동적계획법을 이용한 장애인방송 폐쇄자막 동기화

오주현
KBS 미디어기술연구소
jhoh@kbs.co.kr

Closed Caption Synchronization Using Dynamic Programming

Juhyun Oh
KBS Media Technology Research Institute

요 약

지상파 방송에서는 청각장애인을 위해 폐쇄자막(closed caption) 서비스가 제공되고 있다. 현재의 폐쇄자막 방송은 속기사가 실시간으로 방송을 보면서 입력하기 때문에 지연이 있다. 또한 이렇게 입력된 폐쇄자막은 TV 프로그램 영상과 별도로 저장되기 때문에 영상과 그 시작점이 맞지 않는 경우가 대부분이다. 폐쇄자막을 온라인 서비스 등에 제공하고자 할 때 이러한 문제로 인해 영상과의 동기가 맞지 않아 사용이 어렵다. 본 논문에서는 TV 프로그램의 음성을 인식하여 동기화된 텍스트를 추출하고, 이를 기 저장된 폐쇄자막과 정렬하여 동기화하는 방법을 제안한다. 실제 TV 프로그램과 자막에 적용하였을 때 대부분의 음절과 라인에서 동기화가 정확히 이루어짐을 확인하였다.

1. 서론

국내 지상파 방송에서는 관련 법규에 따라[1] 전체 방송시간에 걸쳐 청각장애인을 위한 자막방송 서비스를 제공하고 있다. 현재의 자막방송 제작은 지상파 방송을 수신한 외부 서비스 기관에서 실시간으로 자막을 입력하고 이를 방송사로 다시 송신하여 지상파 방송에 삽입하는 구조로 이루어진다. 따라서 송수신 지연과 자막입력 지연이 함께 발생한다. 더 큰 문제는 이와 같이 생성된 폐쇄자막을 온라인서비스 등에서 활용하고자 할 때 발생한다. 방송사에서 폐쇄자막을 실시간 저장하더라도, 실제 방송 프로그램과 시작 시간을 맞추는 것이 어렵다. 폐쇄자막은 입력 스트림을 그대로 저장하지만 아카이브에 저장된 방송 프로그램은 사전제작과 광고 삽입 등으로 인해 시간 오프셋(time offset)이 발생하여 실제 방송과 그 시작점이 달라질 수 있기 때문이다. 이를 해결하기 위해 방송사에서는 수동으로 시작점을 맞추는 작업을 하기도 한다.

만약 순수 음성인식을 통해 폐쇄자막을 새로 생성한다면 동기화 문제는 발생하지 않을 것이나, 자막 서비스를 제공할 정도의 음성인식 정확도를 얻을 수 없다는 것이 문제가 된다. 따라서 본 논문에서는 TV 프로그램에서 음성을 인식하여 음성인식 텍스트를 먼저 생성하고, 폐쇄자막과 정렬(align)함으로써 자막 동기화 문제를 해결하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 음성인식 과정에 대해 살펴본 후, 3 절에서는 동적계획법을 이용한 자막 동기화 방법을 설명한다. 4 절에서는 제안된 방법으로 동기화 실험을 수행한 결과를 살펴보고 5 절에서 결론을 맺는다.

2. 음성인식 자막과 폐쇄자막

TV 프로그램의 음성을 인식하여 텍스트로 변환하는 speech-to-text 를 위해서 많은 방법이 제안되어 있으며, Kaldi[2], CMU Sphinx[3], HTK[4] 등 많은 공개 라이브러리가 사용 가능하다. 그렇지만 본 논문에서는 제안한 방법의 빠른 검증을 위하여 상용 클라우드에서 제공하는 API 서비스(AWS Transcribe)[5]를 사용하기로 한다. 일반적으로 음성인식 라이브러리나 서비스는 그림 1 에서 보듯이 단어 단위로 발화시점과 음성인식 신뢰도 등을 제공한다.

