음성인식기술을 이용한 자막생성 연구

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

Subtitle generation using Speech recognition

Chung Hyun, AHN, In Sun Jang
Electronics and Telecommunications Research Institute

요 약

본 논문에서는 동영상, 팟캐스트 오로부터 자막을 생성하여 청각장애인의 미디어 접근권을 향상시키는 음성인식기술을 적용한 자막생성에 대하여 제안한다. 또한 레퍼런스 음성 DB 와 드라마, 팟캐스트 오디오로부터 생성된 자막의 정확도에 대해 평가하였다. 오디오를 이용하여 생성된 자막은 사극의 경우에는 다소 정확도가 낮게 평가되었으나, 전체적으로는 약 80%이상의 정확도를 갖는 것으로 파악되었다.

1. 서론

자막방송은 장애인 방송서비스 중 가장 기본이 되는 방송으로 청각장애인 뿐 아니라 다문화 가정 또는 한글을 배우는 어린이, 한국어를 공부하거나 한국어에 익숙하지 않은 외국인의 TV 시청에 많은 도움을 주며, 외국어 학습용으로도 유용하게 사용되는 멀티미디어 서비스의 하나이다. 또한 방송사 내부적으로는 자막을 이용한 영상검색 등에서 메타데이터로 활용이 가능하여, 방송서비스를 위한 기반정보로서의 활용도도 매우 높은 것으로 평가받고 있다. 현재 국내에선 주요 지상파 방송사가 방송 중인 모든 프로그램에 대한 100% 자막 서비스를 송출하고 있으며, 정확도 98% 이상, 시간 지연 4 초 이내로(미국의 정확도 96%, 시간 지연 6 초 이내) 외국보다 우수한 평가를 받고 있다[1]. 이와 같이 방송은 물론 IPTV, 인터넷 스트리밍 등 다양화된 플랫폼에서 자막의 활용가치가 날로 높아지고 있는 만큼 자막 제작의 효율성을 높이기 위해 기술 개발이 진행되고 있으며, 최근 이슈화되고 있는 인공지능과 결합된 음성인식 기술을 적용하여 시청자가 원하는 장면을 말로 찾을 수 있는 서비스도 실현할 수 있게 될 것으로 기대된다.

현재 방송에서는 속기사에 의한 속기 방식의 자막 구축이 주를 이루고 있지만 음성인식 및 방송부가정보를 활용한 자막 자동생성 시스템에 대한 연구도 활발히 진행되고 있다. 세계 최고의 동영상 서비스인 유튜브는 2009 년부터 이미 동영상에서 나오는 음성을 인식하여 자막을 자동으로 생성하고 번역해주는 기술을 서비스 중에 있다. 국내에서 제작되는 콘텐츠에 대해서는 연기자들의 대사가 별도로 한글자막의 형태로 제공이 되지 않기 때문에 청각장애인의 입장에서는 콘텐츠 소비에 대한 접근성이 많이 제한된다. 또한 국내에서 제작된 영화나 드라마를 해외에 판매할 경우, 영상을 보면서 속기사가 대사를 받아 적고, 이를 다시 번역하여 자막을 만든 경우가 많았다. 그러나, 음성인식과 번역기술이 현실화 되면, 이런 작업은 이제 자동으로 이뤄지게 된다. 각 방송사에서도 방송시스템에 영상물의 음성인식을 통한 자막생성과 데이터베이스화가 중요한 과제가 되고 있다. 영상물과 음성인식의 결합은 단순히 자막을 생성하는 것뿐 아니라 시간 지연이 발생하지 않도록 방송 음성과 자막이 시간적으로 불일치 되는 것을 일치하도록 하며, 부자연스러운 줄 바꿈과 오탈자 등을 개선해 자막에 대한 가독성을 높여주는 고품질 자막 제작 및 송출을 가능하게 할 것이다.

본 논문에서는 구글 음성인식 API 를 이용하여 오디오파일을 입력으로 자막을 생성하는 소프트웨어의 구현과 음성인식에 의한 자막생성 정확도에 대해 검토한다. 소프트웨어는 c#으로 구현되었으며, 오디오파일의 처리를 위해서는 오픈소스인 NAudo 라이브러리를 활용하였다. 음성인식을 위한 레퍼런스로는 ETRI 에서 보유중인 음성 DB 와 드라마, 팟캐스트에서 제공되는 오디오파일을 사용하였다. 음성인식을 위한 오디오 구간의 설정을 위해서는 기본적인 오디오 특징인 STE(short term energy)에너지와 ZCR(zero-crossing rate)를 사용한 EPD(end point detection)기법을 적용하였다.

2. 음성인식 Open API

구글, MS, IBM 은 개발자에게 음성인식을 이용한 어플리케이션의 개발이 가능하도록 API 를 제공하고 있다. 이들 API 는 입력으로 개발자 고유의 APIkey, 인식하고자 하는 오디오 소스언어, 인식하고자 하는 오디오파일을 기본으로 한다. ,예를 들어 구글 음성인식 API 는 다음과 같이 구성된다.

https://www.google.com/speech-api/v2/recognize?output=json&lang=ko-KR&key=APIKey;

이 구조에 따라 WebRequest command set 을 구성하고 오디오 소스스트림을 서버로 전달하면, 주어진 오디오소스에 대한 인식결과는 JSON 으로 반환된다.

