

## Design and Implementation of Sentence-Level Lip-reading with a Korean Morpheme-Based Multimodal AVSR Model (KM-AVSR)

Hee-Dong Yoon\*, Se-Uk Lee\*\*, Dong-Kyu Moon\*\*, Myung-Ho Kim\*\*\*

\*Ph.D Candidate, Dept. of IT Policy and Management, Soongsil University, Seoul, Korea

\*\*Researcher, IT Research Institute, DeakyoCNS Co., Ltd., Seoul, Korea

\*\*\*Professor, Dept. of IT Policy and Management, Soongsil University, Seoul, Korea

### [Abstract]

In this paper, we propose **KM-AVSR**, a **Korean Morpheme-based Multimodal Audio-Visual Speech Recognition (AVSR)** model designed to enhance sentence-level lip-reading accuracy. Lip-reading has become increasingly valuable for understanding speech in noisy environments or in the absence of audio, with promising applications in Korean language education, assistive technologies, and surveillance. To address the challenges posed by the syllabic and agglutinative nature of Korean, **KM-AVSR** adopts morpheme-based subword tokenization. The model independently encodes visual (lip movements) and auditory (raw waveform) inputs using separate encoders, fuses the modalities through a multilayer perceptron, and decodes the output using a hybrid **Connectionist Temporal Classification (CTC)** and Transformer-based decoder. Evaluations on a Korean lip-reading dataset demonstrate that **KM-AVSR** achieves a **Character Error Rate (CER)** of **15.66%**, representing a **39.35% improvement** over a conventional CNN-based AVSR model. These results highlight the effectiveness of morpheme-level subword modeling and hybrid decoding in Korean AVSR.

▶ **Key words:** Korean lip-reading, Audio-visual speech recognition, Morpheme-based subwords, Deep learning, Natural language processing

### [요약]

본 연구는 음성 정보가 결손되거나 소음 환경에 처한 상황에서도 문장 수준의 발화를 예측할 수 있는 립리딩 기술의 필요성에 주목한다. 립리딩은 특히 한국어 교육, 청각장애 지원, 음성 인식 보완 등 다양한 분야에서 활용 가능성이 높으며, 이를 구현하기 위한 멀티모달 시청각 음성 인식 (Audio-Visual Speech Recognition, AVSR) 기술이 각광받고 있다. 본 연구는 한국어 립리딩의 구조적 특성과 음절 기반 문자 체계를 고려하여, 의미 단위의 예측이 가능한 형태소 기반 하위 단어 (subword) 토큰화를 도입한 **KM-AVSR(Korean Morpheme-based AVSR)** 모델을 제안한다. 한국어 립리딩 대규모 데이터셋을 활용한 실험 결과, 제안된 **KM-AVSR**은 문자 오류율(Character Error Rate, CER) 15.66%를 기록하며 기존 모델 대비 약 39.35%의 성능 향상을 보였다. 이러한 결과는 형태소 기반 출력 단위와 하이브리드 디코딩 구조가 한국어 립리딩 정확도 향상에 효과적임을 시사한다.

▶ **주제어:** 한국어 립리딩, 시청각 음성인식, 형태소 기반 하위 단어, 딥러닝, 자연어 처리

- First Author: Hee-Dong Yoon, Corresponding Author: Myung-Ho Kim
- \*Hee-Dong Yoon (hdyoon@soongsil.ac.kr), Dept. of IT Policy and Management, Soongsil University
- \*\*Se-Uk Lee (seuk\_lee@daekyo.co.kr), IT Research Institute, DeakyoCNS Co., Ltd.
- \*\*Dong-Kyu Moon (dongkyu\_moon@daekyo.co.kr), IT Research Institute, DeakyoCNS Co., Ltd.
- \*\*\*Myung-Ho Kim (kmh@ssu.ac.kr), Dept. of IT Policy and Management, Soongsil University
- Received: 2025. 06. 18, Revised: 2025. 07. 04, Accepted: 2025. 08. 04.

## I. Introduction

음성 인식 기술은 다양한 산업 분야에서 인간-기계 간 상호작용을 혁신적으로 개선해왔으나, 여전히 소음 환경, 음성 결손, 다중 화자 상황과 같은 실제 환경에서는 성능 저하의 한계에 직면한다. 이러한 한계를 극복하기 위해 시청각 기반의 음성 인식(Audio-visual Speech Recognition, AVSR) 기술이 각광받고 있다. 그 중 시각 정보 중에서도 입모양 영상을 활용한 립리딩 기술은 음성 정보 없이도 발화 내용을 예측할 수 있어 주목받고 있다 [1]. 립리딩은 화자의 입술 움직임과 표정 등 시각적 조음 정보를 분석하여 발화를 유추하는 기술로, 한국어 발화 교육 및 평가, 청각 장애인 지원, 회의 녹취, 인공지능 비서 등 다양한 실사용 도메인에서 그 활용성이 높아지고 있다 [2].

특히 한국어는 초성-중성-종성으로 구성된 음절 문자 체계와 다양한 문법 변형을 수반하는 교착어라는 구조적 특성을 갖고 있어, 기존의 영어 기반 립리딩 접근 방식을 그대로 적용하는 데 한계가 존재한다 [3][4]. 예를 들면, ‘ㅂ’과 ‘ㅍ’, ‘ㄱ’과 ‘ㅋ’과 같은 시각적으로 유사한 조음 구조로 인해 자소 단위의 예측 정확도는 낮을 수 있으며, 조사나 어미 등 다양한 형태소가 결합된 문장 구조는 문맥 정보를 충분히 활용하지 못하면 자연스러운 문장 생성이 어렵다. 이러한 이유로 인해 한국어 립리딩 시스템은 자소 또는 음소 기반보다는 문법적 정보와 의미 단위를 반영할 수 있는 형태소 기반 접근이 더욱 적합하다는 주장이 제기되고 있다 [5].

본 연구에서는 이러한 문제를 해결하고자, 한국어의 언어적 특성을 반영한 형태소 기반 멀티모달 AVSR 모델인 KM-AVSR(Korean Morpheme-based Audio-Visual Speech Recognition)을 제안한다. 제안 모델은 RGB 기반의 입모양 영상과 원시 음성 파형을 각각 시각 및 청각 인코더로 독립적으로 처리하고, 다층 퍼셉트론(MLP)을 통해 정보를 융합한 후, 하이브리드 디코더(CTC + Transformer)를 통해 문장 수준의 형태소 기반 하위 단어 시퀀스를 생성한다. 출력 단위는 SentencePiece 기반 Byte-Pair Encoding(BPE)을 통해 생성된 1,207개의 형태소 기반 하위 단어로 구성되며, 이는 자소보다 의미 단위를 보존하면서도 미학습 표현에 대한 일반화 가능성을 확보할 수 있도록 설계되었다 [6][7].

