서라운드 신호의 에너지 비율 값의 변화에 따라 비 대사 프레임 여부를 더욱 정교하게 판단할 수 있으며 이때 분류를 위한 임계값으로는 방송 오디오의 모노 신호와 스테레오 신호간의 STER 값을 활용한다. 즉, 센터채널/서라운드 간의 STER이 모노/스테레오 간의 STER에 비해 작은 경우, 방송 오디오 신호가 스테레오 음상 중 서라운드에 에너지가 더욱 집중되어 있음을 의미하므로 해당 프레임은 비 대사 프레임으로 분류한다. 반대로, 센터채널/서라운드 간의 STER이 모노/스테레오 간의 STER에 비해 큰 경우, 해당 프레임은 대사 프레임으로 분류한다.

3. 후 처리

프레임 별 비 대사 구간 검출 결과는 후 처리(post-processing)를 통해 동종의 세그먼트로 처리해야 한다. 이는 프레임 별 비 대사 구간 검출 결과가 대사 문장 간의 비 대사 구간뿐만 아니라 문장 내 어절 간, 어절 내 음절 간, 음절 내 음소 간의 길거나 짧은 휴지 구간까지도 비 대사 구간으로 포함하고 있으므로 이들 중 의미 있는 즉, 화면해설 삽입 가능한 비 대사 구간 여부를 판별하기 위해서는 후 처리가 필수적이다.

본 방법에서는 화면해설 저작에 적합한 규칙 기반의 후 처리를 적용하였다. 화면해설 삽입에 적합한 비 대사 구간 길이를 산정하기 위하여 실제 방영되고 있는 화면해설 방송물을 분석한 결과, 일반적인 화면해설의 길이는 2초 이상이었다. 예외적으로 장면 전환 및 짧은 자막에 대한 화면해설의 경우, 1초 이내의 매우 짧은 화면해설을 삽입하는 경우도 있었지만 그 횟수가 매우 적었다. 이러한 분석 결과에 기초한 규칙 기반의 후 처리를 통해 오디오 세그먼테이션(segmentation)을 수행하며 의미 있는 비 대사 구간의 최소 시간을 2초로 설정하여 그 이상의 지속시간을 갖는 비 대사 구간을 검출한다.

Ⅳ. 실험 및 결과

본 장에서는 실제 방송 오디오를 대상으로 제안하는 방법의 우수성을 입증한다. 입력 신호는 MPEG-2 TS(Transport Stream)로 인코딩된 공중파 방송 스트림 캡처 파일로부터 AC3로 인코딩된 (48kHz 샘플링율, 16비트, 스테레오) 오디오 비트스트림을 추출하여 디코딩한 PCM를 사용하였다. 본 실험에서는 총 6 종류의 드라마 오디오로부터 대사 음성, 배경 음악, 음향 효과 등 다양한 소리가 포함된 10분 내외의 구간을 발췌하여 사용하였다. 캡처한 MPEG-2 TS 방송스트림을 분석한 결과, 많은 경우의 방송스트림 내에 최종 방송을 위한 마스터 오디오뿐만 아니라 현장녹음 오디오가 독립된 트랙으로 포함되어 전송됨을 확인하였다. 이 트랙에는 방송 제작 단계에서 마이크로 수음한 대사 음성 신호가 포함되어 있으므로 이를 활용하여 비 대사 구간 검출 성능을 측정하기 위한 참 값(Ground Truth)을 확보하였다. 본 실험에 사용된 방송 오디오와 현장 녹음 오디오 각각의 파형의 예를 각각 그림 3과 그림 4에 나타내었다.

