청각장애인을 위한 음성-자막 자동 변환 시스템 개발 및 음성 인식률 고도화[1)]

*최미애, **김승현, ***조민애, ****박동영
한국정보통신기술협회
{miae, shk, alps62, dypark}@tta.or.kr
*****김용호, ******윤종후
{solson, jonghoony}@sorizava.co.kr

Development and Enhancement of Automatic Caption Generation System based on Speech-to-Text for the Hearing Impaired

*Choi, Mi-Ae, **Kim, Seung-Hyun, ***Jo, Min-Ae, ****Park, Dong-young
Telecommunications Technology Association
*****Kim, Yong-Ho, ******Yoon, Jong-hoo
Sorizava

요약

인터넷 미디어, OTT, VOD 등 신규미디어가 비장애인의 정보제공 매체로 널리 확대되나, 자막 서비스를 제공하지 않아 청각장애인의 정보 격차가 더욱 심화되고 있다.

청각장애인의 미디어 접근성 제고를 위해 음성인식 서버 및 스마트 폰·태블릿 앱 간 연계를 통해 음성을 인식하여 자동으로 자막을 생성하고 표시하는 음성-자막 자동 변환 시스템을 개발하였고 음성인식률을 높이기 위해 뉴스/시사/다큐 장르 영상 콘텐츠의 음성에 대해 학습용 데이터를 제작하여 음성인식 성능을 고도화 시켰다.

본 논문에서는 청각장애인을 위한 음성-자막 자동 변환시스템 구성과 음성인식률 비교 평가 결과를 보여준다.

1. 서론

음성인식 솔루션은 스마트폰, AI스피커, 네비게이션(앱), 차량 등에서 널리 적용되어 있으며 각기의 목적 및 사용 환경에 맞게 적응되어 활용되고 있다. 단, 사용자의 명령을 마이크를 통해 취득하고 서버로 전송하여, 원격음성인식 서버를 통해 인식하고 명령을 수행하는 방식이 대부분이다.

※ 예시) (사용자)오늘 날씨 어때 -> (음성인식 후 답변) 오늘 서울의 날씨는 …

장애인들의 멀티미디어 접근성 제공을 위한 음성인식 자막 생성 기능은 보다 일반적이고 광범위한 음성인식 성능을 요구하는 어려운 문제이다. 현재 클라우드를 통해 제공하는 대부분의 서비스들은 사용자의 짧은 음성 입력 처리를 위한 용도로서, 음성-자막 간 시간 동기 정보 없이 결과 텍스트만 반환하고 있어, 미디어 자동 자막 생성 용도로 활용이 불가능하다.

국내 인터넷 포털 네이버*와 카카오**에서 서버에 음성을 제공하고 음성인식을 통해 텍스트를 반환하는 클라우드 서비스를 제공하지만 비용을 지불해야 한다.

구글의 클라우드 Speech-To-Text는 스마트폰/PC의 마이크 음성을 실시간으로 서버에 업로드하거나 영상 콘텐츠를 서버에 업로드하고 처리하는 서비스를 실시 중 (9센트/15초 유료 서비스)이다. 구글 실시간 자막 앱은 스마트폰의 마이크 입력 음성을 인식하여 자막을 생성하고 표시하는 앱을 서비스하며, 구글 클라우드 STT 서비스와 연결된다. 유튜브 자동생성 자막은 유튜브 VOD 콘텐츠에 대해 음성을 인식하고 자막을 자동생성하는 서비스를 제공하고 있다.

이와같이, 기존 음성인식 솔루션은 마이크로 입력받는 단문 서비스나 유튜브에 한정되어 자막서비스가 제공되고 있다.

청각장애인용 음성-자막 자동 변환 시스템은 마이크로 취득한 음성이 아닌 재생되는 영상 콘텐츠의 음성을 직접 인식하여, 음성인식 서버 및 스마트 폰·태블릿 앱 간 연계를 통해 미디어 영상 콘텐츠에 대해 실시

1) 본 연구는 방송통신위원회의 "청각장애인용 자막·수어방송 시스템 개발" 과제의 일환으로 수행한 결과임

* 네이버 : 클로바 음성인식 (CSR : Clova Speech Recognition) 서비스 유료 제공 (15초/4원)

** 카카오 : 뉴톤(음성을 문자로 변환)과 뉴톤톡(문자를 음성으로 변환) 서비스 30초/건 2만건/일 무료 제공

간으로 자막 서비스를 제공할 수 있도록 개발하였다.

