Journal of the KIMST, Vol. 27, No. 2, pp. 167-176, 2024
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2024.27.2.167

Research Paper 센서·신호처리 부문

한국어 립리딩: 데이터 구축 및 문장수준 립리딩

조선영[1] · 윤수성[*,2]

[1] 국방과학연구소 국방첨단과학기술연구원 인공지능자율기술센터
[2] 국방과학연구소 국방첨단과학기술연구원 위성체계단

Korean Lip-Reading: Data Construction and Sentence-Level Lip-Reading

Sunyoung Cho[1] · Soosung Yoon[*,2]

[1] *AI & Autonomy Technology Center, Advanced Defense Science & Technology Research Institute, Agency for Defense Development, Korea*
[2] *Defense Satellite Systems PMO, Advanced Defense Science & Technology Research Institute, Agency for Defense Development, Korea*

(Received 5 July 2023 / Revised 22 December 2023 / Accepted 17 January 2024)

Abstract

Lip-reading is the task of inferring the speaker's utterance from silent video based on learning of lip movements. It is very challenging due to the inherent ambiguities present in the lip movement such as different characters that produce the same lip appearances. Recent advances in deep learning models such as Transformer and Temporal Convolutional Network have led to improve the performance of lip-reading. However, most previous works deal with English lip-reading which has limitations in directly applying to Korean lip-reading, and moreover, there is no a large scale Korean lip-reading dataset. In this paper, we introduce the first large-scale Korean lip-reading dataset with more than 120 k utterances collected from TV broadcasts containing news, documentary and drama. We also present a preprocessing method which uniformly extracts a facial region of interest and propose a transformer-based model based on grapheme unit for sentence-level Korean lip-reading. We demonstrate that our dataset and model are appropriate for Korean lip-reading through statistics of the dataset and experimental results.

Key Words : Lip-reading(립리딩), Korean(한국어), Sentence-level(문장 수준), and Transformer(트랜스포머)

1. 서 론

립리딩은 사람이 발언할 때 음성 없이 영상만을 통해 발언하는 내용을 추론하는 기술로써, 동영상의 연속된 모든 프레임을 통해 문장 또는 단어들을 추론하는 기술이다. 따라서 음성 정보가 없는 비디오, 소음환경 및 다중 화자 환경에서 발언 내용을 추론하거나 음성 인식의 성능 향상 등에 활용될 수 있다. 국방 분야에서 이를 활용하는 방안으로는 대외 주요 관심 인물들

* Corresponding author, E-mail: mercury@add.re.kr

의 비공식 발언 분석, 기도비닉이 요구되는 야간 작전이나 은밀하게 이뤄지는 특수 작전 상황 하에서 발성이 아닌 립리딩을 통한 교신 등에 활용될 수 있다. 이러한 립리딩 기술은 근본적으로 영상 내 입술의 움직임만을 통해 인물의 발언 내용을 추론하기 때문에 서로 다른 글자여도 같은 발음이 나타나는 글자들이나 화자에 따라 같은 글자여도 다르게 발음이 나타나는 글자들이 있는 등 어려운 문제로 알려져 있다.

최근 립리딩 분야에도 딥러닝이 활용되어 많은 발전이 이뤄지고 있다. 문장 단위의 립리딩 연구 초기에는 순환 신경망(Recurrent Neural Network, RNN) 및 장단기 메모리(Long Short-Term Memory, LSTM)를 기반으로 하는 시종간(End-to-End) 립리딩 구조들이 제안되었다[2,3]. 그러나 순환 신경망 계열의 구조들은 고정된 크기의 벡터에 모든 정보를 압축함에 따라 유발되는 정보 손실과 기울기(Gradient) 손실이 존재하게 된다. 이를 해결하기 위해 관심 부분에 더 집중하여 학습하는 방식인 어텐션(Attention) 기법 및 어텐션들로만 구성된 트랜스포머(Transformer) 기반 방법 이 제안되어 현재 문장 단위의 립리딩 분야에서 최고 수준의 성능을 보여주고 있다[1].

하지만 이러한 기술들은 영어 기반의 립리딩 알고리즘으로써 문장 추론 시 출력되는 한 글자가 26개의 알파벳 중 하나로 표현되는 영어에 최적화된 알고리즘이다. 영어와 달리 한국어에서의 한 글자는 초성, 중성 및 종성으로 표현되며, 한 글자가 될 수 있는 가짓수가 11,172개로 언어체계가 매우 다르기 때문에, 기존의 알고리즘들을 적용하는 것에 한계가 존재한다. 한국어 립리딩을 위한 기존 기술들의 경우 실험실 환경에서 취득한 데이터를 기반으로 모음을 추론하거나 제한된 수의 단어를 추론하는 기술이다[4,5]. 따라서 다양한 환경에서 데이터가 수집되지 않았기 때문에 실제 환경에서의 성능은 현저하게 낮아지게 된다. 또한 대부분의 립리딩 응용들은 문장 단위의 추론이 요구되는데 기존 한국어 기반 립리딩 방법들은 모음 또는 단어 추론이 목적이므로 문장 단위의 발언 내용 추론이 어렵다는 한계가 있다.

본 논문에서는 실 환경에서의 한국어 립리딩을 위해 대용량의 한국어 립리딩 데이터셋을 구축하고 문장 단위로 발언 내용을 추론할 수 있는 한국어 립리딩 방법을 제안한다. 이를 위해 대용량의 한국어 방송 동영상을 수집하였고, 화자의 발화 단위로 동영상 분할 및 발언 내용에 대한 레이블링을 통해 데이터셋을 구축하였다. 립리딩 모델의 경우 추론 클래스를 음절이 아닌 자소 단위로 분리하여 학습을 수행하도록 하였는데, 이는 26개의 음절을 갖는 영어와 달리 한국어는 11,172개의 매우 많은 음절을 가지고 있으며 영어와 달리 한국어는 문장 구성 시 조사가 존재하여 이를 효과적으로 구분하기 어렵기 때문이다. 따라서 음절에 비해 개수가 적은 자소 기반으로 학습을 수행하는 시종간 립리딩 방법을 제안하였다. 제안하는 방법은 먼저 얼굴 검출 알고리즘을 통해 추출된 얼굴의 주요 지점(Landmark)을 활용하여 입술 영역을 효과적으로 추출하고, 추출된 입술 영역은 트랜스포머에 입력되어 자소 단위로 추론하게 된다. 구축된 데이터셋을 기반으로 제안하는 모델의 학습과 성능 평가를 통해 제안하는 방법의 적정성을 확인하였다.