KM-AVSR은 전이 학습을 기반으로 시각 및 청각 인코더에 영어 기반 Auto-AVSR 모델의 사전 학습 가중치를 초기화 값으로 적용하였으며, 한국어 립리딩 대규모 데이

터셋을 이용하여 미세 조정(fine-tuning)을 진행함으로써 한국어 특화 성능을 확보하였다. 실험 결과, 제안 모델은 기존 CNN + Mel-Spectrogram 기반 AVSR 모델 대비 문자 오류율(Character Error Rate, CER) 기준 약 39.35%의 성능 향상을 보였다. 이는 형태소 기반 출력 단위와 멀티모달 하이브리드 디코딩 전략이 한국어 립리딩 정확도 향상에 효과적임을 실증적으로 입증하는 결과이다.

## II. Preliminaries

### 1. Related works

#### 1.1 Overview of Lipreading Research

##### 1.1.1 Domestic Research Trends

국내 립리딩 연구는 한국어 고유의 언어적 특성과 실제 응용 가능성을 고려하여 최근 빠르게 발전하고 있다 [8]. 초기 연구들은 주로 자소 또는 음소 기반의 립리딩 시스템 개발에 집중하였으나, 한국어의 교착어적 구조와 초성-중성-종성 음절 체계로 인해 시각적 유사성(예: ‘ㅂ’과 ‘ㅍ’, ‘ㄱ’과 ‘ㅋ’ 등)에서 오는 예측의 어려움이 지속적으로 제기되어 왔다.

이에 따라 최근에는 형태소 단위의 예측, 서브워드 기반 립리딩 그리고 실시간 음성 인식 보조 시스템과의 융합 등 한국어 형태론적 특성을 반영한 AVSR 시스템 개발이 활발히 이루어지고 있다 [3]. 특히, 실제 환경의 다양한 발화 변동성과 소음 조건을 고려한 데이터셋 구축과, 실시간 인터랙티브 시스템 적용 가능성에 대한 연구가 증가하고 있다 [4].

이러한 흐름은 한국어 립리딩에서의 시각적 모호성과 문법적 다양성 문제를 해결하기 위한 형태소 기반 AVSR 접근의 중요성을 부각시키고 있으며, 문장 수준의 립리딩 예측으로 연구 범위가 확장되고 있다 [9].

##### 1.1.2 International Research Trends

국외에서는 시청각 기반 음성 인식 분야가 심층 신경망 기반의 대규모 AVSR 모델을 중심으로 빠르게 발전해왔다. 대표적으로 DeepMind의 WLAS(Watch, Listen, Attend and Spell) [10]는 시각-청각 정보를 통합하는 종단형 구조를 도입하여, 대규모 영어 방송 데이터를 활용한 문장 수준 립리딩의 가능성을 입증하였다.

이러한 구조는 이후 Auto-AVSR [11] 등 다양한 AVSR 아키텍처의 기반이 되었으며, CTC와 Transformer 등 하

이브리드 디코더의 도입으로 문맥 정보와 시간 정렬 정보를 동시에 활용하는 방식이 표준화되고 있다. 한편, 중국어(만다린) 기반의 CMLR 데이터셋과 LIBS 시스템 [12]은 형태소 및 음소 수준의 토큰화 전략을 도입하여, 소량의 데이터로도 일반화 가능한 립리딩 성능을 구현하였다.

특히, 데이터 자원이 제한된 언어에 적합한 데이터 효율적 모델 설계와 제로샷 학습 [13] 접근이 강조되고 있다. 이러한 글로벌 연구 동향은 한국어와 같은 교착어 계열 언어에 AVSR 시스템을 설계할 때, 언어 특성에 맞는 토큰화 및 하이브리드 디코딩 구조의 필요성을 시사하며, 본 연구의 KM-AVSR 설계에 중요한 영감을 제공하였다 [14].

## 1.2 Limitations of Prior Lipreading Research and the Novelty of This Study

기존 립리딩 연구는 주로 영어권 성인 화자를 대상으로 한 대규모 병렬 영상-음성 코퍼스(LRS2, LRS3, GRID 등)와 음소/자소 기반 예측에 집중되어 왔다 [10][11]. 이러한 모델들은 영어의 음운 및 문법 체계에 최적화되어 있어, 한국어와 같은 교착어적 구조를 가진 언어에 직접적으로 적용하기에는 한계가 있다. 한국어는 초성-중성-종성 조합의 음절 문자 체계와, 다양한 형태소 결합에 기반한 문법 변형이 빈번하여, 단순한 음소 또는 자소 기반 립리딩 접근으로는 시각적 구분의 어려움과 문맥 정보 부족을 극복하기 어렵다. 국내 립리딩 연구 역시 대부분 자소 기반 또는 단어 단위 예측에 국한되어 있으며, 문장 수준의 립리딩 및 실제 환경의 다양한 변동성을 반영한 연구는 미흡한 실정이다 [15].

이러한 한계를 극복하고자, 한국어의 언어적 특성에 최적화된 형태소 기반 하위 단어 단위를 출력 단위로 채택하고, 시각 및 청각 정보를 각각 독립적으로 인코딩한 후 하이브리드 디코더(CTC + Transformer) 구조를 통해 문맥과 시간 정렬 정보를 동시에 학습하는 KM-AVSR 모델을 제안한다. 형태소 기반 하위 단어 디코딩은 한국어의 교착어적 구조와 다양한 문법 변형을 효과적으로 반영할 수 있으며, 자소 기반 접근 대비 시각적 모호성 문제를 완화하고, 문장 수준의 자연스러운 예측을 가능하게 한다.

KM-AVSR은 한국어 립리딩 분야에서 형태소 기반 멀티모달 AVSR 모델의 필요성과 효과를 실증적으로 제시하는 최초의 연구 중 하나로 [16][17], 기존 연구와의 명확한 차별성을 갖는다. KM-AVSR은 형태소 기반 하위 단어 예측 구조와 하이브리드 디코더를 통해 한국어 립리딩 과제에 특화된 새로운 방향을 제시하는 선도적 연구이다.

## 2. Theoretical Background and Linguistically-Informed Model Design

본 절에서는 제안하는 KM-AVSR 모델의 설계에 반영된 이론적 기반과 언어학적 동기를 설명한다. 한국어는 음절 단위의 문자 체계와 교착어적 형태를 가진 언어로, 이러한 구조적 특성은 립리딩 모델의 설계 시 고려되어야 한다. 이에 따라 본 절에서는 다음 세 가지 핵심 요소를 중심으로 설명을 전개한다: 첫째, 한국어의 언어 특성이 모델 설계에 미치는 영향, 둘째, 형태소 기반 하위 단어 토큰화의 필요성과 장점, 셋째, 한국어 립리딩을 위한 하이브리드 디코더 구조 채택의 타당성.