미디어 영상의 여러 장르 중 1차적으로 뉴스/시사/다큐 영상 콘텐츠의 음성에 대해 학습용 데이터를 제작하여 음성인식률 성능을 고도화시켰다. “음성-자막 자동 변환 시스템의 인식률 성능의 객관적인 평가를 위해 평가용 오디오 DB를 구축하였고 학습 전/후 인식률과 구글, 네이버 음성인식 솔루션과 인식률을 비교하였다.

2. 청각장애인을 위한 음성-자막 자동 변환 시스템 시제품 개발 및 음성 인식률 고도화를 위한 학습

청각장애인을 위한 음성-자막 자동 변환시스템 개발은 음성인식 전문 기술 솔루션 알파케이 음성 인식 엔진[2]을 도입하고 사용자측 단말에서 미디어 재생 앱을 통해 처리하는 방식으로 영상 콘텐츠에서 직접 음성을 취득·인식하고 처리하는 시스템으로 그림 1과 같이 설계하였다. 이 시스템은 인터넷(IP 네트워크)을 통해 수신하는 영상 콘텐츠에 대해 실시간으로 음성을 추출하고 인식하여 자동으로 자막을 생성 표시한다.

그림 1. 청각장애인을 위한 음성-자막 자동 변환 시스템 구성도

음성인식 서버는 양방향 LSTM 딥러닝 알고리즘을 적용한 음성인식엔진을 적용하였으며, 자연어 음성인식 기술이 적용된 높은 정밀도의 음성-문자 변환 인터페이스를 제공한다. 양방향 방식의 은닉층 뉴럴 네트웍은 기존 DNN 방식 및 포워드 방향만을 학습하는 단방향에 비해 입력 시퀀스의 앞뒤 양방향의 가중치를 모두 학습하기 때문에 학습을 통한 인식률의 향상률이 높다.

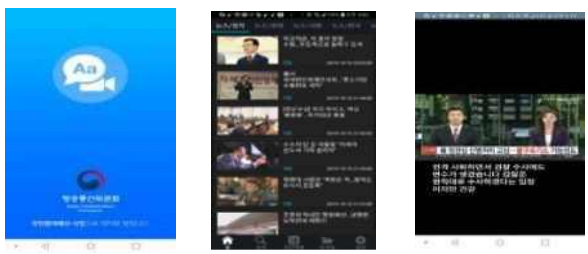

그림 2. 자막 자동 생성 앱 실행, 화면초기화면(왼쪽), 영상리스트화면(가운데), 재생영상에 대한 자막서비스 화면(오른쪽)

자동자막 생성 미디어 재생 앱을 실행 시키면 그림 2의 초기화면과 영상리스트가 나온다. 미디어 재생 앱은 미디어 리스트 정보를 받아서 영상을 재생하고, 추출된 음성에 대한 자막을 받고 동기화하여 실시간으로 영상과 자막을 재생한다. VOD·인터넷 미디어의 다양한 전송포장 방식(MP4, MPEG-2 TS, DASH 등), 음성 부호화 방식(AAC, AC-3 등)에 대응하는 음성 획득이 가능하다. 또한, 자막 On/Off, 화면 표시 방법 등 청각장애인 시청자가 사용하기에 용이한 UI를 적용하였다.

학습서버는 수집된 텍스트 음성정보를 학습 도구를 적용하여 음성을 분석한다. 그림 3은 음성 데이터 분석 및 학습 모델링 구성을 보여준다.

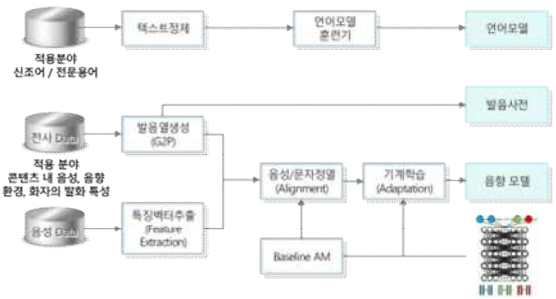

그림 3. 음성 데이터의 분석 및 학습 모델링

음성-자막 자동 변환시스템의 음성인식률 고도화를 위해 자막방송 데이터와 뉴스 기사, 음향 전사 데이터를 활용하여 뉴스, 시사, 다큐, 예능, 드라마의 약 8,000시간 이상의 자막방송 콘텐츠로 440MB의 언어모델 데이터를 생성하였다. 국내 언어모델에 맞게 전문용어, 사투리 등은 텍스트 코퍼스를 추가하였다. 음향 데이터 전사 작업은 뉴스, 시사, 다큐 장르에 대해서 약 1072 시간 진행하였다. 외래어, 전문용어, 신조어 등이 포함된 멀티미디어 콘텐츠의 인식 정확도를 높이기 위한 자연어 음성인식 학습 기술을 적용하여 학습하였다.