2. 관련 연구

2.1 립리딩 데이터셋

문장 수준의 대용량 한국어 립리딩 데이터셋은 알려진바 없으며, 대부분의 립리딩 연구들은 영국 옥스퍼드 대학의 Visual Geometry 그룹에서 구축된 영어 립리딩 데이터셋을 사용한다. 주로 영국의 BBC 방송 프로그램 동영상을 이용하여 데이터가 수집되었으며, 최근에는 YouTube를 이용하여 데이터를 구축하기도 하였다. Lip Reading in the Wild(LRW) 데이터셋[10]는 500개의 단어들에 대해 500,000개의 발화문으로 구성된 데이터셋이다. BBC 방송들로부터 추출된 29개의 프레임들(1.16초 분량)의 동영상 클립들로 구성되어 있으며, 동영상의 중앙 부분에서 발화한다. Lip Reading Sentences(LRS) 데이터셋[11]부터 문장 단위의 데이터셋이 구축되기 시작하였다. 총 17,428개의 단어들로 이루어진 118,116개의 문장에 대한 발화문으로 구성된 데이터셋이며, 각 발화문은 10초 분량의 동영상 클립들로 구성되어 있다. Multi-View-LRS(MV-LRS) 데이터셋[12]는 기존의 LRW 데이터셋[10] 및 LRS 데이터셋[11]에 비해 옆모습의 얼굴 데이터를 포함하고 있다. 총 14,960개의 단어들로 이루어진 74,564개의 문장에 대한 발화문으로 구성되었다. LRS2-BBC 데이터셋[1]는 다양한 BBC 프로그램 방송들로부터 동영상이 수집되었고, 144,482개의 문장에 대한 발화문에 대해 총 224시간 분량의 데이터를 포함하고 있다. LRS3-TED 데이터셋[13]는 YouTube에서 다운로드 된 5,594개의 TED

와 TEDx 영어 강연 동영상으로부터 수집되었고, 총 438시간 분량의 데이터를 포함하고 있다.

앞서 소개한 데이터셋들은 대용량의 데이터셋을 자동으로 구축하기 위한 다단계 파이프라인을 사용한다. 먼저, 각 개별 프레임들에 대해 얼굴 영역을 검출하기 위해 Single Shot MultiBox Detector(SSD)[14]에 기반한 얼굴 검출기를 사용한다. 샷의 경계는 연속적인 프레임들 간 색상 히스토그램들을 비교함으로써 결정하며, 각 샷 내에 얼굴 트랙들이 얼굴 검출 결과로부터 생성된다. 다음으로, 영어 자막 정보가 있는 동영상들은 자막과 음성 신호 간 단어 수준으로 정렬하기 위해 Penn Phonetics Lab Forced Aligner(P2FA) 방법[15] 및 최신 Kaldi 기반의 음성인식 모델을 사용한다. 오디오와 비디오 스트림 간 동기화 및 얼굴의 입술 움직임과 오디오를 매칭하기 위해서는 SyncNet[16]을 사용하는데, 매칭이 안 될 경우의 동영상 클립들은 사용하지 않았다. 마지막으로 자막에서의 구두점들을 이용하여 문장들을 추출하였다. LRS 데이터셋[11]은 Train-val과 Test 데이터셋으로만 구성되어 있으나, MV-LRS[12], LRS2-BBC[1], LRS3-TED[13] 데이터셋들은 커리큘럼 학습을 위한 Pre-train, Train-val, Test 데이터셋으로 구성되어 있다.

2.2 립리딩 방법

초기에는 단어 단위의 인식 연구[10,17]가 수행되었다면, 대용량 데이터셋[1,13]이 구축되고 딥러닝이 발전하면서 문장 단위의 인식을 수행하는 연구[11,18,19]가 많이 되었다. 문장 단위의 시종간 립리딩을 위한 연구 방향은 크게 두 가지로 나뉜다. 먼저, 주어진 시퀀스에 대해 각 출력 심볼마다 출력을 하는 모델인데, 일반적으로 은닉 마코프 모델(Hidden Markov Model, HMM)을 이용하여 디코딩 단계를 적용하거나 프레임 별 레이블을 추론하고 프레임 별 추론값과 출력 시퀀스 간 정렬(Alignment)이 최적으로 되어있는지 찾는 CTC (Connectionist Temporal Classification)[7] 기반 모델이 있다. 두 번째 방향은 시퀀스-시퀀스(Sequence-to-sequence) 모델로써, 모든 입력 시퀀스를 입력받은 후에 출력 문장을 추론하는 방법[11]이다.

초기의 문장 단위의 인식 연구는 LSTM 기반 시퀀스-시퀀스(Sequence-to-sequence)[11] 또는 CTC[18,19]를 이용하여 음성인식용으로 개발된 모델들을 적용하는 경우가 많았다. 최근에는 트랜스포머 기반 방법[1]이 좋은 성능을 보이면서, 트랜스포머 기반 변형된 모델들을 활용하는 경우가 많다. 트랜스포머 기반 변형 모델로써 [20]에서는 컨볼루션(Convolution) 블록들을 이용하였는데, 이웃 프레임들 내 단기간 시간적 의존성 특징들을 추출함으로써 성능을 향상시켰다. [22]에서는 컨볼루션 신경망(Convolution Neural Network, CNN)과 트랜스포머를 결합하여 오디오 시퀀스의 지역적 및 전역적 의존성들을 함께 모델링함으로써 성능을 향상시키는 음성인식 모델인 Conformer 모델[21]을 적용한 립리딩 모델을 제안하여 성능을 향상시켰다.