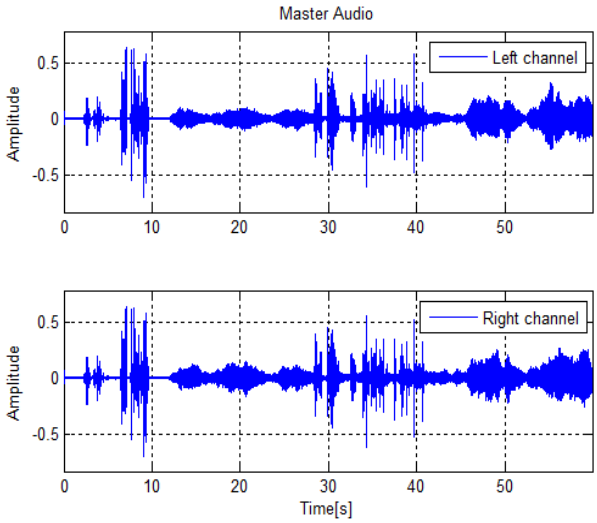

그림 3. MPEG-2 TS로부터 추출된 방송 오디오의 예: 마스터 오디오
Fig 3. An example of the broadcasting audio: Master audio

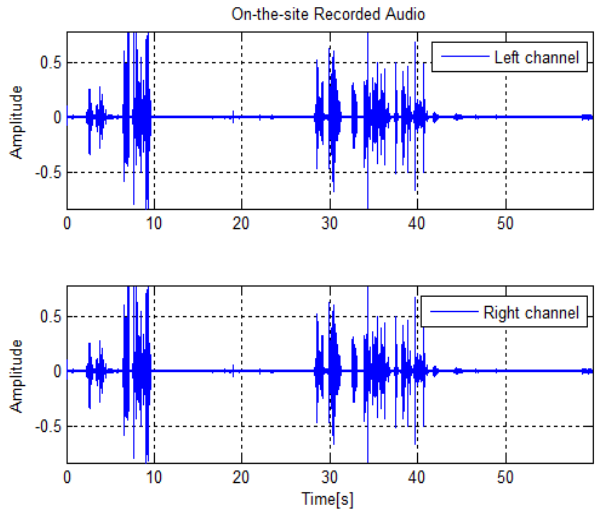

그림 4. MPEG-2 TS로부터 추출된 방송 오디오의 예: 현장녹음 오디오
Fig 4. An example of the broadcasting audio: On-the-site recorded audio

파형에서 볼 수 있듯이, 방송용 마스터 오디오는 현장녹음 오디오를 기반으로 음량 조절과 배경 음악 추가 등의 편집을 거쳐 제작됨을 확인 할 수 있다.

비 대사 구간 검출 성능은 아래의 식을 이용하여 측정하였다.

$$P_{NDSD} = 1 - (MD + FA)/N_{total} \quad (6)$$

여기에서, P_{NDSD}는 비 대사 구간 검출율을 나타나며 MD(Missed Detected non-dialog)는 비 대사에 해당하는 프레임 중 비 대사 프레임으로 분류되지 않은 프레임의 총 개수이다. FA(False Alarm)은 대사 프레임 중 비 대사 프레임으로 분류된 프레임의 총 개수이며 N_{total}은 전체 프레임의 개수이다.

본 실험에서는 STE 및 ZCR을 오디오 특징으로 활용하여 음성 활성도를 측정하고 이를 기반으로 비 대사 구간을 검출하는 기존의 방법과 본 논문에서 제안하는 방법의 성능을 비교하였다.

화면 해설이 삽입되어야 할 구간을 검출하기 위해 50ms 단위로 프레임을 구분하였으며 각각의 프레임에서 오디오 특징을 추출하여 비 대사 검출에 활용하였다. 검출 결과는 표 2와 같다. 방송 오디오인 마스터 오디오 자체에 기존 방법을 적용한 결과 평균 76.56%의 비 대사 구간을 검출하였으며, 방송 오디오로부터 추출한 센터채널 오디오에 STE 및 ZCR을 오디오 특징으로 활용하여 음성 활성도를 측정하고 이를 기반으로 비 대사 구간을 검출한 결과 평균 85.11%의 비 대사 구간을 검출함으로써 전자에 비해 약 8.5% 이상의 성능 향상을 보였다. 이는 방송 오디오의 채널 간 구조 특성을 반영하는 결과로써 방송 오디오 내 대사 음성 신호가 주로 센터채널 음상을 가지고 있음을 보여준다. 본 논문에서 제안하는 방법은 약 92.62%의 비 대사 구간 검출율을 보이며 기존 방법 대비 약 16%, 방송 오디오의 센터채널 음원에 기존 방법을 적용한 결과 대비 약 7.5% 의 성능 향상을 보였다. 이는 방송 오디오의 채널 간 구조를 오디오 특징으로 활용

