JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 101-112, December 2025
https://doi.org/10.9708/jksci.2025.30.12.101

An Open-Source AI-Based Multilingual Translation Lecture Script System for Global University Lectures

DongHyeon Shin*, Oh-Gyu Kwon**, ManKi Min*, Minseo Yoon*, DoHyun Oh*,
Jae-Woo Ryu***, Young Deok Park****

*Researcher, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea
**CEO, MOBIPINTECHNOLOGY Co., Ltd., Daegu, Korea
***Researcher, R&D Center, NEARNETWORKS Co., Ltd., Daegu, Korea
****Professor, Dept. of Computer Engineering, Yeungnam University, Gyeongsan, Korea

[Abstract]

We propose an open-source Artificial Intelligence (AI)-driven multilingual translation system that integrates real-time speech recognition and translation to resolve language barriers and misrecognition of technical terms in university lecture environments. The proposed system combines Whisper-based speech recognition, Bidirectional Encoder Representations from Transformers (BERT) and Levenshtein-distance based correction algorithm, and the DeltaLM multilingual translation model to automate the entire process from noise reduction and sentence refinement to context-aware translation and lecture script generation. In particular, a translation dataset containing domain-specific technical terms in the field of computer engineering is constructed and fine-tuned to improve translation precision in specialized domains. Experimental results show that the Word Error Rate (WER) decreases from 9.2 % to 4.7 %, achieving an improvement of approximately 51 %, while the average Bilingual Evaluation Understudy (BLEU) score increases from 56.3 to 60.2, corresponding to a 6.9 % performance gain. These results confirm that the proposed system achieves consistent translation quality improvements across all language pairs in academic lecture scenarios.

▸ **Key words**: Educational technology, Fine-tuning, Open-Source AI, Speech Recognition, Translation

• First Author: DongHyeon Shin, Corresponding Author: Young Deok Park
*DongHyeon Shin (dkvk486967@gmail.com), Dept. of Computer Engineering, Yeungnam University
**Oh-Gyu Kwon (okkwon1021@gmail.com), MOBIPINTECHNOLOGY Co., Ltd.
*ManKi Min (headwig9898@gmail.com), Dept. of Computer Engineering, Yeungnam University
*Minseo Yoon (nety03@yu.ac.kr), Dept. of Computer Engineering, Yeungnam University
*DoHyun Oh (odh2911@gmail.com), Dept. of Computer Engineering, Yeungnam University
***Jae-Woo Ryu (wodn10110@aaf.co.kr), R&D Center, NEARNETWORKS Co., Ltd.
****Young Deok Park (ydpark@yu.ac.kr), Dept. of Computer Engineering, Yeungnam University
• Received: 2025. 10. 30, Revised: 2025. 11. 29, Accepted: 2025. 12. 02.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 연구는 대학 강의 환경에서 발생하는 언어 장벽과 전문용어 오인식 문제를 해결하기 위해 실시간 음성인식과 다국어 번역 기능을 통합한 오픈소스 AI(Artificial Intelligence) 다국어 번역 시스템을 제안한다. 제안 시스템은 Whisper 기반 음성인식, BERT(Bidirectional Encoder Representations from Transformers) 및 레벤슈타인 거리 기반 교정 알고리즘, DeltaLM 다국어 번역 모델을 결합하여 강의 음성으로부터 잡음 제거, 문장 구조 정제, 문맥 기반 번역, 자막 출력까지의 전 과정을 자동화한다. 특히, 컴퓨터공학 분야의 전문용어를 포함한 번역 데이터셋을 구축하고 파인튜닝을 수행함으로써 전문 도메인에서의 번역 정밀도를 향상시켰다. 실험 결과, WER(Word Error Rate)이 9.2%에서 4.7%로 약 51% 개선되었으며, BLEU(Bilingual Evaluation Understudy Score) 평균 점수가 56.3에서 60.2로 약 6.9% 향상되어 모든 언어 쌍에서 일관된 번역 품질 개선을 확인하였다.

▸ **주제어**: 에듀테크, 파인튜닝, 오픈소스 AI, 음성인식, 번역기

I. Introduction

2019년에 발생한 코로나바이러스감염증-19 팬데믹으로 인해 전 세계 고등교육기관의 운영 체계와 학습 방식에 근본적인 변화가 나타났다. 전통적으로 대면 중심으로 이루어지던 교육은 비대면 및 온라인 기반으로 급격히 전환되었으며, 이는 단순한 위기 대응을 넘어 고등교육 전달 방식의 지속적인 혁신을 촉진하는 계기가 되었다. 국내 대학 역시 이러한 흐름에 발맞추어 LMS(Learning Management System), 실시간 스트리밍 플랫폼, 화상회의 도구 등을 도입하였으며, 상당수의 대학이 비대면 또는 하이브리드 강의를 상시 운영하는 체제로 전환하였다[1].

