원격 개인 농구 기술 피드백 영상 자동 더빙 시스템

임종욱 [1], Ray Kim[2], 윤 영 [3]
[1] 홍익대학교 컴퓨터공학과 학부생
[2]Seons Inc
[3] 홍익대학교 컴퓨터공학과 교수

whddnrkk@g.hongik.ac.kr, jongrinkim89@gmail.com, young.yoon@hongik.ac.kr

Automatic Dubbing System for Remote Personalized Basketball Feedback Video

Jong-Uk Lim[1], Ray Kim[2], Young Yoon [1]
[1]Department of Computer Engineering, Hongik University
[2]Seons Inc.

요 약

본 논문은 전문 스킬 트레이너들의 개인 농구 기술 분석 및 피드백 영상에 더빙을 자동으로 적용하는 시스템을 제안한다. 이 시스템은 농구 용어집 기반 번역, 음성-텍스트 변환 모델 간의 비교 분석, 영상과 더빙 트랙 동기화 알고리즘을 통해 다양한 언어로의 신속한 자동 번역과 더빙을 가능하게 함으로써 선수와 코치 간의 언어 장벽 없는 소통을 지원한다. 본 연구는 자동 더빙 기술에 힘입어 원격 농구 교육 효율성과 질의 재고 및 저변 확산에 기여하고자 한다.

1. 서론

농구는 빠르게 세계화되는 스포츠로, 다양한 언어를 구사하는 선수들에게 전문 스킬 트레이너들이 원격으로 개인 기술에 대한 분석과 조언을 해주는 교육 서비스에 대한 수요가 증가하고 있다. 본 논문은 스킬 트레이너들의 선수 개인의 기술에 대한 피드백 영상을 선수의 언어에 맞게 자동으로 더빙을 적용하는 시스템을 통해 원활한 의사 소통을 도와주는 방법을 선보인다.

자동 더빙 시스템은 원본 음성을 텍스트로 변환 (Speech-To-Text, STT), 해당 텍스트를 다른 언어로 번역, 그리고 번역된 텍스트를 새로운 음성으로 변환 (Text-To-Speech, TTS)하는 과정을 수행한다. 본 연구에서는 다양한 자동 번역 및 TTS 모델의 성능을 비교 분석하여 농구 피드백에 최적화된 모델을 찾아내고자 한다.

더빙 과정에서의 주요 도전 과제 중 하나는 번역된 음성이 원본 비디오와 시간적으로 일치하도록 음성의 속도를 탄력적으로 조절하는 것이다. 이를 위해 본 연구에서는 영상과 더빙 트랙간 동기화 알고리즘을 탐구한다. 또한, 농구 용어집 기반 번역 접근 방식을 통해 농구와 같은 전문 용어가 정확하게 번역되도록 하여 의미 전달의 정확도를 높이는 방법을 제안한다.

본 논문은 농구 피드백을 위한 자동 더빙 시스템의 설계, 구현, 그리고 평가에 대한 포괄적인 논의를 제공함으로써, 언어 장벽을 허물고 원격 스포츠 교육 효율성과 질의 향상 및 저변 확산에 기여하고자 한다.

2. 자동 더빙 프로세스

자동 더빙 프로세스 STT, 자동 번역, voice cloning, 음성가공, 음성 연결 등의 단계로 구성된다.

그림 1 에 도시된 바와 같이 원본 음성으로부터 텍스트 스크립트를 생성하여 번역한 뒤 TTS 로 화자 음성의 특징을 가진 목소리를 생성하고, 음성가공단계

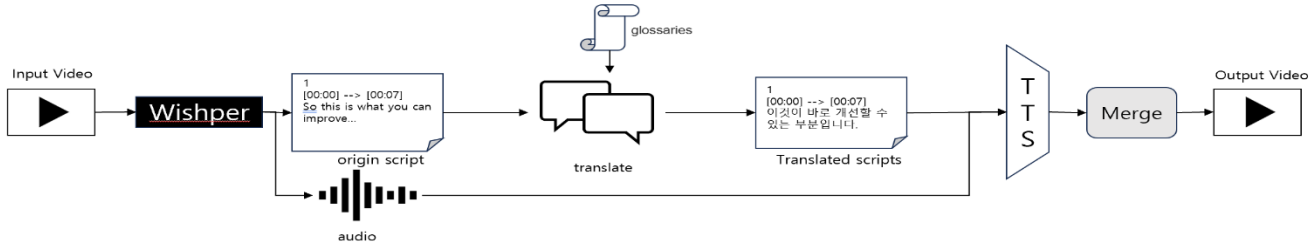

Figure 1 자동 더빙 프로세스

에서 음성의 속도를 조절하여 원본 비디오에 합성한다.

3. 번역

번역 과정은 자동 더빙 시스템의 핵심적인 단계 중 하나로, 원본 음성의 의미를 정확히 다른 언어로 전달하는 데 중점을 둔다. 이 과정은 농구 전문 용어와 은어의 정확한 번역에 있어서 더욱 중요하다. 이를 위해 본 연구에서는 용어집 기반 번역 접근 방식을 평가하였다. 번역결과 평가는 BELU score, 수동 채점 방식을 사용하였으며 [1], Seons [2]에서 제공한 실제 피드백 영상에서 농구 전문 용어가 포함된 28 개의 주요 문장에 대해서 Deepl+용어집 방식이 가장 높은 정확도로 번역한 것을 확인하였다.