3. 데이터셋 구축

3.1 데이터 수집 및 레이블링

데이터셋 구축을 위해 2014년에서 2020년까지 약 7년간 방송된 동영상들을 수집하였다. 총 300 TB 용량의 동영상 데이터들을 수집하였고, 동영상들은 주로 뉴스 보도가 약 50 %, 다큐멘터리나 특집 방송들이 약 41 %, 연속극이나 영화가 약 4 %, 나머지 기타 방송들을 포함하고 있다. 따라서 수집된 동영상들은 실내외 다양한 환경 및 조명 조건에서 촬영된 다양한 얼굴 포즈의 영상들을 포함하고 있다. 수집한 동영상으로부터 Fig. 1과 같은 절차를 통해 립리딩 데이터를

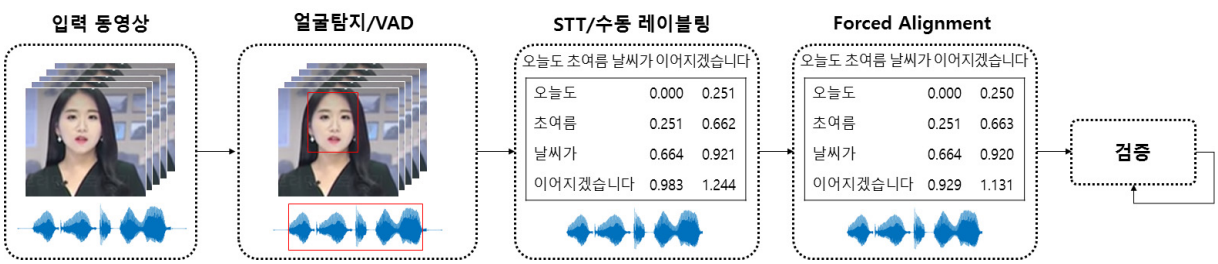

Fig. 1. Overview to construct the korean lip-reading dataset

구축하였다. 먼저, 입력 동영상으로부터 얼굴 및 발화가 동시에 존재하는 구간을 탐지함으로써 립리딩에 유효한 데이터를 추출한다. RetinaFace[6]를 이용하여 얼굴 영역을 검출하고, VAD(Voice Activity Detection)[23]를 이용하여 음성 구간 영역을 검출하였다. 이때 발언 문장 및 동영상 내 발화 간 동기화를 위한 정렬 알고리즘의 입력이 30초로 제한되어 30초 이하로 음성 구간 영역을 분할하였다. 다음으로 추출된 유효 구간에 대해 음성 인식 기술(Speech-To-Text, STT)을 통해 발언 내용 추론 및 문장 내 어절 별 시간 정보를 생성하고, 수동 레이블링을 통해 발언 내용 추론 또는 STT를 통해 추론된 내용 중 오류를 보정하였다. 다음 단계는 추론된 발언 문장과 동영상 내 발화 사이의 동기화를 수행하는 것이다. 앞서 추출된 어절 별 시간 정보는 동영상 내에서 어절을 발화하는 입모양이 나타나는 시간과 차이가 존재할 수 있으므로 서로 매칭이 되도록 정렬을 MFA(Montreal Forced Aligner)[24]를 통해 수행하였다. 마지막으로 수동으로 검증 단계를 거쳤는데, 이는 추론한 문장의 오류나 얼굴이 존재하지 않는 등 유효하지 않은 데이터들을 검출하여 수정/삭제 등의 데이터 클리닝 작업을 수행하였다. 립리딩 데이터에 대한 레이블링 파일에는 다음과 같은 정보들이 포함되어 있다: (1) 원본 동영상의 파일명, (2) 원본 동영상 내 립리딩 데이터의 구간 정보(시작 및 끝 시간), (3) 발언 문장, (4) 립리딩 데이터에 해당하는 동영상의 길이(시간), (5) 발언 문장 내 각 어절 및 어절 별 구간 정보(시작 및 끝 시간).

3.2 데이터 분석

Table 1은 본 논문에서 구축한 한국어 립리딩 데이터셋 및 립리딩 연구에 많이 활용되는 LRS2-BBC[1] 데이터셋에 대한 통계를 보여준다. 한국어 립리딩 데

Table 1. Statistics on the korean lip-reading dataset and LRS2-BBC dataset

데이터셋 명	구분	수량	분량(시간)	명사(단어) 종류 수	어절 수
LRS2-BBC[1]	Pretrain	96 k	195	41 k	2 M
	Train-Val	47 k	29	18 k	337 k
	Test	1,243	0.5	1,693	6,663
한국어 립리딩 데이터셋	Pretrain/Preval	82 k	252.8	20 k	1.45 M
	Train-Val	35 k	35.2	10 k	194 k
	Test	2,343	2.3	2,329	13 k