### 2.1 Language-Specific Design Considerations

립리딩 모델의 설계에서 언어의 형태론적 및 음운론적 특성은 출력 단위의 선택, 예측 전략, 학습 안정성 및 일반화 성능에 핵심적으로 작용한다. 한국어는 초성-중성-종성으로 결합되는 음절 문자 체계와 다양한 형태소 조합을 특징으로 하는 교착어(agglutinative structure)로, 자소 단위 립리딩은 시각적 유사성(예: ‘ㅂ’ vs. ‘ㅍ’, ‘ㄷ’ vs. ‘ㅌ’)으로 인한 고유의 한계가 존재한다. 이러한 언어적 특성은 립리딩 성능 향상을 위해 의미 단위에 기반한 고차원의 출력 표현이 필요함을 시사한다 [18].

본 연구에서 제안하는 KM-AVSR(Korean Morpheme-based AVSR) 모델은 한국어의 형태론적 및 음운론적 특성을 정밀하게 분석하고, 그 결과를 모델 설계에 적극적으로 반영하였다. 구체적으로, KM-AVSR은 morpheme-based subwords(형태소 기반 하위 단어)를 출력 단위로 채택하여 의미 단위 보존, 문맥 정보 반영, 미학습 조합에 대한 일반화 가능성을 동시에 확보한다.

### 2.2 Advantages of Morpheme-Based Subword Units

형태소 기반 하위 단어 단위는 자소보다 예측 단위가 크면서도 단어 수준보다는 세분화되어 데이터 희소성을 완화하고, 한국어의 다양한 형태소 결합과 문법 변형을 효과적으로 모델링할 수 있도록 한다. 이러한 단위는 의미 단위 보존, 문맥 정보 반영, 미학습 조합에 대한 일반화 가능성 등 여러 가지 장점을 제공한다. 형태소 기반 하위 단어 단위는 한국어의 문법적 정보(조사, 어미, 접사 등)와 구조적 패턴을 학습하는 데 효과적이며 [9], Transformer 디코더와 결합될 경우 문맥 정보 예측에 매우 유리하다. 최근 립리딩 관련 연구에서도 형태소 기반 토큰 단위의 효과가 다수 보고되었으며 [16][19], 특히 문장 수준 립리딩에

있어 이 단위는 의미 보존과 정합성 측면에서 가장 적합한 출력 단위로 간주된다.

### 2.3 Hybrid Decoder Design for Korean Lipreading

KM-AVSR의 하이브리드 디코딩 구조(CTC + Transformer)는 한국어 문장의 정렬 정보와 문맥적 연속성을 동시에 학습할 수 있도록 설계되었다 [19][20]. CTC 디코더는 시계열 프레임 정렬을 학습하여 빠른 수렴과 초기 학습 안정성을 제공하고, Transformer 디코더는 문장 구조와 의미 흐름을 고려한 예측을 가능하게 하여 자연스러운 문장 생성을 지원한다. 이러한 하이브리드 구조는 다양한 형태소 결합이 빈번한 한국어의 agglutinative structure에 최적화되어 있다.

이와 같이, KM-AVSR 모델은 morpheme-based subword tokenization을 통해 한국어의 구조적 특성과 시각적/음운적 모호성, 문법적 복잡성, 데이터 불균형 문제를 종합적으로 해결할 수 있도록 설계되었다. 이러한 접근은 KM-AVSR이 한국어 립리딩 과제에서 구조적 타당성과 실질적 성능을 동시에 확보하는 데 핵심적인 기반이 된다.

## III. The Proposed Scheme

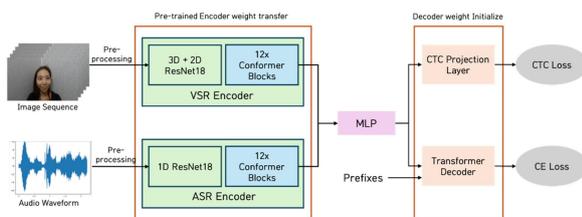


Fig. 1. Overall architecture of the proposed KM-AVSR model for Korean sentence-level lip-reading.

본 연구에서는 한국어 문장 수준 립리딩 예측을 위해 KM-AVSR(Korean Morpheme-based Audio-Visual Speech Recognition) 모델을 설계하고 구현한다. 제안된 KM-AVSR 모델은 RGB 기반의 입모양 영상과 원시 음성 파형을 각각 시각 및 청각 인코더를 통해 처리하고, 이를 융합하여 형태소 기반 하위 단어 시퀀스를 문장 단위로 예측한다.

전체 아키텍처는 그림 1과 같이 구성되며, 시각 인코더(Visual Encoder), 청각 인코더(Audio Encoder), 멀티모달 융합(Multimodal Fusion) 그리고 CTC 및

Transformer의 이중 디코더로 구성된다.

KM-AVSR의 시각 인코더는 3D CNN을 통해 시공간적 조음 특징을 추출하고, 2D ResNet-18을 통해 각 프레임의 세부 패턴을 정제하며, 12개의 Conformer 계층을 통해 시간적 의존성을 효과적으로 모델링한다. 이 구조는 연속된 입모양의 시공간적 변화를 효과적으로 포착하여 정교한 시각 표현을 인코딩한다.

청각 인코더는 원시 음성 파형을 1D ResNet-18과 12개의 Conformer 계층을 통해 처리함으로써 음향적 특성과 시간 의존성을 정밀하게 인코딩한다. 이로써 시각 정보와 상호 보완적인 청각 표현을 확보할 수 있다.

두 인코더의 출력은 시간 축을 기준으로 정렬되며, MLP 계층을 통해 통합된다. 이 멀티모달 임베딩은 각 모달리티의 정보가 정규화된 상태로 결합되어, 하위 디코더에서 활용 가능한 풍부한 표현을 형성한다.

KM-AVSR의 디코더는 두 경로로 구성된다. 첫째, CTC 디코더는 시계열 정렬 기반의 학습을 통해 입력과 출력 간의 프레임 매핑을 수행하며 초기 학습 안정성에 기여한다. 둘째, Transformer 디코더는 self-attention 기반 문맥 예측을 수행하며, 긴 시퀀스에 걸친 언어적 일관성을 보장한다. 이 하이브리드 구조는 두 방식의 장점을 동시에 활용하여 최적의 예측 성능을 도출한다.

학습 시에는 CTC 손실과 교차 엔트로피 손실을 결합한 하이브리드 손실 함수를 사용하며, 가중치는 0.3으로 설정된다. 이 하이퍼파라미터는 두 손실 간 균형을 조정하여 시간 정렬 기반 학습과 문맥 기반 예측이 조화롭게 이루어지도록 한다.