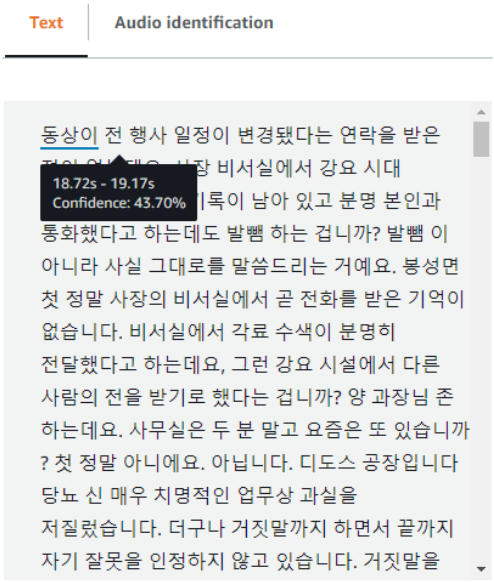

그림 1. 음성인식 결과

그림 2 는 실제로 속기사가 입력한 폐쇄자막을 저장한 파일로, 텍스트 내용을 보면 그림 1 의 음성인식의 정확도를 가늠해 볼 수 있다. 그림 2 의 <SYNC Start···> 정보는 폐쇄자막 라인의 시작점을 밀리초(ms) 단위로 표시한 것인데, 예를 들어 시작 부분의 '본부장님' 이라는 단어는 28.025 초에 시작한다. 그러나 이 단어는 그림 1 의 음성인식 결과('동상이'로 오인식)에서는 18.72 초에 시작하기 때문에 두 텍스트 사이에는 약 10 초의 차이가 있는 것을 알 수 있다. 음성인식 시간 정보를 정확한 것으로 간주한다면, 이 폐쇄자막을 그대로 온라인 서비스 등에 사용할 경우 영상과는 약 10 초의 지연이 발생한다는 의미이다.

```
<SYNC Start=28025><P Class=KRCC>
 -본부장님, 저는 행사 일정이 변경됐다는
<SYNC Start=29088><P Class=KRCC>
 연락을 받은 적이 없는데요.
<SYNC Start=32081><P Class=KRCC>
 -사장 비서실에서 강여원 씨 내선 번호로
<SYNC Start=34074><P Class=KRCC>
 연락한 기록이 남아 있고 분명 본인과
<SYNC Start=37036><P Class=KRCC>
 통화를 했다고 하는데도 발뺌하는 겁니까?
<SYNC Start=37080><P Class=KRCC>
 -네?
<SYNC Start=42092><P Class=KRCC>
 발뺌이 아니라 사실 그대로를 말씀드리는
<SYNC Start=43067><P Class=KRCC>
 거예요, 본부장님.
<SYNC Start=45072><P Class=KRCC>
 저는 정말 사장님 비서실에서 온 전화를
<SYNC Start=46085><P Class=KRCC>
 받은 기억이 없습니다.
<SYNC Start=49022><P Class=KRCC>
 -비서실에서는 강여원 씨에게 분명히
<SYNC Start=52015><P Class=KRCC>
 전달했다고 하는데 그러면 강여원 씨
<SYNC Start=56008><P Class=KRCC>
 자리에서 다른 사람이 전화를 받기라도
<SYNC Start=56083><P Class=KRCC>
 했다는 겁니까?
<SYNC Start=58095><P Class=KRCC>
 양 과장님이세요?
<SYNC Start=59020><P Class=KRCC>
 -네?
<SYNC Start=61020><P Class=KRCC>
 전 아닌데요.
<SYNC Start=65063><P Class=KRCC>
 -그럼 이 사무실에 두 분 말고 여직원이 또
<SYNC Start=66019><P Class=KRCC>
 있습니까?
```

그림 2. 저장된 폐쇄자막

3. 자막 동기화

2 절에서 보듯이 폐쇄자막은 속기사에 의해 오타가 거의 없는 상태로 정확하게 입력되지만, 이후 저장 등의 과정에서 방송 제작 및 아카이빙 프로세스 상 오프셋이 발생하여 시간 동기 정보는 정확하지 않다. 반대로 음성인식을 수행하여 얻은 음성인식 자막은 상대적으로 정확한 시간 동기 정보를 가지고 있지만, 텍스트의 내용은 많은 오류를 내포하고 있다 (표 1 참고).