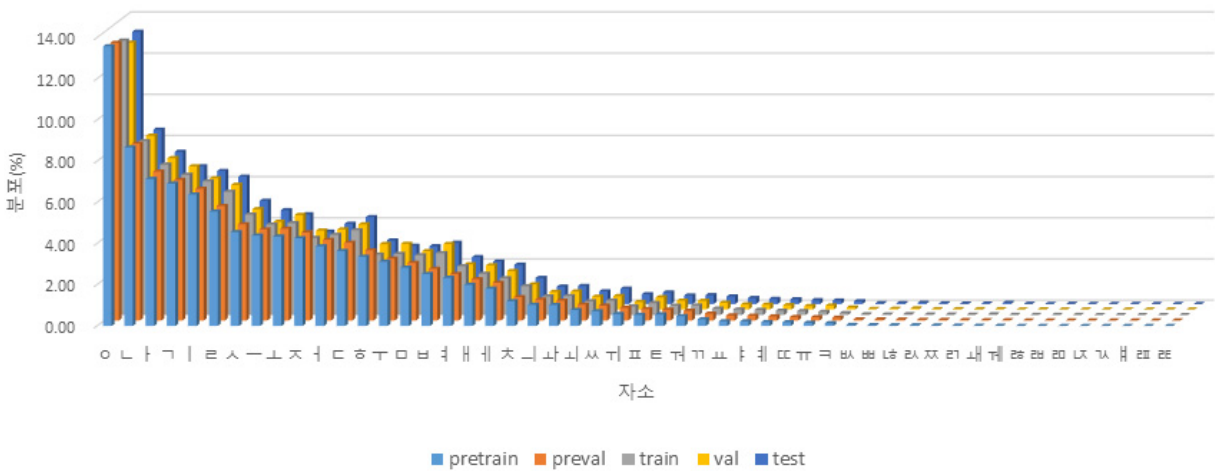

Fig. 2. Statics on graphemes for each data split

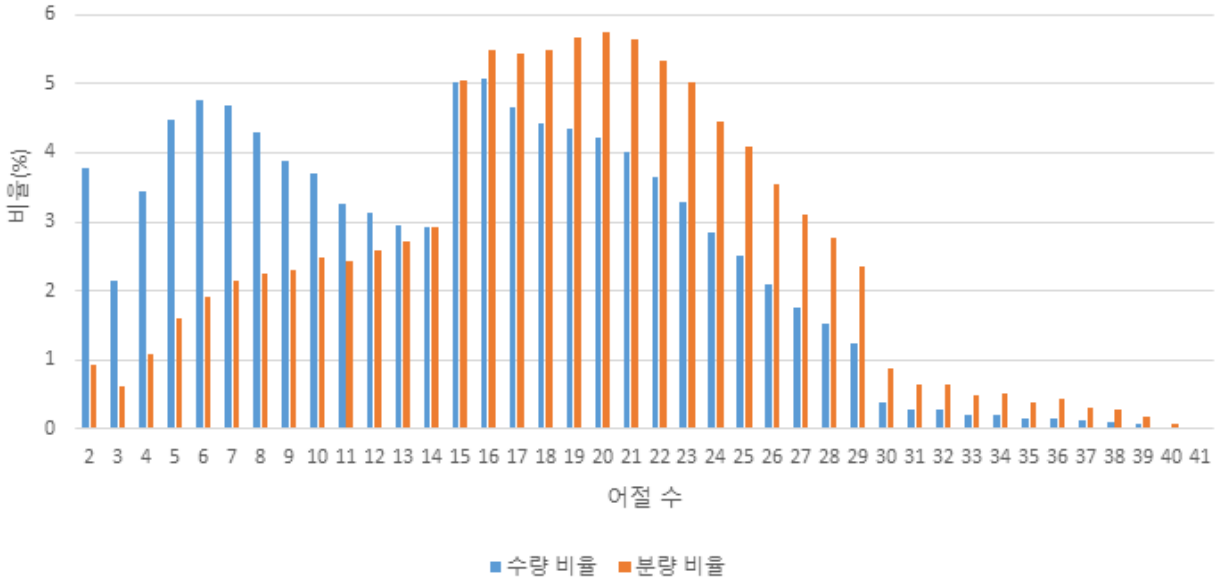

Fig. 3. Statics on the number of word segment

이터셋을 위해 총 346.9시간의 데이터를 구축하였으나, 전처리 알고리즘에 적용한 분량을 제외하고 290.3시간의 데이터를 립리딩 모델의 학습 및 검증에 사용하였다. 립리딩 연구용으로 많이 활용되는 대용량 영어 립리딩 데이터셋인 LRS2-BBC 데이터셋[1]과 비교할 때, 언어 차이로 인해 항목 별 차이는 있지만 해외 벤치마크 데이터셋과 비교할 때 동등한 수준의 데이터셋을 구축한 것을 확인할 수 있다.

Fig. 2는 각 부분 데이터셋 별 자소의 분포를 나타낸다. 모든 부분 데이터셋에서 자소 'ㅇ'의 비율이 가장 높았으며, 다음으로 'ㄴ', 'ㅏ', 'ㄱ', 'ㅣ', 'ㄹ', 'ㅅ' 등의 순으로 자소 비율이 높은 것으로 나타났다. Fig. 3은 어절 수에 따른 수량 및 분량의 분포를 나타낸다. 수량과 분량을 모두 고려했을 때, 어절 수가 15~20개인 데이터가 가장 많았으며 30개 이상의 긴 어절을 포함하는 데이터는 많지 않았다.

4. 자소 단위 기반 한국어 립리딩

본 장에서는 먼저 얼굴 주요 지점 추출 기반의 효율적인 입술 영역 추출 전처리 방법 및 추출된 입술 영역에 대해 사용한 특징을 소개한다. 다음으로 립리딩 모델에 한국어를 적용하기 위한 한국어 자소 분리 및 인덱싱 방법을 설명한다. 마지막으로 Transformer-CTC 기반의 한국어 립리딩 모델에 대해 설명한다.

4.1 전처리

립리딩은 입술 영역의 움직임을 통해 발언 내용을 추론하는 문제이기 때문에 모델의 직접 입력이 되는 입술 영역을 효과적으로 추출하는 방법이 필요하다. 기존 연구들에서 주로 사용되는 데이터셋은 대체적으로 얼굴 영역들이 중앙에 위치하여 영상의 중심을 기준으로 추출하거나 얼굴 주요 지점 추출 알고리즘을 이용하여 입술 중앙 지점을 통해 고정된 크기로 추출한다. 그러나 실 환경의 데이터는 일반적으로 얼굴 영역이 중앙에 위치하지 않을 수 있으며 각 데이터마다 얼굴의 크기가 동일하지 않기 때문에 모든 데이터에 대해 동일한 크기의 입술 영역을 추출할 경우 성능 저하가 발생할 수 있다.

