화면해설방송을 위한 오디오/자막 기반의 무 대사 구간 검출

장인선, 임우택, 안충현
한국전자통신연구원 방송통신미디어연구부문
{jinsn, wtlim, hyun}@etri.re.kr

Audio/Subtitles based Non-Dialog Section Detection for DVS

Inseon Jang, Wootaek Lim, ChungHyun Ahn
Electronics and Telecommunications Research Institute

요 약

화면해설방송이란 시각장애인들이 TV 프로그램, 영화와 같은 미디어에 접근할 수 있도록 해주는 서비스로써 화면을 볼 수 없는 시각장애인들을 위해 상황 변화적 요소와 자막, 그래픽 등의 시각적 요소들을 설명하여 프로그램 내용의 이해를 도와주는 서비스이다. 이러한 화면해설은 대사나 효과음이 없는 부분에 전체 프로그램의 이해를 저해하지 않는 수준에서 삽입된다. 본 논문에서는 화면해설방송 제작을 위한 무 대사 구간 검출 방법을 제안한다. 본 방법은 방송스트림에 포함되어 있는 오디오와 자막 정보를 분석함으로써 화면해설을 삽입할 수 있는 구간을 검출한다. 실제 방송컨텐츠를 이용한 실험을 통해 본 방법을 검증하고 성능을 확인한다.

1. 서론

미디어 및 방송통신 융합 기술의 발달에 따라 일반 사용자들은 더욱 다양하고 많은 미디어 컨텐츠를 소비하고 있다. 하지만, 디지털 디바이드는 더욱 심화되고 있으며 이는 시청각 장애인 및 고령자, 다문화가정과 같이 정상적인 방송시청에 어려움을 겪는 소외계층에서 더 뚜렷하게 나타난다. 이러한 소외계층의 방송접근권 향상을 위해 다양한 노력이 진행되고 있다[1][2]. 일 예로, 장애인에 대한 안정적이고 체계적인 방송접근권을 보장하기 위하여 방송통신위원회는 2011 년 12 월 “장애인방송 편성 및 제공 등 장애인방송 접근권 보장에 관한 고시”를 제정하였다[3]. 고시는 장애인방송 제공 의무 대상 사업자, 장애인방송의 편성 비율, 그리고 기술표준 준수 의무화 및 이행시기 등 세부 이행방법을 포함하고 있다. 이 고시에 따르면 중앙지상파 방송사와 보도 및 종합편성 채널에서는 각각 2014 년과 2016 년까지 전체 방송 프로그램의 10%를 화면해설방송으로 편성하도록 의무화되었으며 케이블 및 IPTV 사업자에도 2016 년까지 5~7%의 의무편성이 규정되었다.

장애인 방송 제작비 지원에 따라 국내 화면해설 방송의 편성 비율은 지속적으로 증가하는 추세이나 여전히 부족한 실정이다. 이는 화면해설 제작 특성상 시간과 비용이 상당히 소요되기 때문이다. 즉, 화면해설 방송을 제작하기 위해서는 전문 작가가 미리 화면해설 방송용 대본을 작성하고, 성우는 화면해설 대본을 녹음하며, 프로듀서는 이를 이용하여 원본 오디오에 믹싱하는 작업이 필요하다.

화면해설 제작에 소요되는 인적, 시간적 노력을 줄이고 화면해설 컨텐츠의 양적 증가를 통해 시각 장애인들의 방송 접근성을 향상시키기 위한 연구가 진행되고 있다[4][4].

본 논문에서는 화면해설 제작을 위한 무 대사 구간 검출 방법을 제안한다. 본 방법은 MPEG-2 TS 방송스트림에 포함되어 있는 오디오와 자막 정보를 분석함으로써 화면해설을 삽입할 수 있는 구간을 검출하며 이는 화면해설 대본 작성 혹은 화면해설 삽입 시 활용되어 효율적인 화면해설 컨텐츠 저작을 가능하게 한다.