출력 시퀀스는 SentencePiece 기반의 BPE 알고리즘을 이용해 1,207개의 형태소 기반 하위 단어 단위로 토큰화된다. 이 단위는 한국어의 형태론적 및 음운론적 특성을 반영하며, 발화 오류와 음운 변이에 유연하게 대응할 수 있다.

추론 단계에서는 Transformer 디코더만을 사용하며, 이는 문맥 기반의 문장 생성에서 더 높은 예측 성능과 일관된 언어 흐름을 제공하기 때문이다. 빔 서치(beam search) 알고리즘(빔 너비 5)을 적용하여 <sos>, <eos> 토큰까지의 서브워드 시퀀스를 생성한다. 최종 결과는 문자 오류율(CER)을 기준으로 평가된다.

제안된 모델은 한국어의 교착어적 구조와 복잡한 조음 특성, 그리고 문맥 정보를 효과적으로 학습할 수 있도록 설계되었으며, 이는 문장 수준 립리딩 예측을 위한 멀티모달 AVSR 구조의 새로운 기준을 제시한다.

## 1. Data Pre-processing

본 연구에서는 한국어 화자의 입모양 및 음성 데이터를 멀티모달 입력 시퀀스로 활용하기 위하여, 시각 정보와 청각 정보를 각각의 특성과 목적에 맞게 체계적으로 전처리 하였다. 시각 정보는 정제되고 정렬된 입모양 영상 시퀀스로, 청각 정보는 정규화된 원시 음성 파형 시퀀스로 처리 된다. 두 입력 시퀀스는 모두 문장 단위로 시간 정렬되어 멀티모달 통합 과정의 기반이 된다. 본 절에서는 각 스트림별 전처리 절차와, 멀티모달 통합을 위한 정렬 및 정규화 과정의 중요성을 중점적으로 기술한다.

### 1.1 Visual Stream

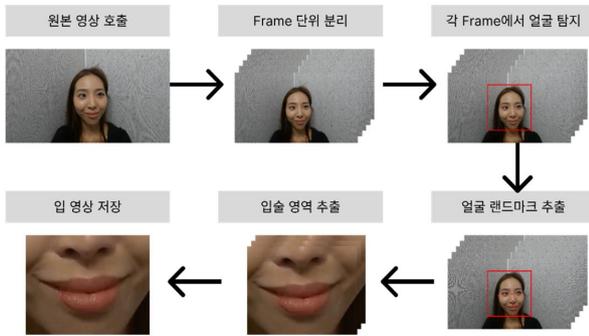


Fig. 2. Visual stream preprocessing pipeline.

시각 입력은 화자의 RGB 기반 정면 영상으로부터 구성되며, 정규화된 입모양 시퀀스를 추출하기 위해 그림 2에 도시된 전처리 파이프라인을 따른다. 먼저, 원본 입력 시퀀스는 발화 구간 또는 자막 정보를 기준으로 문장 단위로 분할되고, 이후 프레임 단위로 세분화된다. 이어서 RetinaFace [21] 또는 MediaPipe [22]와 같은 얼굴 검출 알고리즘을 활용하여 얼굴 영역을 탐지한 후, 얼굴의 방향성과 위치의 일관성을 확보하기 위해, 눈과 코의 랜드마크를 기준으로 정렬 과정을 수행한다.

이후 랜드마크 기반으로 입모양 영역(ROI)을 지정한 뒤, 96x96 픽셀 해상도로 잘라낸 정규화된 입모양 영상 이미지를 시퀀스로 구성한다. 모든 프레임은 평균과 표준편차를 기준으로 정규화되며, 이는 입력 시퀀스 간 휘도 및 조명 차이를 최소화하여 모델의 학습 안정성과 일반화 성능을 향상시킨다. 이와 같이 정렬 및 정규화 과정은 멀티모달 입력 시퀀스 간의 일관된 표현 확보 및 융합 단계의 효과적 통합에 핵심적인 역할을 한다.

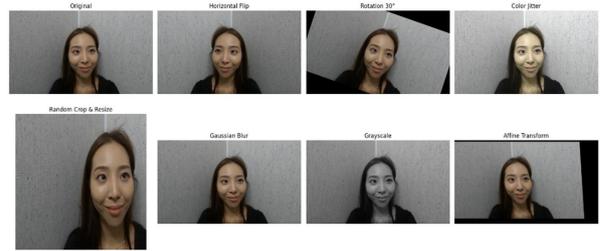


Fig. 3. Examples of visual data augmentation.

더불어, 학습 데이터의 다양성 및 일반화 성능 제고를 위하여 그림 3과 같이 수평 반전(horizontal flipping), 무작위 추출(random cropping), 프레임 위치 조정(frame shift) 등 다양한 시각적 데이터 증강(data augmentation) 기법을 적용한다.

### 1.2 Audio Stream

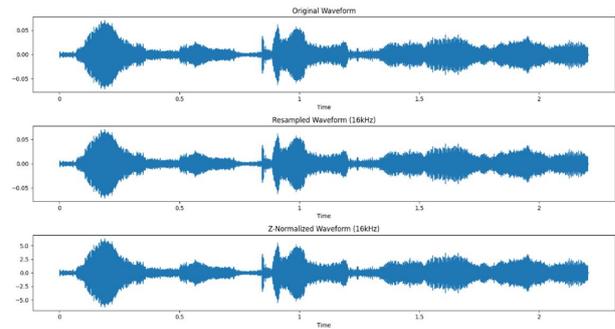


Fig. 4. Audio stream preprocessing pipeline.

청각 입력은 각 문장 단위로 분할된 입력 시퀀스에서 음성 신호를 추출하여 구성하며, 전처리 과정은 그림 4에 도시된 바와 같이 이루어진다. 음성 신호는 16kHz로 리샘플링되고, 별도의 음향 특징(MFCC, Spectrogram 등)을 추출하지 않고, 음성 인식 모델에 직접 입력되는 형태로 원시 음성 파형 자체를 활용한다.

모델의 학습 안정성과 다양한 녹음 조건에 대한 적응력을 확보하기 위하여 z-normalization을 적용하여 발화 강도에 따른 음압 차이를 보정한다. 또한, 실제 환경에서의 조음 누락, 잡음, 강제 변화 등 다양한 발화 변형에 대한 모델의 강건성을 강화하기 위해, 발화 구간 내 일부 영역을 무작위로 마스킹하는 adaptive time masking 기법을 적용하여, 모델의 강건성을 강화한다.

시각 및 청각 입력 시퀀스는 모두 25fps 기준으로 시간 정렬되며, 동일 시점에서 시작하여 종료되도록 구성된다. 이로써 두 모달리티 간의 시계열 정합성(temporal alignment)이 확보되어, 이후 멀티모달 융합 및 디코딩 단계에서 안정적인 시계열 표현 통합이 가능해진다.