본 논문에서는 이러한 기존의 전처리 방법의 문제를 완화하기 위한 방법을 제안하였다. 이때 본 제안 방법은 [25]에서 제안한 방법을 보완하였으며, Fig. 4는 제안하는 전처리 방법을 보여준다. 먼저 입력 동영상으로부터 추출된 각 프레임에 대해 얼굴 검출 알고리즘[6]을 적용하여 얼굴의 주요 지점을 추출한다. 추출된 얼굴 주요 지점 중 입술의 가장 자리에 해당되는 지점들을 활용하여 입술 중앙 지점을 추정하고, 입

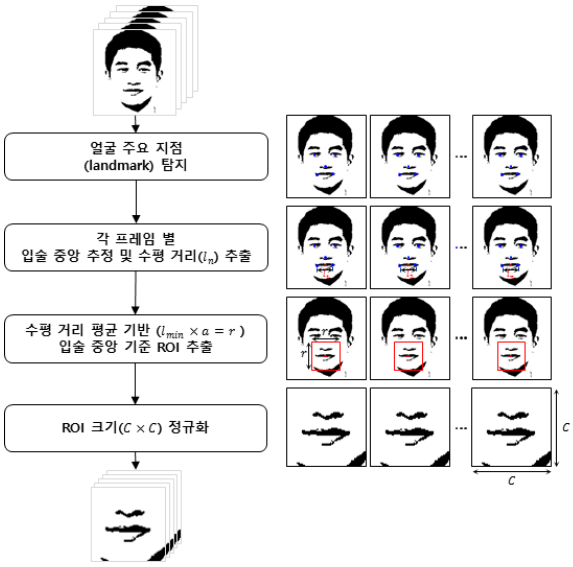

Fig. 4. The proposed method for pre-processing

술 중앙 지점을 기준으로 상하좌우 특정 크기만큼의 정사각형 영역을 입술 영역으로 추출하였다. 정사각형 영역에 대한 크기 선정 시, 모든 프레임의 얼굴 주요 지점 중 입술 영역의 가장 자리 지점 사이의 거리들 $(l_1, l_2, \cdots, l_n)$의 평균값($l_{\min}$)을 활용하였다. 입술 가장 자리 간 평균값($l_{\min}$)을 기반으로 더 넓은 입술 영역을 추출하기 위해 일정 상수(a)를 곱하여 ROI 영역의 크기(r)을 계산하였다. 추출된 ROI 영역은 각 데이터 마다 크기가 달라지므로 학습 시에는 동일한 크기 ($C \times C$)가 되도록 정규화 하였다. 여기서, a는 실험적으로 2.5로 정하였고, C는 112이다.

4.2 입술 영역 특징 추출

4.1절에서 추출된 입술 영역 이미지들로부터 입술의 움직임을 표현하는 특징을 추출하기 위해 시공간 시각적 시종간 방법[8]을 이용하였다. 이 모델은 입력 이미지들에 대해 5프레임의 필터 너비를 갖는 3D 컨볼루션을 적용하고, ResNet-18[9]을 이용하여 시간 스텝별 단일 차원의 텐서가 될 때까지 점진적으로 공간적인 차원을 줄인다. 결과적으로 각 입력 이미지 별 512 차원의 특징 벡터를 추출하였다.

4.3 한국어 자소 분리 및 인덱싱

기존의 립리딩 방법들은 주로 영어를 기반으로 하는 방법들이 대부분으로 알파벳 26개를 클래스로 고려하여 모델을 설계하기 때문에 영어에 최적화된 방법들이다. 그러나 한국어의 경우 11,172개의 글자 종류를 가지고 있어서, 영어 기반의 립리딩 알고리즘을 그대로 적용할 경우 모델 학습 시 추정해야할 파라미터의 수가 많기 때문에 학습에 어려움이 발생한다. 따라서 한국어 데이터를 전처리하는 과정이 필요하고, 한국어의 특성을 고려하여 영어와는 다른 단위로 출력 클래스를 설계해야 한다. 한국어 기반 립리딩 방법들이 일부 연구된 사례가 있으나 극히 적고[4,5], 이는 음절 이외에 단어 또는 형태소 등의 단위로 학습하는 방법 또는 19개의 자음과 6개의 모음으로 구성된 기본 단음절 114개를 활용한 한국어 립리딩 방법이다[5]. 따라서 학습 시에 사용되지 않은 단어, 형태소, 단음절 등이 나타날 경우 인식 가능 범위가 제한되는 문제(Out Of Vocabulary, OOV)가 발생할 수 있다. 또한 알고리즘 자체도 제약된 환경에서 촬영된 데이터에 대해 딥러닝 방식이 아닌 기존의 특징 기반 유사도를 측정하는 방식이어서 실 환경 데이터에 대해 적용할 만한 수준이 아닌 것으로 판단된다.

본 논문에서는 영어 기반의 립리딩 방법에서 사용되는 알파벳 단위 모델 설계의 직접적인 한국어 립리딩 적용 한계를 해결하기 위해, 자소 단위 기반의 한국어 립리딩 방법을 제안하였다. 이를 위해 53개의 자소 맵핑 표를 기반으로 기존의 립리딩 방법에서 성능이 좋다고 알려진 Transformer-CTC 모델[1]을 활용하여 한국어 립리딩 모델을 설계하였다. Transformer-CTC 모델의 경우 매 프레임 단위로 글자 단위를 추론하기 때문에 단어 또는 형태소 단위 보다는 자소 단위가 더 적합하다고 할 수 있다.

Table 2는 본 논문에서 사용한 자소 및 자소에 맵핑되는 인덱스를 보여준다. 한국어에서 한 글자는 초성, 중성 및 종성으로 구성될 수 있는데, 초성은 19개의 자소, 중성은 21개의 자소, 종성은 27개의 자소가 사용된다. 이때 초성과 종성에 모두 사용될 수 있는 자소 16개를 제외하면 총 51개의 자소가 글자를 구성하는 데 필요하다. 51개의 자소, 1개의 공백 및 1개의 EOS(End Of Sentence)를 이용하여 자소 맵핑 테이블을 구성하였고, 학습 시에는 입력된 텍스트로부터 자소 분리 및 인덱싱을 통해 학습되도록 하였다. 추론 시에는 출력된 인덱스로부터 자소로 변환하고 문장으로 다시 변환하는 과정을 통해 결과를 획득하였다.