표 2. 제안하는 비 대사 구간 검출 방법 적용 결과
Table 2. Results of the proposed non-dialog section detection

Soap opera title	Non-dialog section detection ratio [%]		
	Speech activity based detection		Proposed
	Master audio	Center channel audio	
구가의 서	56.81	74.68	88.59
구암 허준	72.46	82.61	96.69
굿닥터	83.28	89.41	92.5
상어	79.97	88.70	90.38
잘났어 정말	82.13	85.93	92.37
불의 여신 정이	84.74	89.37	95.19
Total average	76.56	85.11	92.62

함으로써 대사/비 대사 여부를 더욱 효과적으로 분류할 수 있음을 보여준다.

그림 3의 방송 오디오에 센터채널 추출을 적용한 결과 신호 파형은 그림 5와 같다. 그림 4와의 비교를 통하여 대사 음원이 방송 오디오의 센터채널에 주로 위치하며 또한 배경 음악이나 음향 효과와 같은 대사 이외의 음향은 스테레오 음상 전역에 걸쳐 분포함을 확인 할 수 있다.

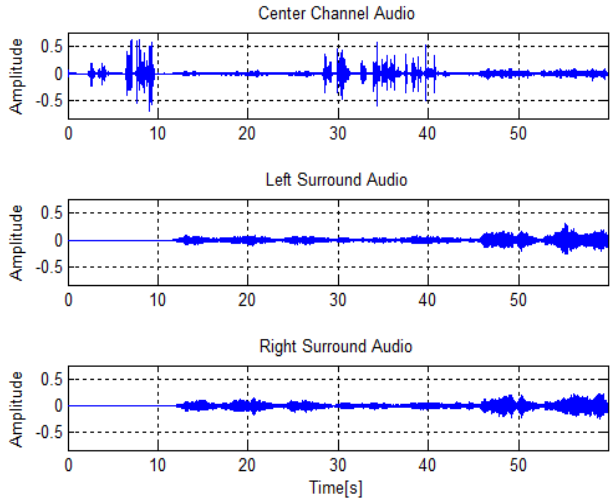

그림 5. 센터채널 추출 결과 신호
Fig. 5. Results of the center channel extraction

상기의 센터채널 신호와 서라운드 신호의 STER은 그림 6과 같다. 첫 번째 그림은 STER 값의 전 범위에 대해 나타낸 것이며 두 번째 그림은 정밀한 분석을 위해 STER 값을 [0, 10]의 범위로 확대한 것이다. 대사 음성만 존재하는 초반 10초 구간에서는 대사 음성의 음상이 거의 센터채널에만 존재하므로 센터채널/서라운드 간의 STER 값이 매우 크며 그 이외의 구간에서는 배경 음악이 주로 서라운드 채널로 분리 되어 STER 값이 작게 나타났다. 특별히, 대사 음성과 배경 음악이 섞여 있는 28초에서 39초 사이의 구간에서는 대사 음성이 주를 이루는 프레임에서는 센터채널/서라운드 간의 STER 값이 매우 크게 나타나고 그 이외의 프레임에서는 작은 값을 나타남을 확인 할 수 있다.