이와 같은 변화는 학습 접근성과 시간, 공간적 유연성 측면에서 긍정적인 성과를 가져왔으나, 모든 학습자에게 동등한 학습 경험을 보장하지는 못한다. 특히 국내 고등교육기관에 재학 중인 외국인 유학생의 경우, 강의 언어(한국어 또는 영어)에 대한 숙련도 차이로 인해 학습 내용 이해에 어려움을 겪는 사례가 지속적으로 보고되고 있다[2]. 한국교육개발원의 2024년 통계에 따르면 국내 고등교육기관 재학 외국인 유학생 수는 208,962명으로 역대 최고치를 기록하였으며, 이 중 다수가 중국, 베트남, 몽골 등 비영어권 출신이다[3]. 이러한 언어적 다양성은 교육적 형평성과 포용성 강화를 위한 새로운 과제로 작용하고 있다.

일부 대학에서는 영어 강의 개설 또는 이중 언어 자막 제공을 통해 유학생 학습 지원을 시도하고 있으나, 이러한 방식은 근본적인 한계를 지닌다[4, 5]. 영어는 비영어권 학습자에게도 제2외국어로 작용하므로 학습 이해도 향상에 한계가 있으며, 교수자에게는 언어적 부담이 가중되어 강의 품질이 저하될 수 있다. 또한 전문 통역 인력의 상시 투입은 인력 수급과 비용 측면에서 현실적으로 지속 가능하지 않다[6]. 이에 따라, 자동화된 방식으로 학습자의 언어 환경에 적응할 수 있는 실시간 다국어 번역 시스템의 필요성이 점차 부각되고 있다.

그러나 기존 실시간 번역 시스템은 고등교육 강의 환경에 직접 적용하기에는 여러 기술적 제약이 존재한다[7]. 일반적인 STT(Speech-to-Text) 기술은 잡음 환경, 복잡한 문법 구조, 전문용어 및 외래어 처리에서 여전히 높은 오류율을 보이고 있으며, 자동 번역 시스템 또한 문맥 단절, 낮은 번역 품질, 다국어 미지원, 실시간 지연 등 다양한 문제를 내포하고 있다[8]. 또한 STT, 번역, 자막 생성, 영상 동기화 등 각 모듈 간의 처리 연계성이 미흡하여, 이를 통합적으로 제공하는 실시간 일체형 구조는 아직 완성되지 않은 단계에 머물러 있다[9].

이러한 기술적 한계는 여러 측면에서 나타난다. 우선, 강의실 내에는 학생들의 대화, 냉난방기, 프로젝터 등으로 인한 비정형적 잡음이 빈번하게 발생하여 강의자의 음성 신호가 왜곡되기 쉽다. 이러한 환경적 요인은 STT 인식률을 저하시켜 자막 품질에 직접적인 영향을 미친다[10]. 또한, 대학 강의에서는 전문용어와 외래어가 빈번히 사용되며, 한국어의 복잡한 문법 구조와 발음 변이로 인해 STT의 인식 정확도가 저하된다. 그 결과 번역 문장이 왜곡되고 학습자의 이해도가 떨어지는 문제가 발생한다[11]. 뿐만 아니라 STT가 생성한 텍스트에는 문장 경계 불명확, 불필요한 공백 및 특수문자 등이 포함되어 번역 과정에서 문맥 단절을 유발하며, 자막의 자연스러운 흐름을 방해한다[12]. 또한 대부분의 실시간 자막 시스템은 한국어와 영

어 자막만 제공하고 있어, 중국어, 베트남어 등 주요 비영어권 유학생 집단을 충분히 지원하지 못한다[3]. 이러한 제약으로 인해 다수의 유학생이 강의 핵심 내용이나 전문용어의 뉘앙스를 정확히 이해하지 못하고, 결과적으로 학업 성취도에서 불균형이 발생한다. 마지막으로 음성인식, 번역, 자막 생성 및 배포 과정이 독립적으로 수행되면서 시스템 간 지연과 호환성 문제가 발생하고, 학습자가 별도의 소프트웨어를 설치하거나 복잡한 접근 절차를 거쳐야 하는 비효율성이 존재한다.

