TTS 를 이용한 화면해설 방송 제작 방법

임우택, 양승준. 안충현
한국전자통신연구원
wtlim@etri.re.kr

Descriptive Video Service using Text to Speech

Wootaek Lim, Seung-Jun Yang, ChungHyun AHN
Electronics and Telecommunications Research Institute

요 약

본 논문에서는 기존의 화면해설 방송 제작 방법을 보완하기 위한 TTS 를 이용한 화면해설 방송 제작 방법을 제안한다. 우선 화면해설 방송이 삽입 될 수 있는 구간을 검출하기 위해 에너지 값과 스펙트럼 도심 값을 이용하여 묵음구간을 검출하고 검출된 구간에 TTS 를 이용하여 화면 해설을 삽입하였다. 제안한 방법을 이용하면 기존의 화면해설 방송 제작에 소요되는 인적, 시간적 노력을 줄일 수 있을 뿐만 아니라 화면해설 방송 콘텐츠의 양적 증가를 통해 시각 장애인들의 방송 접근성을 향상시키는 효과를 가져올 수 있다.

1. 서론

최근 방송통신융합 환경과 미디어 기술의 발달은 시청자로 하여금 더욱 더 다양한 콘텐츠를 제공받고 선택할 수 있도록 한다. 그러나 이와 같이 방송 환경이 발전함에 따라 장애인들의 디지털 디바이드(digital divide)는 심화되게 된다. 이러한 정보 격차를 줄이고 모든 사람들에게 보편적 서비스를 제공하기 위한 노력이 국내는 물론 미국, 영국, 프랑스, 일본 등 여러 나라에서 진행되고 있다. 국내의 장애인 방송 편성 비율 또한 장애인 방송 제작비 지원과 함께 현재까지 지속적으로 증가하는 추세이나 여전히 부족하며, 특히 화면해설 방송의 경우 제작에 시간과 비용이 상당히 소요되어 아직까지 편성률이 미미한 실정이다 [1].

이상의 문제점을 보완하기 위하여 본 논문에서는 TTS 를 이용한 화면해설 방송 제작방법을 제안한다. 먼저 화면해설 방송 삽입이 가능한 구간을 검출하기 위해 오디오 신호에서 묵음 구간을 검출하고, 검출된 구간을 판단하여 선택적으로 화면해설을 삽입한다. 이 때 화면 해설은 TTS 를 이용하여 삽임 함으로서 결과적으로 화면해설 방송을 작가 혼자서 제작할 수 있는 방안을 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 기존의 화면해설 방송 제작방법에 대해 살펴본 후, 3 절에서는 제안하는 방법에 대하여 설명한다. 4 절에서는 제안한 방법을 이용한 실험을 통해 그 가능성을 확인하고, 5 절에서는 본 논문에 대한 결론을 맺는다.

2. 기존의 화면해설 방송 제작 방법

화면해설 방송이란 시각장애인들이 TV 프로그램, 영화와 같은 미디어에 접근할 수 있도록 해주는 서비스 이다. 즉, 화면을 볼 수 없는 시각장애인들을 위해 자막, 배우들의 행동, 배경 등의 화면 변화 요소를 설명 함으로서 프로그램의 내용을 이해할 수 있도록 도와주는 서비스 이다 [2]. 이러한 화면해설 방송은 대사나 효과음이 없는 부분에 전체 프로그램의 이해를 저해하지 않는 수준에서 삽입된다.

기존의 화면해설 방송 제작은 다음과 같은 절차로 진행된다. 먼저 화면 해설이 필요한 프로그램이 선정되면, 전문적인 작가가 프로그램의 내용을 전달할 수 있는 화면, 배경, 배우들의 동작, 표정 등의 중요한 시각적 요소들을 바탕으로 대본을 작성한다. 이렇게 작성된 화면 해설 대본을 전문 성우가 음성으로 녹음하여 오리지널 오디오와 합성된 화면해설 방송용 오디오 트랙이 만들어 진다. 이러한 합성 작업이 끝나면 최종적으로 방송으로 송출된다. 이 과정은 전문적인 화면해설 작가가 미리 프로그램을 보면서 대본 작업을 한 이후에, 성우와 작가가 다시 프로그램을 확인하며 대사가 없는 구간에 화면 해설 방송을 녹음하는 과정을 거친다. 이는 인적, 시간적 노력이 많이 소요되며 현실적으로 화면 해설 방송이 보급화 되는 데에 큰 제한 점으로 작용한다.

3. 제안하는 방법

현재 제작되는 화면해설 방송의 경우 전문적인 작가가 미리 화면해설 방송용 대본을 작성하고, 작성된 대본을 바탕으로 성우가 재 녹음해야 하는 번거로운 과정을 거친다. 이러한 문제점을 보완하기 위해 본 논문에서는 TTS 를 이용한 화면해설 방송 제작 방법을 제안한다. 먼저 화면 해설이 삽입되어야 할 구간을 검출하기 위하여 50ms 단위로 프레임을 구분하였다. 구분한 각각의 프레임에서 신호의 에너지와 스펙트럼 도심 값을 추출하여 묵음구간을 검출하고, 검출된 구간 중 화면해설이 가능한 구간을 선택하여 TTS 합성을 통해

쉽게 화면 해설방송 제작이 가능함을 확인하였다.