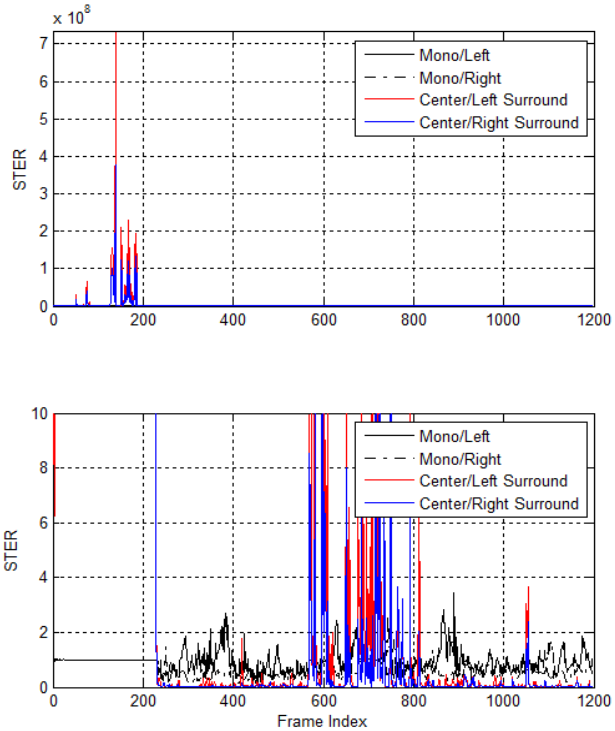

그림 6. 센터채널/서라운드 신호 간의 STER 추출 결과
Fig. 6. STER of the center channel/surround signal

상기의 STER 계산 결과와 센터채널 신호의 프레임별 STE, ZCR 측정 결과를 조합한 다음 화면해설 삽입에 유의

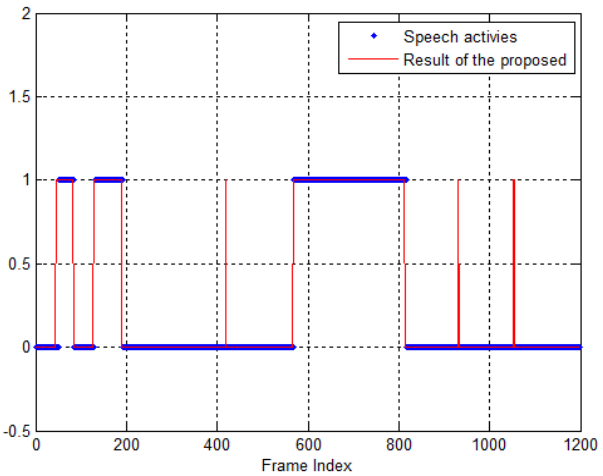

그림 7. 비 대사 구간 검출 결과의 예
Fig. 7. An example of the non-dialog section detection

한 2초 이상의 비 대사 구간을 검출하는 후 처리 결과, 비 대사 구간 검출율은 97.92% 이었다. 그림 7은 상기 예제에 대한 비 대사 검출의 최종 결과이다. 그림에서 '1'과 '0'은 해당 프레임이 각각 대사 프레임과 비 대사 프레임임을 의미한다.

실제 방송 오디오를 대상으로 비 대사 구간 검출 실험 결과, 검출 오류의 원인은 다음과 같았다. 현장 녹음에서 수음된 차 문 닫는 소리, 초인종 소리 등의 음향 효과는 비 대사에 해당함에도 불구하고 음상이 주로 센터채널에 위치하므로 비 대사로 분류되지 않았으며, 보컬이 있는 배경 음악이 삽입된 경우에는 상용 음악에서의 보컬 음상이 일반적으로 센터에 위치하므로 비 대사 구간으로 분류되지 않아서 검출율 저하의 원인이 되었다. 또한, 배경 음악이 있는 상황에서의 대사 음량이 작은 경우, 음성 프레임임에도 불구하고 비 대사 구간으로 분류되는 한계가 있었다.