## 2. Target Tokenization

멀티모달 립리딩 예측 모델에서 출력 단위의 선택은 모델의 예측 정확도, 일반화 성능, 그리고 학습 안정성에 중대한 영향을 미친다. 특히 한국어와 같이 초성-중성-종성으로 구성된 음절 문자 체계를 갖는 교착어에서는, 자소나 음소 단위의 예측이 시각적으로 유사한 조음 구조로 인해 높은 오류율을 유발할 수 있다. 예컨대, ‘ㄴ’과 ‘ㄹ’, ‘ㄱ’과 ‘ㅋ’은 입모양만으로는 명확히 구분하기 어려운 자음이며, 이는 립리딩 기반 예측의 정밀도를 저해한다.

본 연구에서는 이러한 언어적 특성과 립리딩 입력의 제약을 고려하여, 문장 수준 립리딩 예측의 출력 단위로 형태소 기반 하위 단어 단위를 채택하였다. 이는 단순한 음소 단위보다 의미 중심의 표현을 가능하게 하며, 단어 수준보다 더 세분화되어 다양한 조합을 통해 미학습 표현에 대한 일반화도 가능하게 한다.

서브워드 단위 구축을 위해 SentencePiece의 BPE (Byte-Pair Encoding) 알고리즘을 적용하였고, 총 1,207개의 형태소 기반 하위 단어 어휘 집합을 생성하였다 [6]. 이 단위는 다음과 같은 장점을 지닌다:

- 발화 오류나 음운 변이에 유연하게 대응 가능
- 자소 기반보다 의미 단위를 보존하며, 단어 기반보다 조합 가능한 표현이 풍부
- 학습 데이터에 존재하지 않는 표현도 기존 서브워드 조합을 통해 예측 가능

또한 형태소 기반 하위 단어 단위는 한국어의 문법적 정보(조사, 어미, 접사 등)와 구조적 패턴을 학습하는 데 효과적이며, Transformer 디코더와 결합될 경우 문맥 정보 예측에 매우 유리하다. 최근 립리딩 관련 연구에서도 형태소 기반 토큰 단위의 효과가 다수 보고되었으며, 특히 문장 수준 립리딩에 있어 이 단위는 의미 보존과 정합성 측면에서 가장 적합한 출력 단위로 간주된다.

Table 1. Examples of Korean tokenization levels for the sentence "나는 학교에 간다".

tokenization unit	Description	Example
Word	Space-based segmentation	나는/학교에/간다
Morpheme	Smallest meaningful unit	나/는/학/교/에/가/ㄴ/다
Syllable	Korean syllables	나/는/학/교/에/간/다
Phoneme	Initial, medial, final sounds	ㄴ/ㅏ/ㅍ/ㄹ/ㅡ/ㄴ/ㅎ /ㅏ/ㄱ..

표 1은 문장 "나는 학교에 간다"에 대해 다양한 토큰화 수준을 적용한 예시를 제시하며, 각 단위의 정보 보존과

예측 정밀도 영향을 비교한다. 단어 단위는 의미 단위 전 반을 유지하지만, 한국어의 형태적 변형(어미, 조사 등)로 인해 OOV(Out-of-Vocabulary) 발생률이 높고, 새로운 형태에 대한 일반화가 어렵다. 음절 및 자소 단위는 OOV 문제를 줄이는 데 도움이 되지만, 너무 세분화되어 문맥 및 의미 정보가 축소되어 성능이 저하될 수 있다.

반면, 형태소 기반 단위는 의미를 가진 최소 단위로, 어미, 조사 및 어간 등 한국어 특유의 형태 변화를 반영하며, OOV 발생률을 효과적으로 낮추면서 정보 보존에도 유리하다. 실제로 한국어 대용량 ASR에서 형태소 기반 단위 사용시 OOV 비율을 영어 수준으로 낮출 수 있음을 보고하였으며 [24], subword 기반 문장 임베딩 실험에서 OOV 대응력과 문장 표현 정확도 측면에서 형태소 기반 하위 단어가 음절 및 자소 기법보다 우수하다는 것을 확인하였다 [25]. 또한 형태소 인식 정보를 반영한 BPE기반 subword tokenization이 구조적 의미 표현과 예측 정밀도를 향상 시킴을 보고하였다 [26].

이처럼 OOV 대응, 문맥/의미 보존, 예측 정밀도 세 가지 기준에서 형태소 기반 하위 단어 단위가 한국어 입모양 발화 예측에 가장 적합한 해법임이 이론적 및 실험적으로 뒷받침된다. 따라서 본 연구는 형태소 기반 하위 단어를 최적의 토큰화 단위로 채택하였으며, 이는 제안 모델의 핵심 구조 설계 방향과도 일치한다.

## 3. Model Architecture

본 연구에서는 멀티모달 입력(입모양 영상과 음성 신호)을 받아 형태소 기반 하위 단어 시퀀스를 예측하여 문장 수준 립리딩 예측을 수행하는 종단형 구조의 한국어 입모양 립리딩 예측 모델을 구현한다. 전체 구조는 그림 1에 요약되어 있으며, 인코더 단계에서는 시각 및 청각 입력을 각각 전방 네트워크로 처리하고, 후방에서는 융합 표현을 기반으로 이중 디코더 경로를 통해 서브워드 시퀀스를 생성한다.

### 3.1 Audio-Visual Feature Extraction and Concatenation

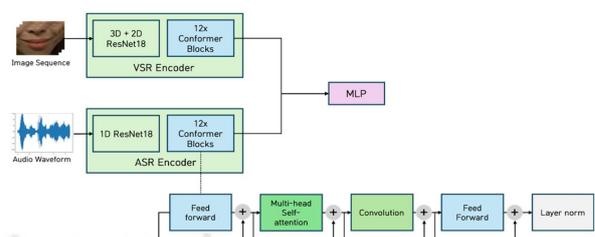


Fig. 5. Audio-Visual Speech Recognition Encoder architecture.

멀티모달 입력은 RGB 기반의 입모양 영상과 원시 음성 파형으로 구성되며, 각각 그림 5와 같이 시각 인코더(VSR Encoder)와 청각 인코더(ASR Encoder)를 통해 인코딩된다.

- 시각 인코더는 3차원 합성곱 계층(3D-CNN)으로 시공간적 조음 특징을 추출한 후, 2D ResNet-18 블록과 12개의 Conformer 계층을 통해 시계열 정보를 정제한다.

- 청각 인코더는 1D 합성곱 기반 ResNet-18과 12개의 Conformer 계층으로 구성되며, 음색 및 주파수 기반의 시간 정보를 효과적으로 인코딩한다. 모든 오디오 표현은 25fps로 샘플링을 고정하여 시각 표현과 시간 정렬을 유지한다.