Table 2. Grapheme to index mapping table

공백	1	ㄱ	2	ㄴ	3	ㄷ	4
ㄹ	5	ㅁ	6	ㅂ	7	ㅅ	8
ㅇ	9	ㅈ	10	ㅊ	11	ㅋ	12
ㅌ	13	ㅍ	14	ㅎ	15	ㄲ	16
ㄸ	17	ㅃ	18	ㅆ	19	ㅉ	20
ㅏ	21	ㅑ	22	ㅓ	23	ㅕ	24
ㅖ	25	ㅗ	26	ㅘ	27	ㅙ	28
ㅚ	29	ㅛ	30	ㅜ	31	ㅝ	32
ㅞ	33	ㅟ	34	ㅠ	35	ㅡ	36
ㅢ	37	ㅣ	38	ㄳ	39	ㄵ	40
ㄶ	41	ㄺ	42	ㄻ	43	ㄼ	44
ㄽ	45	ㄾ	46	ㄿ	47	ㅀ	48
ㅄ	49	ㄽ	50	ㄽㅌ	51	ㄽㅍ	52
EOS	53	-	-	-	-	-	-

4.4 한국어 립리딩 방법

Fig. 5는 제안하는 한국어 립리딩 방법에 대한 전체적인 흐름도를 나타낸다. 4.1절의 전처리 단계에서 추출된 입술 영역들의 시퀀스에 대해 4.2절의 방법을 통해 추출한 특징들은 4.3절의 방법을 통해 입력 텍스트로부터 자소 단위로 분리하고 각 자소를 인덱싱하여 추출된 인덱스 배열과 함께 Transformer-CTC 모델에 입력된다. 본 논문에서는 문장의 문맥을 고려하기 위해 트랜스포머를 사용하였으며, 모델 구성 및 학습 시 수렴 측면에서 효율적이고 잡음에 강인하다고 알려진 CTC[7]를 활용하였다. 본 논문에서 사용된 Transformer-CTC 모델은 여러 개의 멀티헤드-어텐션(Multi-head Att-ention)과 전방향(Feed-forward) 네트워크들로 구성된 인코더, 같은 구성으로 이루어진 디코더 및 1D 컨볼루션과 Softmax 층으로 이루어져 있다. 인코더와 디코더의 레이어 개수는 각각 6개를 사용하였다. 모델에 입력된 특징들은 인코더를 통해 인코딩된 컨텍스트(Encoded Context)를 추출하게 되고, 추출된 컨텍스트 정보는 디코더로 입력된다. 디코더의 출력값은 1D 컨볼루션과 Softmax 레이어를 거쳐 모든 입력 이미지들

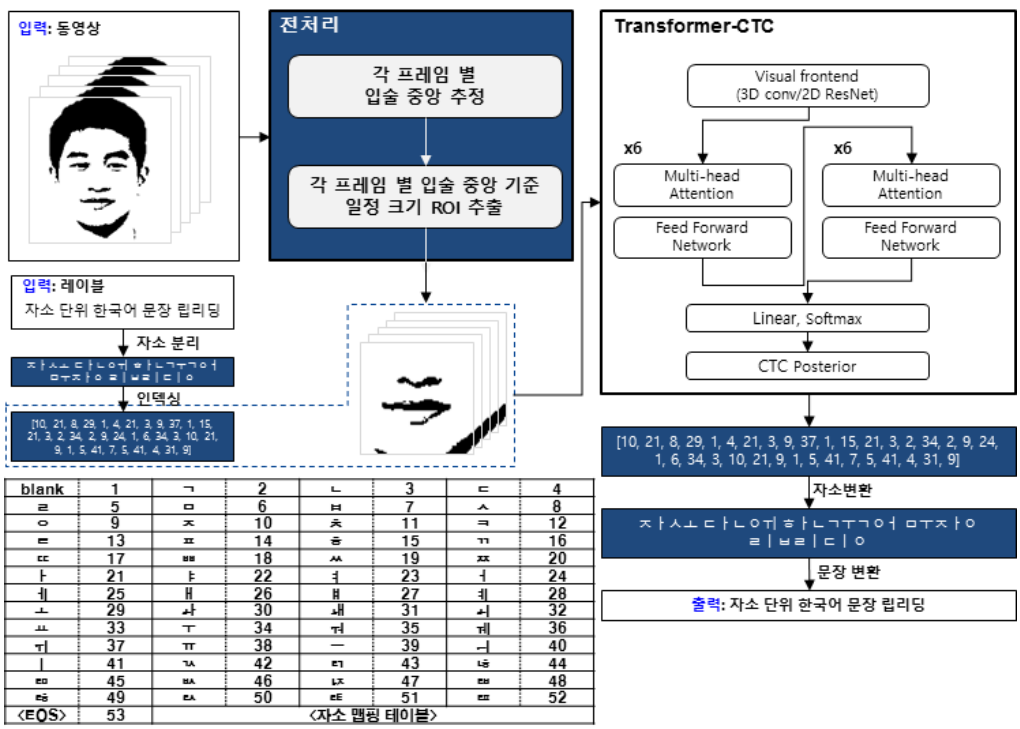

Fig. 5. Overview of the proposed method for korean lip-reading

에 대한 CTC 사후확률(Posterior Probability)을 출력하게 된다. 이 CTC 사후확률을 이용하여 인덱스 배열로 바꾸고 이는 다시 자소, 그리고 문장 변환을 통해 최종적인 결과를 획득하게 된다. 전체적인 모델 학습은 CTC 손실값(Loss)을 이용하여 학습되며, 이때 자소 단위의 GT(Ground Truth)를 기반으로 손실 값을 계산하고 모델을 최적화하게 된다.