표 1. 폐쇄자막과 음성인식 자막의 정확성

	텍스트 정보	시간 동기 정보
폐쇄자막	**정확**	부정확
음성인식 자막	부정확	**정확**

따라서 폐쇄자막의 텍스트를 음성인식 자막의 시간 정보로 동기화하는 것이 본 논문에서 제안하는 방법이다. 두 자막의 시간 정보 동기화는 두 개의 문자열을 정렬함으로써 가능하다. 두 개의 긴 문자열을 정렬하는 작업은 흔히 생물정보학(bioinformatics) 분야에서 유전체의 염기서열을 분석하기 위해 사용된다. 이 때 그림 3 과 같은 동적계획법(dynamic programming)의 일종인

Needleman-Wunsch 알고리즘[6] 등을 사용할 수 있다.

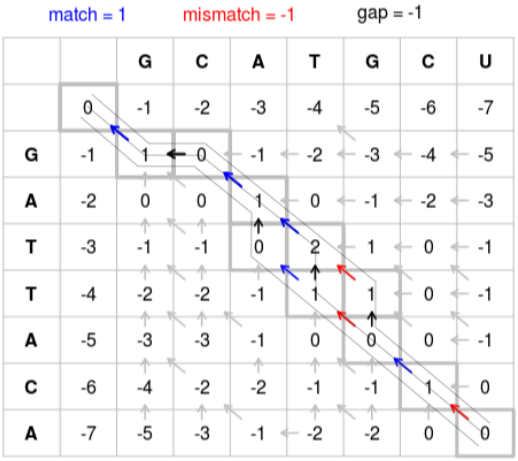

그림 3. Needleman-Wunsch 알고리즘 [7]

제안된 동기화를 위한 대략의 프로세스는 다음과 같다.

1. 영상에서 음성만 추출하여 음성인식 수행 및 결과 저장
2. 폐쇄자막 파일로부터 텍스트의 문장부호 등을 제외한 문자만을 추출
3. 음성인식 자막에서 문자만을 추출
4. 문자열 정렬 기법을 이용하여 2 와 3 의 문자열을 정렬
5. 원 폐쇄자막 파일에서 동기화(sync) 정보의 단위가 되는 라인별로 첫 음절에 대응하는 음성자막의 동기(시간) 정보를 읽음
6. 5 의 동기 정보를 결과물인 보정 폐쇄자막에 기입하여 저장

4. 실험 결과

제안된 방법으로 KBS 드라마 '꽃길만 걸어요' 제 68 회(약 28 분 24 초 분량) 영상의 자막 동기화를 수행하였다. 클라우드 기반 음성인식에는 음성 전송 시간을 제외하고 분석에 약 4 분 7 초의 시간이 소요되었다. 이 음성인식 자막과 폐쇄자막을 동기화한 결과의 일부를 그림 4 에 나타내었다.