두 인코더의 출력은 시간 축을 기준으로 정렬된 후, MLP 구조를 통해 하나의 통합된 시퀀스 표현으로 투영된다. 이 표현은 영상과 음성 정보를 모두 포함한 시계열 임베딩이며, 이후 디코딩 단계의 입력으로 사용된다.

이러한 멀티모달 인코더 설계는 소음 환경에서도 음성 정보가 손상될 수 있는 한계를 시각 인코더의 입모양 영상 특징이 보완하도록 의도되었다. 즉, 시각 정보와 음성 정보를 통합함으로써 소음 환경에서도 더 높은 예측 정밀도를 유지할 수 있는 강건성을 확보할 수 있다.

### 3.2 Dual Decoding and Inference Strategy

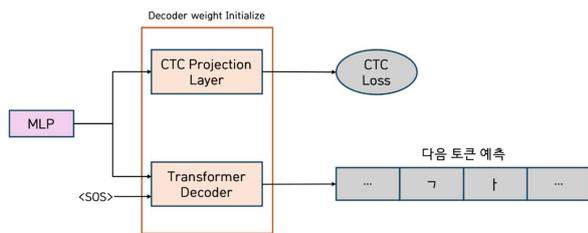


Fig. 6. Audio-Visual Decoder Process.

통합된 멀티모달 표현은 그림 6과 같이 MLP의 출력을 CTC 디코더 경로와 Transformer 디코더 경로의 두 가지 디코딩 경로를 통해 병렬적으로 처리된다.

- CTC 디코더 경로는 시간 축 상에서 독립적인 프레임별 예측을 수행하며, 발화와 정렬되지 않은 상태에서도 효과적인 학습이 가능하다. 초기 학습 안정성과 단기 의존성 모델링에 강점을 갖는다.

- Transformer 디코더 경로는 위치 인코딩과 self-attention 기반 구조를 통해 문맥 정보를 활용한 예측을 수행한다. 이전에 생성된 토큰과 인코더 출력을 입력으로 받아, 순차적으로 다음 토큰을 생성하며, 장기 의존 관계와 문장 구조를 효과적으로 학습할 수 있다.

학습 과정에서는 CTC 손실 함수와 교차 엔트로피 (Cross Entropy, CE) 손실 함수를 결합한 하이브리드 손실 구조를 사용한다. 이를 통해 시간 정렬 기반 학습과 문맥 기반 예측의 장점을 모두 활용한다. 학습 손실 함수는 CTC 손실( $Loss_{CTC}$ )과 교차 엔트로피 손실( $Loss_{CE}$ )을 가중 평균하여 다음과 같이 정의된다:

$$Loss_{total} = \alpha \cdot Loss_{CTC} + (1 - \alpha) \cdot Loss_{CE} \quad (1)$$

여기서  $\alpha$ 는 두 손실 간의 가중치를 조절하는 하이퍼파라미터로, 본 연구에서는  $\alpha = 0.1$ 으로 설정하였다.

추론 시에는 Transformer 디코더만을 사용하며, 빔 탐색(Beam Search) 알고리즘을 적용하여 <sos>, <eos> 토큰까지의 형태소 기반 하위 단어 시퀀스를 생성한다. 빔 너비(beam width)는 5로 설정하였으며, 누적 확률이 가장 높은 시퀀스가 최종 예측으로 선택된다.

### 4. Transfer Learning and Model Initialization

최근 립리딩 연구에서는 언어에 종속되지 않는 시각 및 청각 표현을 사전 학습하고, 이를 새로운 언어로 전이하는 전이 학습(Transfer Learning) 기반 접근이 활발히 시도되고 있다 [12]. 본 연구에서도 이러한 접근을 반영하여, 영어 기반 Auto-AVSR 모델에서 사전 학습된 가중치를 활용해 한국어 립리딩 모델을 효과적으로 초기화하는 방법을 제안한다.

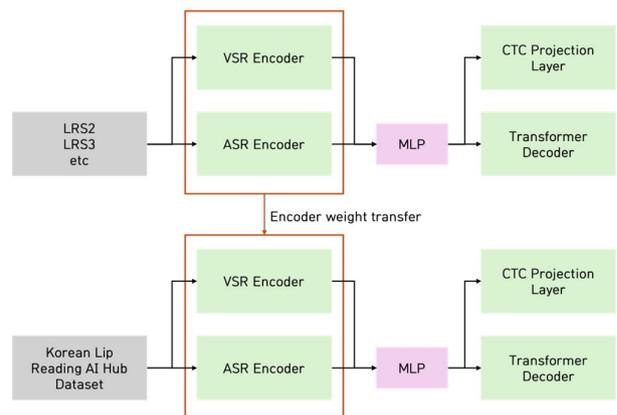


Fig. 7. Transfer learning and initialization strategy of the proposed model.

그림 7에 도시된 바와 같이, 시각 인코더(Visual Encoder)와 청각 인코더(Audio Encoder)는 영어 화자 기반의 대규모 시청각 데이터셋(LRS2, LRS3 등)에서 사전 학습된 가중치를 초기화 값으로 사용한다. 이 초기화는 언어 비의존적인 조음 및 음향 표현을 한국어 립리딩 과제로 안정적으로 전이하며, 초기 학습 수렴 속도와 성능 안

정성 향상에 기여한다.

반면, 디코더 모듈은 언어 고유의 문법과 형태소 구조를 반영해야 하므로, 한국어 형태소 기반 하위 단어 시퀀스를 효과적으로 처리할 수 있도록 Transformer 디코더 및 CTC Projection 계층은 새롭게 초기화한다. 이후 전체 모델은 한국어 립리딩 데이터셋을 기반으로 미세 조정(Fine-tuning)되어 최적화된다.

이러한 인코더-디코더 분리 초기화 전략은 다음과 같은 이점을 제공한다:

- 시각 및 청각 표현의 일반적인 저수준 특징을 효과적으로 전이
  - 한국어 특화 문법 및 형태소 구조를 반영한 고차원 예측 성능 확보
  - 다양한 연령대와 발화 특성에 대한 적응 가능성 향상
- 따라서 제안하는 전이 학습 전략은 한국어 립리딩 모델의 초기화를 효율화하고, 실질적인 문장 예측 성능을 높이는 데 기여할 수 있다.

## IV. Experimental Results

### 1. Dataset Description

본 연구에서 제안하는 KM-AVSR 모델의 한국어 립리딩 성능을 검증하기 위하여, 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 립리딩(입모양) 음성인식 데이터를 활용하였다 [23]. 해당 데이터셋은 다양한 연령, 성별, 직업군의 한국어 화자가 다양한 소음 환경에서 발화한 영상을 포함하며, 총 약 6,000명 규모로 구성으로 세부 분포는 표 2와 같으며, 본 연구에 활용된 데이터는 AI 허브에서 다운로드 받을 수 있다.

