|  |
| --- |
| **1. 주제**  **음성 및 입 모양 분석을 활용한 한글 자막 생성**  **(나)반, 5팀, 20180367, 박상혁** |

|  |  |
| --- | --- |
| **2. 요약**  SNS 및 동영상 플랫폼을 통해 많은 영상물이 업로드 되고있다. 이 프로젝트에서는 자막이 제공 되지 않는 영상물에 대한 음성 및 이미지 분석을 통하여 한글 자막을 생성하는 것을 목표로 한다. 입 모양 인식 알고리즘 Lipreading, 음성 인식을 위한 wav2letter 등의 알고리즘을 사용하여 동영상의 프레임들과 음성을 통해서 자막을 생성한다면, 정보공유 및 지식 전달의 매개체로써 영상 매체들이 보다 정확한 정보 전달이 가능해지고, 청각 장애인에게도 효과적으로 내용을 전달할 수 있게 된다. | **3. 대표 그림**  영상 매체의 정확한 내용 전달을 위한 부정확한 자동 자막에 대한 수정 작업이 필수적이다.  자동 생성 자막의 경우 대다수의 영상물에 대한 정확한 스크립트를 제공되지 않아 불명확한 정보를 전달하기도 한다.    그림 1. 잘못된 자동 자막 수정  (출처: 숭실대 AI 모빌리티 사업단 재직자 대상 특강 유튜브 채널)  음성데이터 학습으로 다음과 같이 부정확한 자막을 수정한다.  “주로 병부 하고” 🡪 “주로 연구하고”  "검출 적으로” 🡪 “검출 쪽으로”  그림 2. 원본 Script 및 한국어 발화  음성이 포함된 동영상  (출처: ITSub잇섭 유튜브 채널)  Script가 존재하는 동영상의 음성, 영상을 각각 분리하여 학습을 진행 시킨다.  음성 데이터와 스크립트의 학습률을 높이는 방향으로 진행한다.  입술 모양을 학습은 자막 수정 작업을 보조한다. |

|  |
| --- |
| **4. 서론**  영상 매체는 다수의 대상에게 생동감 있는 정보를 빠른 속도로 전달이 가능하지만, 정보를 깊이 있게 전달하는 데에는 한계가 존재한다. 이를 보완하기 위한 수단인 자동 생성 자막의 경우 대다수의 영상물에 대한 정확한 스크립트를 제공되지 않아 불명확한 정보를 전달하기도 한다. 이를 보완하기 위해 자동 생성 자막을 영상인식, 음성인식, 기계학습의 방식을 활용하여 개선하고자 한다. 각각의 방식에 대해 영상인식은 LipReading, 음성인식은 wav2letter++, 기계학습은 DNN을 기반으로 학습을 진행한다. |

|  |
| --- |
| **5. 본론**   * 필요한 기술 요소   한국어 발화 음성과 이미지가 포함된 동영상  동영상 소스에서 추출한 영상 데이터 및 음성 데이터  Wav2Letter - 음성 Feature 들에 대한 학습을 진행  LipReader – 영상 프레임을 학습하여 한국어 발음에 대한 입술 모양 라벨링   * 구현 방법 및 개발 방향   <데이터 준비 및 전처리>  한국어 발화 음성과 이미지 및 Script가 포함된 동영상을 기본 데이터셋으로 한다.  동영상 소스는 학습 데이터인 원본 Script가 포함된 영상과,  테스트 데이터인 Script를 가지지만 배제한 영상으로 구분한다.  동영상 소스에서 추출한 영상 데이터 및 음성 데이터를 추출한다.  <학습 진행>  Script가 포함된 영상들에 대한 다음의 학습과정을 진행한다.   1. Wav2Letter - 음성 Feature 들에 대한 학습을 진행한다 2. LipReader – 영상 프레임을 학습하여 한국어 발음에 대한 입술 모양 라벨링한다. 3. 학습율은 기존 Script에 대하여 (1)의 결과의 정확도를 기준으로 하고, (2)의 학습 결과를 parameter로 추가하여 (한글 자음, 모음에 대한 사용 시점의 정확성을 지표로 선정) 역전파 후 기계학습을 진행한다. 4. 학습된 모델을 Script가 존재하지 않는 동영상에 대하여 적용하여 성능을 판단한다.     **[ 시스템 개요 그림 ]** |

|  |
| --- |
| **6. 결론**  원본 Script가 존재하지 않는 한국어 발화 동영상에 대해 정확한 자막 생성을 목표로 한다.  Wav2Letter (음성 데이터 학습) 와 LipReader (영상 데이터 학습) 을 병합하여 기존 자막 생성 방식에서 존재 하지 않았던 입술 모양을 Parameter로 추가하여 자동 자막 생성의 정확도를 향상 시킨다.  한국어 발화 동영상과 원본 Script 여부에 따라 구분한다.  동영상 데이터에서 음성 및 영상 데이터를 분리한다.  LipReader 에 대한 코드 구현 및 학습 진행  Wav2Letter의 Parameter로 LipReader의 학습 결과 데이터를 추가  Wav2Letter 학습 및 검증  모델의 정확도 확인 |

**7. 출처**

[1] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, Ronan Collobert, “Wav2Letter++: The Fastest Open-Source Speech Recognition System”, Facebook AI Research, 2018.

[2] Hank Liao, Erik McDermott, and Andrew Senior, “LARGE SCALE DEEP NEURAL NETWORK ACOUSTIC MODELING WITH SEMI-SUPERVISED TRAINING DATA FOR YOUTUBE VIDEO TRANSCRIPTION”, Google Inc., 2013.

[1] 허균, 임꺽정, “홍길동의 얼굴 분석,” 한국OOO논문지, 제5권, 제6호, pp. 1-10, 2006.