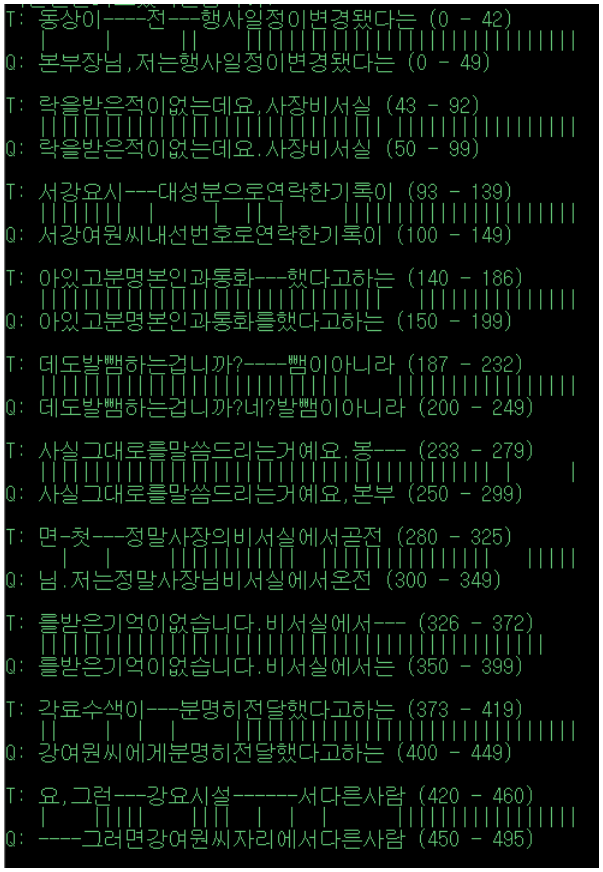

그림 4. 자막 동기화 결과 표시

위 동기화 결과에 대해 음절 단위의 동기화 정확도를 평가하였다. "선물은받았어"와 "선물을받아서"처럼 음성인식 오류로 인해 음절이 다르게 인식되었더라도 "은-을", "았-아"와 같이 정렬이 정확히 되었다면 성공으로 간주하였다. 또한 동기화 결과가 한 음절이라도 어긋나는 경우 실패로 간주하였다. 약 5 천 음절로 이루어진 위 예의 자막에서 이와 같은 기준을 적용하였을 때 제안한 방법은 79.2%의 음절단위 동기화 정확도를 나타내었다. 실험에 사용된 폐쇄자막에는 실제 음성과 관계없는 "(휴대전화 통화연결음)"과 같은 보조정보도 들어있어 향후 전처리를 통해 이를 제외하면 정확도는 더 높아질 것으로 보인다. 실패로 처리된 결과 중에서도 실제로는 2 음절 이내의 오류가 대부분을 차지하기 때문에 실제 동기화 결과에는 거의 영향을 주지 않았다. 자막 음절 당 평균 동기화 오차는 약 0.19 음절로 나타났는데, 이 결과로부터 자막의 대부분이 음절 단위로 정확하게 정렬되었음을 알 수 있다.

위 음성인식 자막과 폐쇄자막을 정렬하여 동기화된 폐쇄자막을 얻는 데는 약 78 초가 걸렸다. 앞으로 실제 현장 적용을 위해서는 음성 데이터 전송과 분석결과의 수신이 필요한 클라우드 음성인식을 내부(on premise) 구현으로 바꿀 필요가 있을 것으로 판단된다.

5. 결론

국내 청각장애인 시청자들을 위하여 지상파 방송에서는 빠짐없는 폐쇄자막 방송 서비스를 제공하고 있지만, 방송 프로그램이 재활용되는 온라인 서비스 등에서는 이와 같은 서비스 접근권이 제공되지 못하고 있는 것이 현실이다. 이러한 문제는 방송 프로그램과 폐쇄자막이 시간적 동기가 어긋나 있기 때문이며, 이를 해결하기 위하여 생물정보학에서 사용되는 동적계획법 기반으로 음성인식 문자열과 폐쇄자막 문자열을 정렬함으로써 동기화 정보를 생성하는 방법을 제안하였다. 제안된 방법은 KBS 에서 기 구현하여 사용하던 시스템과 비교하여 속도와 메모리 사용량을 개선함으로써 효과적으로 온라인 서비스 등을 위한 동기 보정 자막을 제공할 수 있을 것으로 기대된다. 향후 음절비교를 세분화하는 방법을 연구하여 성능을 개선하고, 시간이 오래 걸리는 음성인식 과정을 별도로 수행하지 않고 동기화를 위한 정보만을 추출하여 처리시간을 단축하는 방법을 연구할 예정이다.

Acknowledgement

본 연구 논문은 과학기술정보통신부 및 정보통신기획평가원의 출연금으로 수행하고 있는 한국전자통신연구원 시청각 장애인의 방송시청을 지원하는 감성표현 서비스 개발[2019-0-00447]의 연구결과입니다.

참고문헌

[1] “장애인방송 편성 및 제공 등 장애인방송접근권 보장에 관한 고시,” 방송통신위원회고시, 제 2011-53 호

[2] Kaldi Speech Recognition Toolkit, https://github.com/kaldi-asr/kaldi, 2020.

[3] OPEN SOURCE SPEECH RECOGNITION TOOLKIT, https://cmusphinx.github.io/, 2020.

[4] HTK, http://htk.eng.cam.ac.uk/, 2020.

[5] Amazon Transcribe, https://aws.amazon.com/ko/transcribe/, 2020.

[6] Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 48 (3)

[7] https://en.wikipedia.org/wiki/Needleman-Wunsch_algorithm