모든 영상은 사람을 정면에서 촬영되어진 데이터로 구성되었으며, 시간 정보에 따라 문장 단위로 분할되어 모델 학습 및 평가에 사용되었다. 이 데이터는 한국어 발화의 다양성과 소음 조건의 일반화를 반영하는 데 적합하며, 본 연구에서는 학습 및 평가 모두에 해당 데이터셋으로 8:2의 비율로 각 비율과 분포가 균일하게 구성하여 사용하였다.

### 2. Evaluation Metric

모델 성능 평가는 문자 오류율(CER)을 기준으로 수행하였다. CER은 예측된 시퀀스와 참조 문장 간의 삽입(Insertions), 삭제(Deletions), 대체(Substitutions) 오류를 기반으로 계산되며, 다음과 같이 정의된다.

$$CER = \frac{S + D + I}{N} \quad (2)$$

여기서  $S$ 는 대체 오류 수,  $D$ 는 삭제 오류 수,  $I$ 는 삽입 오류 수,  $N$ 은 정답 문장의 총 문자 수를 의미한다. CER은 발화의 세밀한 조음 차이와 발화 오류를 민감하게 반영하므로, 한국어 발화 평가에 적합한 지표이다.

Table 2. Demographic and Environmental Statistics of the Korean Lipreading Dataset.

Category	Attribute	Ratio Percentage
Noise Environment	No Noise	29%
	Daily Noise	14%
	Traffic Noise	14%
	Industrial Noise	14%
	Natural Noise	14%
	Other Noise	14%
Sex	Male	50%
	Female	50%
Speaker Type	General	75%
	Professional	25%
Age	Teens	10%
	20s	26%
	30s	25%
	40s	23%
	50s	8%
	60s	7%

### 3. Implementation Details

모델 학습은 NVIDIA A100 GPU 80GB RAM 환경에서 PyTorch 프레임워크를 기반으로 수행되었다. 최적화에는 AdamW 옵티마이저를 사용하였으며, 학습률은  $1e-3$ , warm-up epoch는 5, weight decay는 0.03으로 설정하였다. 전체 학습은 총 50 epoch 동안 진행되었다.

시각 및 청각 인코더의 은닉 차원(hidden dimension)은 768로 설정하였고, self-attention은 12개의 헤드로 구성된 multi-head attention 구조를 사용하였다. 인코더는 Macaron-style 인코더 구조, 상대적 위치 인코딩(relative positional encoding), Swish 활성화 함수, 커널 크기 31의 convolution 모듈이 적용되었으며, 시각 스트림과 청각 스트림의 입력 계층으로는 각각 conv3d와 conv1d를 사용하였다.

시청각 표현의 융합은 8,192 차원의 은닉 공간에서 batch normalization을 포함한 MLP 구조로 수행되었고, 추론 단계에서는 Beam Search를 사용하였으며 빔 너비(beam width)은 32로 설정하였다.

#### 4. Experimental Cases and Comparison

본 연구에서 제안하는 KM-AVSR 모델의 성능을 검증하기 위해 단일 실험 시나리오를 수행하였다. 실험은 한국어 대규모 립리딩 데이터를 학습과 평가에 모두 사용하는 구성으로, 제안 모델의 기본적인 예측 성능을 확인하는 데 목적이 있다.

비교 대상으로는 기존 CNN 기반 AVSR 구조(CNN 기반 시각 인코더 + Mel-spectrogram 기반 청각 인코더 + Transformer 디코더)를 사용한 모델을 설정하였으며, 동일한 학습 조건과 데이터셋에서 성능을 비교하였다. 실험 결과는 다음과 같다.

Table 3. Character Error Rate (CER) Comparison of Conventional AVSR and Proposed KM-AVSR Models.

Scenario	Method	CER(%)
Korean Lip-reading Data	CNN-based AVSR	25.81
	<b>KM-AVSR (Proposed)</b>	<b>15.66</b>

표 3의 결과는 제안하는 KM-AVSR 모델이 한국어 립리딩 과제에서 기존 CNN 기반 AVSR 대비 약 39.35%의 성능 향상(CER 감소)을 달성했음을 의미하며, 형태소 기반 하위 단어 예측 구조와 멀티모달 하이브리드 디코딩 전략의 유효성을 실증적으로 보여준다.

#### 5. Discussion

본 연구에서는 한국어 형태소 기반의 문장 발화 예측 멀티모달 AVSR 모델(KM-AVSR)이 립리딩 성능 향상에 효과적임을 실증적으로 보였다. 기존 CNN 기반의 AVSR 모델 대비 39.35% 향상된 문자 오류율(CER) 15.66%를 기록함으로써, 제안된 형태소 기반 하위 단어 단위 및 하이브리드 디코딩 구조의 유효성이 입증되었다.

형태소 기반 하위 단어는 의미 단위의 예측과 문장 구조 학습에 효과적이며, 자소 기반 단위의 시각적 모호성을 극복하는 데 기여하였다. 특히, SentencePiece 기반의 서브워드 분할 전략은 미학습 표현에 대한 일반화 능력을 향상시키는 데 중요한 역할을 하였다.

또한, 시각 및 청각 인코더의 독립적 설계와 Conformer 계층의 통합은 시공간적 조음 정보와 음향 정보를 장기 시계열 관점에서 효과적으로 모델링할 수 있게 하였으며, 멀티모달 융합을 통한 통합 표현은 문맥 기반 예측에서 Transformer 디코더의 성능을 극대화하였다.

한편, 한국어 대규모 립리딩 데이터셋을 단일 출처로 사

용하였기 때문에 다양한 도메인에 대한 일반화 가능성을 확정적으로 논의하기에는 한계가 존재한다. 향후에는 뉴스, 인터뷰, 강연 등 다양한 환경에서 수집된 실제 립리딩 영상에 대한 테스트가 필요하며, 이를 통해 모델의 실사용 가능성과 범용성을 더욱 강화할 수 있을 것이다.

## V. Conclusions

본 연구에서는 한국어 문장 수준 립리딩 예측을 위한 형태소 기반 멀티모달 AVSR 모델인 KM-AVSR을 제안하였다. 제안된 모델은 RGB 기반의 입모양 영상과 원시 음성 파형을 각각 전용 인코더에서 독립적으로 처리하고, 이를 MLP를 통해 융합한 후, 하이브리드 디코딩 구조(CTC + Transformer)를 통해 형태소 기반 하위 단어 시퀀스를 예측하는 구조를 갖는다. 또한, SentencePiece 기반 BPE 알고리즘을 활용하여 한국어의 교착어적 언어 구조에 적합한 1,207개의 형태소 기반 하위 단어 어휘를 구축하고 이를 출력 단위로 설정하였다.