5. 실험 환경 및 결과

5.1 실험 환경

데이터 증강(Data Augmentation). 4.1절의 전처리 과정을 통해 추출된 입술 영역은 회색조(Grayscale) 영상으로 변환되었고, Pretrain 및 Train 데이터에 대하여영상의 좌우를 반전시키는 데이터 증강이 수행되었다.

커리큘럼 학습(Curriculum Learning). 일반적으로 시퀀스-시퀀스 학습은 타임 스텝의 개수가 클 때 매우 느리게 수렴하는 것으로 알려져 있다. 커리큘럼 학습은 수렴 속도가 빠르고 과적합(Overfitting)을 줄이는 효과가 있기 때문에, [1]과 같이 전체 문장을 바로 학습하지 않고 점진적으로 시퀀스의 길이를 늘려가는 커리큘럼 학습을 수행하였다. 먼저 본 학습 이전에 사전 학습을 수행하는데, 어절 수를 11단계로 구분하여 학습한다. 11단계는 1, 2, 3, 5, 7, 9, 13, 17, 21, 29, 37개의 어절을 의미한다. 즉, 1개의 어절을 이용하여 모델을 학습한 후에 학습된 모델 기반으로 2개의 어절을 이용하여 학습하는 등 어절의 개수를 늘려가며 반복적으로 학습하게 된다. 최종적으로 37개의 어절로 구성된 시퀀스를 이용하여 학습된 모델은 본 학습을 통해 최종적으로 파라미터 미세 조정(Fine-tuning)을 수행하게 된다. 사전 학습은 3장에서 구축된 데이터셋의 Pretrain / Preval 집합을 이용하였으며, 본 학습은 Train / Val 집합을 이용하였다.

구현 파라미터. 트랜스포머는 멀티헤드-어텐션 레이어에서 8개의 어텐션 헤드를 갖도록 하였고, 인코더 및 디코더 블록은 6개를 사용하였다. 또한 모든 레이어에서 0.1의 확률값을 갖는 드롭아웃(Dropout)을 사용하였다. 립리딩 모델은 β_1 및 β_2에 대한 기본 파라미터 값 및 10^{-4}의 초기 학습률을 가지고 ADAM 최적화 방법을 이용하여 학습되었다.

성능 평가 방법. 일반적으로 영어 립리딩의 경우 성능 평가를 위해 단어 단위 오류율(Word Error Rate, WER)을 사용하며, 아래의 식 (1)과 같이 정의된다:

$$WER = (S + D + I)/N \tag{1}$$

위 식에서, S, D 및 I는 각각 대체(Substitution), 삭제(Deletion), 삽입(Insertion) 오류의 개수를 나타내며, N은 정답에 대한 단어의 개수를 나타낸다. 한국어의 경우 조사가 존재하여 단어 단위로 오류율을 측정하는 것이 어렵기 때문에, 본 논문에서는 한국어 립리딩을 위한 성능 지표로써 자소 단위 오류율(Grapheme Error Rate, GER)과 음절 단위 오류율(Character Error Rate, CER)을 사용하였다. 각 오류율 계산 시 단위는 자소 또는 음절이 되며, 수식 (1)을 통해 각 단위 별로 오류의 개수를 세어 오류율을 계산하였다.

5.2 실험 결과

Table 3은 한국어 립리딩 데이터셋의 Test 집합에 대한 자소 단위 오류율 및 음절 단위 오류율을 나타낸다. 대표적인 영어 립리딩 데이터셋인 LRS2-BBC 데이터셋[1]에서의 기존 방법들의 단어 단위 오류율이 약 50~70 %인데, 이와 비교해서 본 논문에서 구축한 한국어 립리딩 데이터셋에서의 오류율은 약 30~37 %로 다소 성능이 높게 나오는 경향이 있는 것을 확인하였다. 이는 영어 데이터셋보다 본 논문에서 구축한 한국 데이터셋에서의 명사(단어) 종류 수가 절반 이하이기 때문에 성능이 다소 높게 나오는 것으로 추측된다.

Fig. 6은 문장 내 어절 수량에 따른 자소 단위 오류율(GER) 및 음절 단위 오류율(CER)을 나타낸다. 어절 수가 4개 이상인 경우에는 어절 수가 증가하면서 일반적으로 성능이 높아지는 경향을 보였는데, 이는 어절 수가 적을 경우 문장의 문맥 정보가 적기 때문에 발언 내용을 추론하기 더 어려운 문제가 되기 때문인 것으로 판단된다. 어절 수가 4개 일 때 자소 단위 오류율 및 음절 단위 오류율이 가장 높았으며, 어절 수가 많아질수록 오류율이 줄어드는 것을 확인하였다.

Table 3. Grapheme error rate(GER) and character error rate(CER) on korean lip-reading dataset

성능 지표	GER	CER
오류율	30.71 %	37.27 %

Table 4. Examples of our results. GT: Ground Truth

GT 추론 결과	그는 이렇게 피력했습니다 그**들**은 이렇게 피력했습니다
GT 추론 결과	석방할 것을 요구하는 사설을 실었습니다 **ㄱ**방할 것을 요구하는 사설을 실었습니다
GT 추론 결과	금메달 세 개 은메달 세 개 동메달 두 개를 쟁취했습니다 금메달 세**계** 금메달 세**계** 동메달 **수괴**를 쟁취했습니다
GT 추론 결과	오늘 우리나라는 장마 전선의 영향으로 대부분 지역에서 오늘 우리나라는 장**바전 서**의 영향으로 대부분 지역에서
GT 추론 결과	이에 대해서 당시 청와대는 동의한다는 내용의 회답을 주었다고 합니다 이에 대해서 당시 청**화되**는 동**지않**다는 내용의 회**담**을 주었다고 합니다

Table 4는 다양한 종류의 발화문에 대해 제안하는 한국어 립리딩 방법을 이용하여 추정한 결과의 예를 보여준다. '이렇게', '오늘', '우리나라는', '대부분', '합니다' 등의 자주 사용되는 단어나 표현은 정확하게 잘 추정하는 것을 보였다. 그러나 '세 개' → '세계', '회답을' → '회담을', '장마 전선의' → '장바전 서의', '청와대는' → '청화되는' 등과 같이 서로 다른 단어임에도 불구하고 입모양이 비슷하여 잘못된 단어로 추론되는 것을 보였다. 이밖에도 입모양을 작게 하여 발음하거나 얼굴이 정면이 아닌 경우에는 잘못된 추론 결과를 보였다.