V. 결 론

본 논문에서는 방송 오디오에서의 비 대사 구간 검출 방법을 제시하였다. 음성과 비 음성을 구분하기 위해 에너지 기반 분류 방식과 높은 차원의 오디오 특징을 기반으로 분류기를 학습하여 분류하는 방식이 개발되어 왔으나, 다양한 음원이 존재하는 방송 오디오에 적용하는데 있어서 각각의 방식은 구조와 성능 측면에서 장단점이 공존하였다. 본 논문에서는 방송 오디오의 채널 간 구조 특성을 활용하여 대사 음성 성분이 모여 있는 센터채널 신호로부터 음성 활성여부를 판단하고 센터채널과 서라운드 신호 간의 에너지 비율을 활용하여 비 대사 여부를 판단함으로써 DB 구축과 분류기의 훈련 과정에 대한 수고 없이 낮은 차원의 오디오 특징을 활용하여 보다 높은 성능의 비 대사 구간을 검출하는 방법을 제시하였다. 제안된 기술은 실제 방송 컨텐츠를 이용한 실험에서 92% 이상의 검출율을 보였으며 기존 기술과의 성능 비교를 통해 그 우수성이 검증되었다.

본 연구는 화면해설방송 제작을 위해서 기존에 소요되었던 인적, 시간적 노력을 줄이고 화면해설 방송물의 양적 증가를 이루데 일조할 것으로 기대되며 향후 더욱 정교한 비 대사 구간 추출을 위해 다양한 오디오 특징에 대한 연구와 비디오 처리 기술과의 접목 등을 수행할 예정 이다.

참 고 문 헌 (References)

[1] Korea Employment Agency for the Disabled, 2013 the disables statistics, Ministry of Employment and Labor, April 2013.
[2] M. Park, ITU Activities for improving ICT accessibility of disabled people, Policy of Broadcasting and Telecommunication, vol 25, no. 12, July 2013.
[3] ITU-T BT.2207-2 (11/2012) Accessibility to broadcasting services for persons with disabilities. (http://www.itu.int/pub/R-REP-BT.2207-2-2012)
[4] Korean Association for Broadcasting & Telecommunication Studies, Study on improving the media accessibility of broadcasting alienation class including the blind and the deaf, Korea Communications Commission, Dec. 2010.
[5] Korea Communications Commission Announcement issue 2011-53, "Announcement of broadcasting access right guarantee for the disabled, which is including organizing and providing the broadcasting for the disabled, " Dec. 2011.
[6] http://www.miranda.com/family/12/Audio_or_Video_Description
[7] A. Szarkowska, "Text-to-speech audio description: towards wider availability of AD", Journal of Specialised Translation 15, pp. 142-163, 2011.
[8] W. Lim, C. Ahn, "Descriptive video service using text to speech," in Proc. Conference of the Korean Society of Broadcast Engineers, June 2013.
[9] B. Elizalde, G. Friedland, "Lost in segmentation: three approaches for speech/non-speech detection in consumer-produced videos," in Proc. ICME, SanJose, USA, July 2013.
[10] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselý, and P.l Matejka, "Developing a speech activity detection system for the DARPA RATS program," in Proc. Interspeech, 2012.
[11] H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," in Proc. ICASSP, pp. II 5-8, 2003.
[12] L. Lu, S. Li, and H. J. Zhang, "Content-based audio segmentation using support vector machines," in Proc. ICME, pp. 749 - 752, 2001.
[13] G. Jung, Management of TV System and Image Production, Cheongmoongak publishing co., 2009.

저 자 소 개

장 인 선

- 2001년 2월 : 충북대학교 전기전자공학부 정보통신공학 학사
- 2004년 2월 : 포항공과대학교 컴퓨터공학과 석사
- 2004년 8월 ~ 현재 : 한국전자통신연구원 선임연구원
- 주관심분야 : 음성/오디오 신호처리, 객체기반 오디오, 장애인 방송 오디오

안 충 현

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

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 부교수
- 주관심분야 : 음성/오디오 신호처리, 유무선 통신