최근 발표된 자료에 의하면 “구글 나우”이나 “IBM 왓슨”의 인식오류는 약 8% 정도로 평가되고 있다[2][3]

3. 음성인식 기반 자막생성

음성인식구간을 정확하게 찾아서 입력하는 것은 정확한 음성인식 결과를 얻기 위해 매우 중요하다. 기존의 특징추출 방법 중 대표적인 에너지와 영교차율을 이용한 방법은 비록 잡음의 개입에 취약하다[4]. 본 연구의 관점은 정확한 음성구간의 추출 보다는 자막생성 자체에 중점을 두고 있어 비록 잡음에 취약하지만 알고리즘이 간단한 STE, ZCR 을 기반으로 음성인식 구간을 추출하였다. 아래 그림은 C#으로 구현된 음성인식기반의 자막생성 소프트웨어의 GUI 를 나타낸다. 소프트웨어는 C#으로 구현되었으며, 오디오 처리를 위해서 NAudio 라이브러리를 사용하였다전처리 없이 먼저 자막생성을 위한 오디오 구간추출을 적용하여 대상 오디오 구간을 설정하면 해당 구간에 대해 음성인식 규격에 맞도록 오디오를 자르고 포맷변환을 하여 서버로의 요청을 하기 위한 코드를 생성한다. 서버로 부터 반환된 인식결과는 시작시간-종료시간, 텍스트의 형태로 리스트뷰를 구성한다.

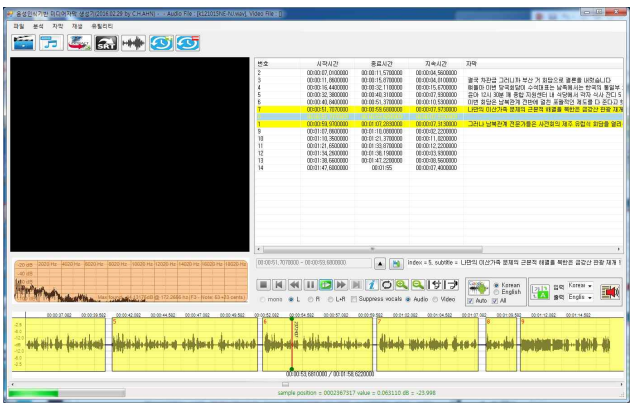

그림 1. 음성인식기반 자막생성 SW

4. 자막의 정확도 평가

음성인식을 통해 생성된 텍스트의 정확도를 평가하기 위해서 일반적으로는 다음과 같은 WER(Word Error Rate), WA(Word Accuracy)가 많이 사용된다[4].

구분	정의	reference	generated
대체(Substitution;S)	음성인식과정에서 다른 단어로 인식되거나, 의미가 다른 단어로 대체된 것	O	O
삽입(Insertion; I)	음성인식과정에서 단어가 추가된 것	X	O
삭제(Deletion; D)	음성인식과정에서 단어가 삭제된 것	O	X

$$\text{Word Accuracy}(WA) = \frac{total\ words - S - D}{total\ words} \quad (1)$$

$$\text{Word Error Rate}(WER) = \frac{S + I + D}{total\ words} \quad (2)$$

식(1)과 (2)를 적용하여 실제 청취를 통해 작성한 텍스트와 음성인식을 통해 생성된 텍스트를 분석한 결과는 표(1)과 같다. 순수하게 음성만 들어 있는 오디오의 경우에는 전체적으로 인식율이 높게 나타난 반면, 배경음이 일부 들어 있는 드라마 대화체의 경우에는 인식률이 다소 낮았으며, 특히 사극의 경우에는 정확도가 더 낮게 나타났다. 이러한 부분은 음성인식학습에 사용된 DB 가 현대어 위주로 되어 있기 때문인 것으로 생각된다.

표 1. 음성인식 정확도 평가

구분	문장수 (단어수)	인식률	
		WA	WER
낭독체(남성/에트리)	26(1363)	0.845	0.168
낭독체(여성/에트리)	19(979)	0.779	0.230
대화체(남성/에트리)	30(201)	0.963	0.040
대화체(여성/에트리)	30(201)	0.898	0.136
대화체(드라마/멜로-상어)	133(2261)	0.778	0.232
대화체(드라마/사극-정이)	61(1290)	0.701	0.319
낭독체+대화체(CBS 김현정)	99(2960)	0.834	0.201
낭독체+대화체(JTBC손석희)	115(3719)	0.812	0.218

5. 결론

자막은 청각장애인들에게 있어서는 TV 나 영상물을 시청함에 있어 반드시 필요하며, 현재 우리나리 지상파 TV 의 경우 거의 100% 자막방송이 이루어지고 있는 반면, 인터넷 동영상에 대해서는 거의 한글자막이 제공되고 않고 있다. 자막 수작업으로 제작되기 때문에 많은 비용과 시간이 소요되어 이에 대한 해결책으로 음성인식을 기반으로 하는 자동자막 생성시스템을 고려하였다. 물론 음성인식 시스템을 자막방송에 바로 활용하기에는 아직 인식률, 수정인력 운용방안 등의 문제가 남아있으나, 음성인식기술의 발전에 발맞추어 음성인식 시스템을 도입한 자막생성 시스템에 대해서도 지속적인 기술개발이 필요하다.

6. 감사의 글

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [B0192-16-1001, 시청각장애인 방송접근권 향상을 위한 디지털자막·음성해설 서비스 기술 개발]

7. 참고문헌

[1]http://www.cas21c.com/caption/caption01.asp

[2]http://googleresearch.blogspot.kr/2015/09/google-voice-search-faster-and-more.html

[3]https://developer.ibm.com/watson/blog/2015/05/26/ibm-watson-announces-breakthrough-in-conversational-speech-transcription/

[4] 김승희, 박준, 김상훈, 2014, 자동통역기술, 서비스 및 기업동향, 전자통신동향분석, 29, 4, p39-48.