본 논문의 구성은 다음과 같다. 2 절에서는 제안하는 오디오와 자막 분석을 통한 무 대사 구간 검출 방법에 대하여 설명하고 3 절에서는 제안한 방법을 이용한 실험을 통해 그 성능을 확인한다. 마지막으로, 4 절에서는 본 논문에 대한 결론을 맺는다.

2. 무 대사 구간 추출 방법

가. 오디오 기반 무 대사 구간 추출 방법

AC-3 디코더에서 출력된 PCM 데이터를 입력 받아 매 프레임마다 대사 존재 여부를 판단하기 위한 특징을 추출하였다. 한 프레임의 길이는 50ms 이며, 오디오 특징으로는 STE (Short Time Energy)와 ZCR (Zero Crossing Rate)을 사용하였다. STE 는 각 프레임에 대한 신호의 크기 값을 나타낸다. 에너지는 묵음 구간을 추출할 때 사용되는 대표적인 특징 값으로서, 배경 잡음이 크지 않는 경우에는 보편적으로 음성 구간의 에너지가 묵음구간보다 크게 추출된다. ZCR 은 신호의 부호가 바뀌는 양을 나타낸다. 이 값은 파열음(percussive sound)을 분류하는데 주로 사용되는데 유성음인 경우 ZCR 이 낮게 되고, 무성음의 경우 ZCR 이 높아진다. 이러한 성질을 이용하여 VAD (Voice Activity Detection)에 활용된다.

각각의 프레임 별로 추출 된 두 가지 특징 값을 바탕으로 묵음구간을 판별할 임계값(threshold)을 계산한다. 임계값은 임계값 계산을 위한 설정 구간 내 각 특징 값들의 평균값으로 설정하였다. 그 후 매 프레임마다 계산된 STE 와 ZCR 값을 각 임계값과 비교하며, 두 특징값 중 하나라도 임계값에 비해 크면

대사에 해당하는 프레임으로 판단하고, 그렇지 않은 경우에는 무 대사에 해당하는 프레임으로 판단한다.

이렇게 검출된 결과는 후 처리를 통해 동종의 세그먼트로 처리된다. 이는 상기 방법이 대사의 문장 내 어절 간, 어절 내 음절 간, 음절 내 음소 간의 길거나 짧은 휴지 구간까지도 대사 여부를 판별하므로 의미 있는 무 대사 구간 검출을 위해서는 후 처리가 필수적이다. 화면해설방송을 분석한 결과, 일반적인 화면해설의 길이는 2 초 이상이며 예외적으로, 장면 전환에 대한 해설의 경우 1 초 이하의 매우 짧은 길이의 화면해설을 삽입하는 경우도 있지만 그 횟수가 매우 적었다. 이러한 사실에 기초하여 규칙 기반의 후 처리를 통해 오디오 세그멘테이션(segmentation)을 수행하며 의미 있는 무 대사 구간의 최소 시간을 2 초로 설정하여 그 이상의 지속시간을 갖는 무 대사 구간을 검출한다.

나. 자막 기반 무 대사 구간 추출 방법

자막 방송을 위한 텍스트 데이터는 비디오 TS 패킷 내 비디오 ES (Elementary Stream) 의 Picture User Data 영역에 저장되어 전송된다. 따라서 해당 부분을 분석하여 자막 텍스트를 추출하고 관련 PES 헤더의 PTS (Presentation Time Stamp) 정보를 추출하여 무 대사 구간을 검출한다.

자막 텍스트 추출은 문장 단위로 수행하며 한 문장을 계산할 때 아래와 같이 2 가지의 방법으로 문장을 구분한다.

i. 문장 끝 부호가 있는 경우

자막데이터 버퍼를 임의로 설정하고 버퍼에 .!? 등과 같은 문장의 마지막을 의미하는 캐릭터가 존재할 경우, 현재 버퍼에 저장된 문자들과 첫번째 문자를 포함하는 TS 패킷의 PTS 그리고 문장 끝 부호 캐릭터를 포함하는 TS 패킷의 PTS 을 알 수 있다. 따라서, 이를 기반으로 첫 캐릭터를 포함하는 TS 패킷의 상대 시간과 문장 끝 부호를 포함하는 TS 패킷의 상대 시간 그리고 문장 전체를 반환한다.

ii. 임의의 버퍼 사이즈를 초과한 경우

자막데이터 버퍼를 임의로 설정하여 버퍼가 모두 찼을 경우 버퍼의 첫번째 문자를 포함하는 TS 패킷의 상대 시간 즉, (현재 문자의 패킷의 PTS – 첫 자막 패킷의 PTS)과 버퍼의 마지막 문자를 포함하는 TS 패킷의 상대 시간, 그리고 문자의 집합을 반환한다.