가. 묵음구간 검출(Silence Detection)

화면해설 방송은 대사가 없는 구간에 작가와 성우의 주관적인 판단을 통해 해설을 삽입하게 된다. 이러한 번거로움을 보완하기 위하여 화면해설 삽입이 가능한 묵음구간을 검출하는 과정을 수행한다. 묵음 구간을 검출하는 방법으로는 신호의 에너지와 스펙트럼 도심 특징 값을 이용한 방법을 사용하였다.

신호의 에너지(Signal Energy)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. i-번째 프레임의 신호를 x_i 라고 할 때, 신호의 에너지는 식 1 과 같이 계산된다.

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|x_i(n)|^2 \quad (1)$$

스펙트럼 도심(Spectral Centroid)은 스펙트럼 에너지 분포의 무게중심을 나타내는 특징 값으로 스펙트럼 에너지의 대부분이 집중하는 주파수 영역을 결정한다. 따라서 스펙트럼 도심 값은 소리의 밝기를 나타내는 척도가 되고 시간 축에서의 ZCR(Zero Crossing Rate) 특성과 밀접한 관련을 갖는다 [3]. 스펙트럼 도심은 식 2 와 같이 계산되며, 이때 $X_i(k)$는 i-번째 프레임의 DFT 계수이고 N 은 프레임의 길이이다. 일반적으로 음성이 있는 구간에서 스펙트럼 도심은 높은 값을 갖는다.

$$C_i = \frac{\sum_{k=1}^{N}(k+1)X_i(k)}{\sum_{k=1}^{N}X_i(k)} \quad (2)$$

이렇게 각각의 프레임 별로 추출 된 두 가지 특징 값을 바탕으로 후처리 과정을 통해 묵음구간을 판별할 임계값(Threshold)을 설정한다.

나. Text to Speech

TTS(Text to Speech)란 글자, 문장, 숫자 등의 정보를 음성합성을 통해 사람의 발성과 유사하게 재생시켜 주는 것이다. 본 연구에서는 구글에서 제공하는 음성합성을 이용하여 실험을 진행하였다.

4. 결과

실험에 사용한 샘플은 방영중인 드라마 콘텐츠의 일부를 발췌하여 사용하였으며 그림 1 은 사용한 샘플의 시간 축 신호이다. 이 샘플을 바탕으로 신호의 에너지와 스펙트럼 도심을 계산한 결과는 각각 그림 2,3 과 같으며 묵음 구간 검출 결과는 그림 4 와 같다. 그림 5 는 선택 구간에 TTS 를 삽입한 신호로 화면 해설이 효과적으로 삽입된 것을 확인할 수 있다.

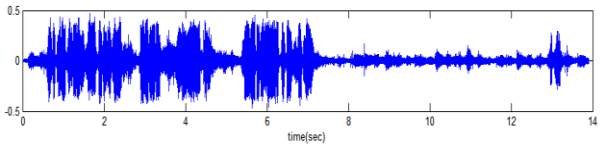

그림 1. 실험에 사용 된 샘플 오디오 신호

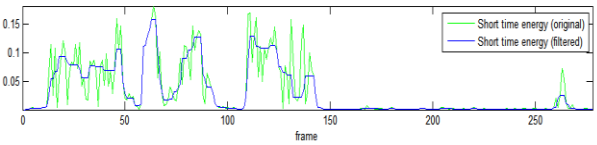

그림 2. 에너지 값 추출 결과

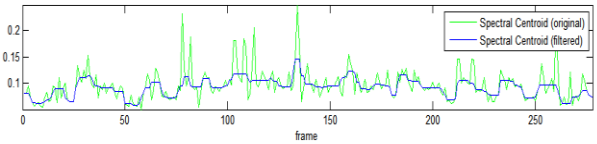

그림 3. 스펙트럼 도심 값 추출 결과

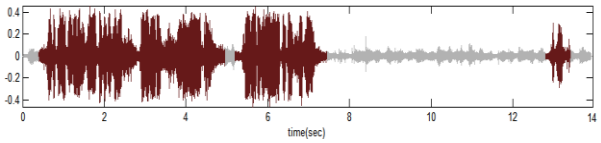

그림 4. 묵음구간 검출 결과

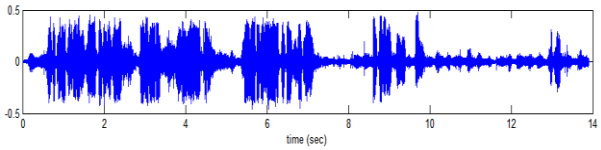

그림 5. TTS 음성합성 결과

5. 결론

본 논문에서 제안한 TTS 를 이용한 화면해설 방송 제작 방법은 기존의 화면해설 방송의 제한 점인 인적, 시간적 소요를 보완하여 화면해설 방송의 보급화를 목표로 하였다. 앞에서 확인한 바와 같이 묵음구간 추출, TTS 등을 활용하여 화면해설 작가의 화면해설 대본 작성만으로 화면해설 방송이 가능함을 보였다.

Acknowledgement

본 연구는 미래창조과학부가 지원한 2013 년 정보통신·방송(ICT) 연구개발사업의 연구결과로 수행되었음. [감성기반 사용자 맞춤형 UI/UX 방송시스템 기술 개발]

참고문헌

[1] 홍종배, “장애인 방송접근성 표준화 동향”, *TTA journal*, Vol.137, p52-56, 2011.

[2] 이은향, “시청각 장애 보조 방송 서비스 표준화 동향”, TTA journal, Vol.137, p57-61, 2011.

[3] Shao, X., Xu, C., Wang, Y., Kankanhalli, M. S., "Automation Music Summarization in Compressed Domain", *IEEE ICASSP*, Vol.4, 2004.