번역의 일례를 들면 다음과 같다.

- 원본 텍스트: It's because every time someone **drives**, you're just **hacking** them, right?
- 기본 GPT: 그것은 누군가 운전할 때마다, 넌 그저 그들을 해킹하고 있기 때문이야, 맞지?
- GPT + 용어집: 그것은 누군가 드라이브할 때마다, 넌 그저 그들을 해킹하고 있기 때문이야, 맞지?
- Deepl + 용어집: 상대가 **돌파**할 때마다 **파울**을 범하기 때문이죠?

	GPT	GPT+용어집	Deepl+용어집
수동 채점	7.1	7.4	7.8
BELU score	0.4	0.6	0.7

<표 1> 농구 전문 용어 포함한 번역 정확도

4. TTS 모델 분석 비교

본 연구에서 분석한 TTS API 의 특징을 표 2 에 정리하였다.

	Google TTS	Seamless M4T	Bark voice cloning	Xttx_2
음성 데이터 양	20~30 분	-	1 분 미만	1 분 미만
지원 언어	56 개	35 개	13 개	13 개

<표 2> 라이브러리 특징

제시된 TTS 성능은 널리 사용되는 주관적 음질 평가 지표인 MOS (Mean Opinion Score)로 측정하였다. MOS 방식은 생성된 음성 샘플을 평가자 그룹 별로 각 지표마다 1~5 점을 부여하여 문장마다 개별점수를 부여한 뒤 평균을 구하는 방식이다. 평가에 사용된 음성샘플은 10 개이며 지표는 노이즈, 발음, 대사 정확도이다. 대사의 정확도는 음절 오류률(Character Error Rate, CER)로 평가하였다 (식 1).

$$CER = (S_c + D_c + I_c) / N_c \qquad \text{식 1}$$

S_c 는 변경된 글자 수, D_c 는 삭제된 글자 수, I_c 는 삽입된 글자수, N_c 는 전체 글자 수이다.Google TTS 가 가장 높은 성능을 낸 것으로 확인하였다.

	Google TTS	SeamlessM4T	Bark voice cloning	Xttx_2
MOS 점수	4.45	2.7	3.24	4
CER(%)	0	0.43	0.45	0.27

<표 3> 성능 평가 결과

5. 영상과 더빙 트랙간 동기화 알고리즘

생성된 음성이 원본 멀티미디어 콘텐츠와 재결합될 때, 생성된 음성의 시간적 길이가 원본 시퀀스의 길이와 일치하지 않아서 음성 합성의 결과물이 후속 대사와의 겹침을 초래하거나, 내용 전달의 흐름을 방해하는 문제가 발생하였다. 이 문제를 해결하기 위하여 TTS 로 생성된 음성의 시간적 길이를 조정하는 메커니즘의 개발은 필수적이다. 생성된 오디오의 음성 높낮이를 유지시면서 음성 길이에 비례하여 배속인자를 조절하는 Time-Stretching 기법을 Algorithm 1 과 같이 개발하여 생성된 더빙 트랙과 영상의 동기화 문제를 해결하였다.

Algorithm 1 Calculate Speech Rate for Time-Stretching

Require: $startTime \geq 0$, $originalDuration > 0$, $generatedDuration > 0$, $nextStartTime > startTime$
Ensure: $adjustedDuration$ is the duration of the speech after applying the speech rate and time-stretching
$speechRate \leftarrow 1.0$
$timeUntilNextPhrase \leftarrow nextStartTime - (startTime + originalDuration)$
if $generatedDuration > originalDuration$ **and** $timeUntilNextPhrase < (generatedDuration - originalDuration)$ **then**
  $speechRate \leftarrow originalDuration / generatedDuration$
end if
{Time-Stretching Process}
$timeStretchedDuration \leftarrow adjustedDuration / speechRate$
if $speechRate < 1.0$ **then**
  {The speech will be played faster, reducing the duration.}
end if

6. 결론

본 논문에서 제안된 농구 피드백 영상의 자동 더빙 처리 시스템은 원격 스포츠 교육과 훈련 시 언어 장벽을 허물고 정확하게 의사를 전달하는 새로운 접근법을 제시하였다. 본 연구에서 다양한 전문 용어집 기반 번역과 TTS 모델을 적용하고 성능을 비교 분석하였다. Seons 에서 제공한 피드백 영상의 자동 서비스 결과 Deepl 과 전문 용어집 및 Google TTS 활용했을 때 가장 정확한 피드백 더빙 영상이 생성됨을 실험적으로 확인하였다. 또한 영상-더빙트랙간 동기화 알고리즘의 적용을 통해 자동 더빙 재생의 자연스러움을 향상시켰다.

참고문헌

[1] 최지수 “한국어 번역평가 참조 방향에 관한 고찰: 기계학습 분류 성능 평가지표를 활용하여” 언어와 정보사회 49 pp. 231-250 (2023)
[2] Seons, https://seons.co/