영상 콘텐츠의 다양한 장르 중 1차 개발로 뉴스/시사/다큐 장르의 음성에 대해 학습용 데이터를 제작하여 음향 모델과 언어 모델을 학습시켜 음성-자막 자동 변환시스템의 음성 인식률의 성능을 고도화시켰다

3. 시험환경 구축

음성인식 기술을 활용한 자동 자막 생성 기술은 다양하게 활용되어 발전하고 있으나 객관적인 성능 검증용으로 활용할 수 있는 음성 DB가 없어 품질 제고 및 타 시스템과 성능 비교에 어려움이 있다.

음성인식 기반 자동 자막 생성 시스템의 객관적인 성능 평가 및 타 시스템과 성능 비교를 위해 검증용 음성 DB를 제작하였다.

평가용 음성 DB는 뉴스/시사/다큐 장르에서도 더 다양한 어휘와 방송 환경을 반영하기 위하여 시사, 경제, 역사, 자연, 스포츠 등 분야별로 구축하였고 국내 한국어의 특성에 자주 쓰이는 특성어를 포함, 평가용 음성-자막 DB를 구축하였다.

평가용 음성 DB는 표1과 같이 개별 30초, 혹은 2분 단위로 뉴스•시사•다큐•드라마•예능 장르와 줄임말•고유명사•외래어•숫자•신조어•다수화자•축약•감탄사 등 분야/특성어 분류 체계로 총100시간 분량 DB를 제작하였다.

2) 소리자바에서 ETRI 음성인식 엔진을 적용한 AI 음성인식엔진 알파케이

[표 1] 평가용 음성 DB 리스트

구분	분류	총시간	비고
뉴스/시사/다큐	뉴스	20	각 30초단위
	역사	6	
	자연	6	
	스포츠	6	
	경제/사회	10	
	기술	6	
	예술/철학	5	
	GEO	4	
	*특성어	16	
드라마		10	
예능		10	
효과음		1	각 2분단위
계		100	1,132개

*특성어 분류	시간
숫자	2
고유명사	2
줄임말	1
신조어	2
영어	0.5
외국어(영어외 외국어)	0.5
외래어	2
축약된 발성	1
추임새	1
감탄사	1
다수화자(대화, 동시)	3
계	16

평가용 음성과 자막(답안) 등 평가용 DB를 이용하여 음성-자막 변환 시스템과 네이버나 구글 음성-자막 변환 시스템과 비교 평가를 하고 학습 전/후의 음성인식률을 비교할 수 있었다.

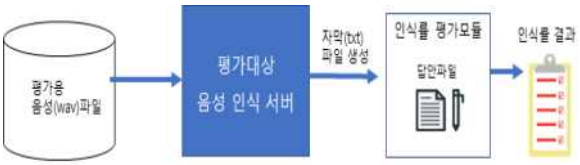

그림 4. 평가용 음성을 이용한 인식률 평가 흐름도

음성-자막 변환 시스템의 음성인식률을 평가하기 위해 음성 평가용 DB 100시간 중 무작위로 6시간을 선택하여 구글, 네이버의 음성인식 엔진과 비교하였고, 30시간을 선택하여 음성-자막 변환 시스템의 성능을 검증하였다. 또한 외래어, 줄임말, 신조어 등 특성어를 제작하여 음성-자막 변환 시스템의 국내 환경에 맞는 특성어에 대한 성능을 검증하였다. 아래의 음절 단위 인식률 식을 적용하여 음성 인식률을 계산하였다.

$$\text{음절 단위 인식률} = \left(1 - \frac{S+D+I}{N}\right) \times 100$$

(N:전체 음절개수, S: 잘못인식된 음절수, D: 빠진 음절수, I: 잘 못 추가된 음절 수)

4. 음성인식률 평가 결과

뉴스/시사/다큐 각 2시간씩 임의로 선택하여 학습 전(기본 모델)과 최종 학습 결과, 그리고 구글, 네이버의 총 4가지 엔진을 비교 평가하였다.

[표 2] 장르별 음성인식률 산출 결과(각 장르별 2시간 분량)

장르	학습전	학습 최종	구글	네이버
뉴스	86.72%	91.78%	64.66%	79.09%
다큐	85.30%	89.21%	64.32%	76.69%
시사	84.35%	89.94%	71.88%	76.76%

국내 동향에 맞게 이미 학습되어진 학습전 기본모델이 구글이나 네이버의 음성인식 엔진보다 음성 인식률이 높았으며, 뉴스/시사/다큐 분야의 학습을 통해 음성인식률 결과가 더 향상되어졌음을 표2와 같이 확인할 수 있었다.