임의로 설정된 버퍼 사이즈를 크게 늘린다면, 문장 끝 부호 캐릭터가 존재하기 전까지의 모든 자막데이터를 모으는 중에 문장 끝 부호 캐릭터가 나타나게 되며 이 경우, 버퍼 첫 캐릭터를 포함하는 TS 패킷의 상대 시간과 문장 끝 부호 캐릭터를 포함하는 TS 패킷의 상대 시간 그리고, 문장 전체를 반환할 수 있다.

3. 실험 및 결과

MPEG-2 TS 방송스트림을 캡쳐하여 분석한 결과, 많은 경우 마스터 오디오 트랙뿐만 아니라 대사 오디오 트랙이 포함되어 전송됨을 확인할 수 있었다. 이 오디오 트랙의 경우 배경 음악 등 스튜디오에서의 믹싱 이전의 현장 녹음 오디오이므로 주로 대사 성분이 포함되어 있다.

본 실험에서는 자막 방송 중이며 대사 오디오 트랙이 포함된 공중파 드라마의 방송스트림의 저장하여 사용하였으며, 총 3 가지의 컨텐츠의 3 분 분량에 대하여 오디오 및 자막 기반의 무 대사 구간 검출을 수행하였다. 화면해설을 위한 유효 무 대사 구간 검출을 위해 오디오의 경우, 2 초 이상의 무대사 구간 검출을 수행하였으며 자막의 경우에는 3 초 이하의 오차는 무시하였다.

실험 결과는 표 1 과 같다. 오디오 분석을 통한 무대사 구간 검출에서는 대사 오디오 트랙에 포함된 음향 효과와 음량이 작은 음성으로 인해 오차가 발생하였으며, 자막 분석을 통한 무 대사 구간 검출에서는 PTS 즉, 화면에 자막이 디스플레이 되는 시간 정보로 무 대사 구간을 검출하므로 실제 대사와의 시간 오차가 발생하였다.

표 1. 오디오/자막 분석을 통한 무 대사 구간 검출 결과

	무 대사 구간 검출율 [%]	
	시작 시간	끝 시간
오디오	80.56	86.11
자막	80.28	97.53

4. 결론

본 논문에서는 화면해설방송 저작을 효율적으로 하기 위해 방송컨텐츠 내 오디오 및 자막 분석을 통하여 무 대사 구간을 검출하는 방법을 제시하였으며 실제 방송컨텐츠를 이용한 실험을 통해 그 우수성을 검증하였다.

본 연구는 화면해설방송 제작을 위해서 기존에 소요되었던 인적, 시간적 노력을 줄이고 화면해설 컨텐츠의 양적 증가를 이루데 일조할 것으로 기대되며 향후 좀 더 정교한 무 대사 구간 추출과 마스터 오디오 트랙 기반의 무 대사 구간 추출에 대한 연구를 수행할 예정이다.

감사의 글

본 연구는 미래창조과학부가 지원한 2013 년 정보통신 · 방송(ICT) 연구개발사업의 연구결과로 수행되었음.

참고문헌

[1] 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)

[2] http://www.itu.int/en/ITU-T/jca/ahf

[3] 방송통신위원회고시 제 2011-53 호, “장애인방송 편성 및 제공 등 장애인 방송접근권 보장에 관한 고시”, 2011.12.26.

[4] Szarkowska, Agnieszka, “Text-to-speech audio description: towards wider availability of AD”. Journal of Specialised Translation 15, 142-163, 2011.

[5] 임우택, 안충현, “TTS 를 이용한 화면해설 방송 제작 방법,” 한국방송공학회 하계학술대회, 2013.