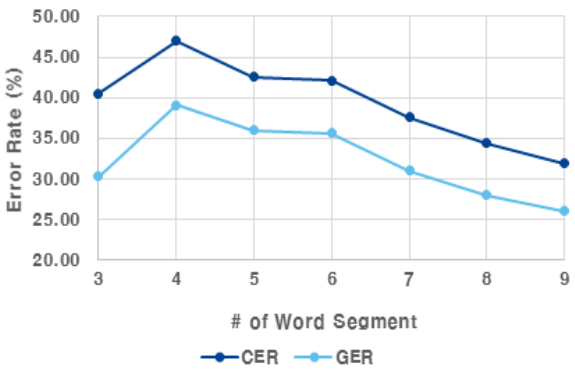

Fig. 6. Grapheme error rate(GER) and character error rate(CER) with respect to the number of word segments in the sentences

6. 결 론

본 논문에서는 문장 수준의 한국어 립리딩을 위한 대용량 데이터셋을 구축하였다. 구축된 데이터셋을 기반으로 립리딩 모델을 학습하기 위해, 먼저 입술 영역의 가장 자리 지점 간 거리들의 평균값을 이용함으로써 발화자의 얼굴 위치나 크기에서의 변형이 있어도 입술 영역의 범위를 일정하게 추출할 수 있는 전처리 방법을 제안하였다. 또한 자소 단위의 트랜스포머-CTC 모델을 통해 학습 시 수렴에 효율적이고 컨텍스트 정보를 활용함으로써 성능을 향상시키는 문장 단위의 한국어 립리딩 방법을 제안하였다. 구축된 데이터를 분석하고 이를 활용한 립리딩 모델의 학습 및 성능 평가를 통해 구축된 데이터 및 제안하는 모델의 적정성을 확인하였다.

후 기

이 논문은 2024년 정부의 재원으로 수행된 연구 결과임.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A.

Zisserman, “Deep audio-visual speech recognition,” IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 44, Issue. 12, pp. 8717-8727, 2018.

[2] S. Petridis, M. Pantic, “Deep complementary bottleneck features for visual speech recognition,” ICASSP, pp. 2304-2308, 2016.

[3] M. Wand, J. Koutn, J. Schmidhuber, “Lipreading with long short-term memory,” ICASSP, pp. 6115-6119, 2016.

[4] S. O. Kim, K. H. Lee, “Design & implementation of speechreading system using the face feature on the korean 8 vowels,” Korea Society of Computer and Information Winter Conference pp. 135-140, 2008.

[5] M. A. Lee, “A lip-reading algorithm using optical flow and properties of articulatory phonation,” Journal of Korea Multimedia Society, Vol. 21, No. 7, pp. 745-754, 2018.

[6] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, S. Zafeiriou, “RetinaFace: Single-shot multi-level dense localisation in the wild,” CVPR, pp. 5203-5212, 2020.

[7] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, “Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks,” ICML, pp. 369-376, 2006.

[8] T. Stafylakis, G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” Interspeech, pp. 3652-3656, 2017.

[9] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” CVPR, pp. 770-778, 2016.

[10] J. S. Chung, A. Zisserman, “Lip reading in the wild,” ACCV, pp. 87-103, 2016.

[11] J. S. Chung, A. Senior, O. Vinyals, A. Zisserman, “Lip reading sentences in the wild,” CVPR, 2017.

[12] J. S. Chung, A. Zisserman, “Lip reading in profile,” BMVC, pp. 155.1-155.11, 2017.

[13] T. Afouras, J. S. Chung, A. Zisserman, “LRS3-TED: a large-scale dataset for visual speech recognition,” In arXiv preprint arXiv:1809.00496, 2018.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, “SSD: Single shot multibox detector,” ECCV, pp. 21-37, 2016.

[15] J. Yuan, M. Liberman, “Speaker identification on the scotus corpus,” Journal of the Acoustical Society of America, Vol. 123, No. 5, pp. 3878-3882, 2008.

[16] J. S. Chung, A. Zisserman, “Out of time: automated lip sync in the wild,” In Workshop on Multi-view Lip-reading, ACCV, pp. 251-263, 2016.

[17] T. Stafylakis, G. Tzimiropoulos, “Combining residual networks with lstms for lipreading,” Interspeech, 2017.

[18] Y. M. Assael, B. Shillingford, S. Whiteson, N. Freitas, “Lipnet: Sentence-level lipreading,” arXiv: 1611.01559, 2016.

[19] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, N. Freitas, “Large-scale visual speech recognition,” Interspeech, pp. 4134-4139, 2019.

[20] X. Zhang, F. Cheng, S. Wang, “Spatio-temporal fusion based convolutional sequence learning for lip reading,” ICCV, pp. 713-722, 2019.

[21] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” Interspeech, pp. 5036-5040, 2020.

[22] P. Ma, S. Petridis, M. Pantic, “End-to-end autio-visual speech recognition with conformers,” ICASSP, pp. 7613-7617, 2021.

[23] H. Dinkel, S. Wang, X. Xu, M. Wu, K. Yu, “Voice activity detection in the wild: a data-driven approach using teacher-student training,” IEEE/ACM Trans. on Audio, Speech and Language Processing, Vol. 29, pp. 1542-1555, 2021.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using Kaldi,” Interspeech, pp. 498-502, 2017.

[25] S. S. Yoon, T. Y. Chun, D.-J. Jung, H. S. Song, “A study on data preprocessing for lip-reading of national defense data,” KIMST Annual Conference, pp. 367-368, 2022.