그림 4. 학습전/학습최종/구글/네이버 음성인식률 비교

외래어, 줄임말, 신조어 등 국내 언어 환경에 더 밀접한 특성어별 음성인식률 평가 결과는 표 3과 같다.

[표 3] 특성어별 음성인식률 산출 결과(전체 1.5시간 분량)

특성어	줄임말	고유명사	외래어	숫자	신조어	추임새	다수화자	영어	축약	감탄사	없음
학습전	83.95	87.90	86.86	85.18	85.07	83.67	84.78	89.40	81.59	81.43	92.20
학습최종	93.92	94.51	92.37	93.19	93.70	92.50	91.56	95.36	91.35	91.64	97.35
구글	74.57	68.64	70.41	68.18	73.11	67.82	69.45	88.16	63.18	22.03	66.70
네이버	79.21	81.56	79.08	79.03	76.62	74.76	74.70	91.45	71.93	67.31	82.72
파일개수	10	113	78	90	5	95	33	1	11	3	5

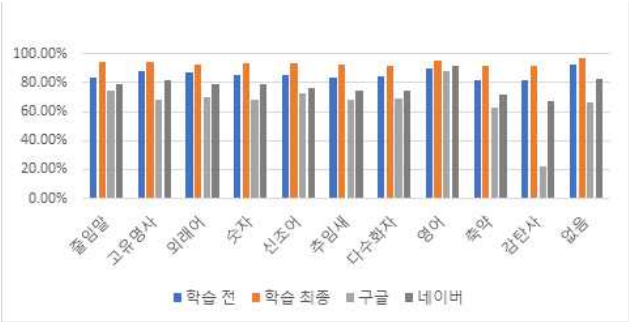

그림 5. 특성어별 음성인식률 비교

특성어별 인식률도 국내환경에 맞게 학습시킨 음성-자막 변환 시스템이 구글, 네이버 음성인식 엔진보다 높았고 학습 결과 학습 전보다 5~10%정도 향상됐음을 알 수 있다.

좀 더 많은 데이터 평가로 객관성을 확보하여 음성-자막 변환 시스템의 음성인식률을 검증하기 위해서, 평가용 오디오 파일을 30시간으로 늘려 뉴스 10시간, 다큐 10시간, 시사 10시간을 평가하였다. 평가 결과는 표 2와 같다.

[표 2] 장르별 음성인식률 산출 결과(각 장르별 10시간 분량)

장르	학습 최종
뉴스	92.44%
다큐	90.46%
시사	90.92%

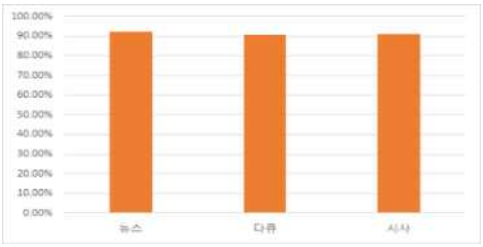

그림 5. 장르별 음성인식률 평가 결과 그래프

뉴스/시사/다큐, 각 10시간 총 30시간의 음성인식률을 평가한 결과 91%의 인식률을 확인할 수 있다. 이는 평가용 오디오 DB에 따라 다소 차이가 날 수 있으나 평균적으로 90%가 넘었음을 확인 할 수 있었다.

5. 결론

급변하는 스마트미디어 환경에서의 청각장애인의 방송접근권 확대 및 정보 격차 해소에 기여하고자 단말(스마트폰 또는 태블릿)의 영상 콘텐츠에서 직접 음성을 인식하여 음성인식 서버 및 스마트폰 앱 간 연계를 통해 영상에 자막서비스를 제공하는 청각장애인용 음성-자막 자동 변환 시스템을 개발하였다. 이 시스템은 기존 음성인식 솔루션을 활용하여 국내 스마트미디어 및 언어 환경에 맞는 학습데이터를 생성하여 학습을 통해 음성 인식률을 향상 시켰다. 본 논문에서는 평가용 오디오 DB에서 랜덤하게 오디오를 추출하여 네이버, 구글의 음성인식엔진과 비교하고 학습전보다 학습후의 음성인식률이 향상되었음을 평가 결과를 통해 보여주었다. 음성-자막 자동 변환 시스템은 청각장애인이 접근이 힘든 인터넷 영상, VOD 등에 대해 자막 서비스를 제공하는데 활용 할 수 있다.

[참조 문헌]

[1] 멀티미디어 콘텐츠에 대한 음성인식 기반 자동 자막 생성 시스템 시제품 제작 사업 최종보고서, TTA, 소리자바