한국어 대규모 립리딩 데이터를 기반으로 한 실험에서, 제안된 KM-AVSR 모델은 문자 오류율 15.66%를 기록하며, 기존 CNN + Mel-Spectrogram 기반 AVSR 모델(CER 25.81%) 대비 약 39.35% 향상된 성능을 달성하였다. 이는 형태소 기반 출력 단위와 멀티모달 하이브리드 디코딩 전략이 한국어 립리딩 정확도 향상에 효과적임을 실증적으로 입증한다.

향후 연구에서는 다양한 도메인의 립리딩 영상 데이터에 대한 테스트와, 한국어 특화 언어 모델 또는 사후 보정(post-processing) 기법과의 결합을 통해 실사용 가능성을 더욱 향상시키는 방안이 모색될 수 있다. 더불어, 본 연구에서 제안한 KM-AVSR 구조는 형태소 기반 언어 구조를 갖는 다른 언어로도 확장 가능성이 높으며, 다국어 립리딩 모델 설계의 기반 프레임워크로 활용될 수 있다.

## ACKNOWLEDGEMENT

This work was supported by the IT Research Institute of Daekyo CNS Co., Ltd. as part of its internal research program on AI-based multimodal learning systems.

## REFERENCES

- [1] J. Hong, M. Kim, D. Yoo, and Y. M. Ro, "Visual Context-driven Audio Feature Enhancement for Robust End-to-End Audio-Visual Speech Recognition," in Proc. Interspeech, 2022.
- [2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2346-2359, 2018. DOI: 10.1109/TPAMI.2017.2781401
- [3] J. S. Kim and S. Y. Lee, "Real-time lip reading system for isolated Korean word recognition," *Pattern Recognition*, vol. 36, no. 11, pp. 2731-2743, 2003.
- [4] H. Yang, S. Kim, D. Lee, and H. Kim, "Korean lip-reading with audio-visual fusion using deep neural networks," in Proc. Korea Computer Congress (KCC), 2020, pp. 450-452.
- [5] P. Ma, S. Petridis, and M. Pantic, "End-to-End Audio-Visual Speech Recognition with Conformers," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2021, pp. 7613-7617.
- [6] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66-71, 2018. DOI: 10.18653/v1/D18-2012
- [7] Y. Chung et al., "Byte-Pair Encoding is Suboptimal for Language Model Pretraining," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 108-113, 2020. DOI: 10.18653/v1/2020.aacl-main.17
- [8] D. Kim, Y. Lee, and S. Yoon, "Towards Korean Visual Speech Recognition: Dataset Construction and Baseline Evaluation," in Proc. Interspeech, 2023.
- [9] K. Park and J. Shin, "Effective Subword Tokenization for Korean Language Modeling," in *Proceedings of the Workshop on Subword and Character Level Models*, 2022.
- [10] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv preprint arXiv:1611.01599, 2016.
- [11] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2023, pp. 1-5.
- [12] Y. Lan, J. Liu, Z. Huang, S. Ma, and W. Hu, "Learning phoneme-level lip representations for Mandarin visual speech recognition," in Proc. ACM Int. Conf. Multimedia, 2021, pp. 2733-2741.
- [13] J. H. Yeo, M. Kim, C. W. Kim, S. Petridis, and Y. M. Ro, "Zero-AVSR: Zero-Shot Audio-Visual Speech Recognition with LLMs by Learning Language-Agnostic Speech Representations," arXiv preprint arXiv:2503.06273, 2025.
- [14] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in Proc. EUROSPEECH, 1997, pp. 713-716.
- [15] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2018, pp. 6548-6552.
- [16] S. Lee and J. Kim, "On the Effectiveness of Morpheme-based Subwords for Korean ASR," in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021.
- [17] Y. Yoon and K. Cho, "Challenges and Opportunities in Korean Multimodal Speech Recognition," in Proc. ACL Anthology Workshop, 2023.
- [18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, Apr. 1990. DOI: 10.1121/1.399423
- [19] S. Watanabe, T. Hori, S. Kim, J. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1240-1253, 2017.
- [20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," Proc. Interspeech, pp. 2613-2617, Sep. 2019. DOI: 10.21437/Interspeech.2019-2680
- [21] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage dense face localisation in the wild," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5203-5212, Jun. 2020. DOI: 10.1109/CVPR42600.2020.00525
- [22] F. Zhang, V. Bazarevsky, A. Vakunov, et al., "MediaPipe Hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020.
- [23] National Information Society Agency, "AI Hub: Lipreading and Audio-Visual Speech Recognition Data," <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=538>, 2020.
- [24] Kwon, O.-W. & Park, J. (2003). Korean large vocabulary continuous speech recognition with morpheme-based recognition units. *Speech Communication*, 39(3-4), 287-300...
- [25] Cho, D., Lee, H., & Kang, S. (2021). An empirical study of Korean sentence representation with various tokenizations. *Electronics*, 10(7), 845...
- [26] Jeon, T., Yang, B., Kim, C., & Lim, Y. (2023). Improving Korean NLP tasks with linguistically informed subword tokenization and sub-character decomposition. arXiv preprint arXiv:2311.03928.

## Authors



Hee-Dong Yoon received the B.S. degree in Computer Science from Kookmin University, Korea, in 2000, and the M.S. degree in Management Information Systems from Kookmin University, Korea, in 2011.

He is currently a Ph.D. candidate in IT Policy and Management at Soongsil University, Korea. Hee-Dong Yoon is currently a Ph.D. candidate in the Department of IT Policy and Management at Soongsil University, Seoul, Korea. His research interest include IT strategym digital transformation and AI-based decision support systemss in both public and enterprise domains.



Se-Uk Lee received the B.S. degree in Applied Statistics from Sejong University, Korea, in 2018, and the M.S. degree in Artificial Intelligence Convergence Technology from Soongsil University, Korea, in 2025

He is currently a researcher at the IT Research Institute of Daekyo CNS Co., Ltd. His research interests include large language models and affective computing with a focus on developing multimodal AI systems for education and speech understanding.



Dong-Kyu Moon received the B.S. degree in Electronic Engineering from Daejin University, Korea, in 2020, and the M.S. degree in Artificial Intelligence Convergence from Dankook University, Korea, in 2024

He is currently a researcher at the IT Research Institute of Daekyo CNS Co., Ltd. His research interests include large language models, vision-based deep learning, affective computing and human-computer interaction wiith a focus on developing multimodal AI systems for education and speech understanding.



Myung-Ho Kim received the B.S. in Department of Computer Science and Engineering from Soongsil University, Korea, in 1989. M.S. and Ph.D. degrees in Department of Computer Engineering from

Postech University, Korea, in 1991 and 1995, respectively. He is currently a professor in the Dept. of Software, Soongsil University. He is interested in Machine Learning, Deep Learning and Block chain.