본 연구는 이러한 문제를 해결하기 위해 고등교육 강의 환경에 최적화된 실시간 다국어 번역 시스템을 제안한다. 제안 시스템은 강의자의 음성을 실시간으로 인식하여 텍스트로 변환한 뒤, 문맥 기반 다국어 번역을 수행하고 이를 자막 형태로 즉시 제공한다. 특히 음성인식 단계에서는 소음 환경에서의 인식 안정성을 강화하고, 전문용어 처리 및 문장 단위 세분화 알고리즘을 적용함으로써 번역 품질과 문맥 일관성을 향상시킨다. 또한 중국어, 베트남어 등 주요 언어를 포함한 동시 다국어 자막 출력 기능과, 음성인식, 번역, 자막 및 배포 과정을 단일 플랫폼 내에서 처리하는 End-to-End 시스템 아키텍처를 구현하여 실시간성과 사용자 편의성을 동시에 확보한다. 이를 통해 제안 시스템은 고등교육 현장에서 학습자의 언어 장벽을 완화하고, 학습 몰입도와 이해도를 향상시킬 뿐만 아니라 교육의 형평성과 포용성을 증진할 수 있을 것으로 기대한다.

본 논문의 기여는 다음과 같다.

- 본 시스템은 음성 신호에 노이즈 필터링과 고음부 강조 기능을 적용하여, 잡음이 많은 환경에서도 안정적인 자막을 생성할 수 있다.
- 고등교육 환경에서 자주 사용되는 분야별 전문용어를 학습함으로써, 기존 번역 시스템이 처리하기 어려웠던 전문용어 번역의 한계를 개선하고, 자연스럽고 정확한 다국어 번역을 가능하게 하였다.
- 다국어 번역 기능을 지원함으로써 비영어권 유학생들도 언어적 제약 없이 강의 내용을 실시간으로 이해할 수 있는 학습 환경을 제공한다.

이러한 기술적 기여는 다음과 같은 교육적 기대 효과로 이어질 수 있다. 먼저, 강의 음성의 안정적 인식과 자동 교정을 통해 명확한 자막이 포함된 강의를 효율적으로 제공할 수 있으며, 학습자는 실시간으로 정확하고 자연스러운 자막을 통해 학습 내용을 보다 명확하게 이해할 수 있다. 또한, 전문용어 인식과 다국어 번역 기능을 통해 비영어권 유학생을 포함한 다양한 학습자가 언어적 한계에 제약 없이 동일한 학습을 할 수 있어 교육의 형평성과 언어적 포용성이 강화된다. 마지막으로, 제안된 시스템은 고등교육 환경에서 강의자의 음성 발화부터 번역과 자막 생성까지의 전 과정을 하나의 End-to-End 방식으로 처리함으로써, 교육의 질과 효율성을 높이고 다문화, 다언어 학습자에게 포용적인 학습 환경을 제공한다.

II. Related Work

2.1. MMSE for Noise Reduction

마이크를 통해 입력된 음성 신호에서 원하는 신호를 효과적으로 추출하기 위해 불필요한 잡음을 제거하는 과정이 필요하다. 노이즈를 제거하기 위한 대표적인 방법 중 하나는 MMSE(Minimum Mean Square Error) 추정 기반의 로그 스펙트럼 필터링이다[13]. 이 기법은 시점별 주파수 성분을 파악할 수 있고 사람의 청각을 고려하는 과정으로 음질을 향상시키는데 효과적이다. 이 기법은 음성 신호를 짧은 시간 단위로 프레임 분할한 후, 각 프레임에 대해 STFT(Short-Time Fourier Transform)을 적용하여 시간 도메인 신호를 주파수 도메인으로 변환함으로써 FFT(Fast Fourier Transform)보다 특정 시점의 주파수 변화를 보다 정밀하게 분석할 수 있으며, 이후 음성인식 과정에서 유용하게 활용될 수 있다[14]. 이후 노이즈 스펙트럼을 추정하고 주파수 성분에서 로그 스펙트럼을 계산하여 사람의 청각 스케일에 맞춘다. 이를 통해 얻은 정보로 MMSE 필터를 적용하여 백색 잡음이나 주변 환경 소음에서 높은 SNR(Signal-to-Noise Ratio) 개선 효과를 보인다. MMSE를 이용한 전처리를 이용한 음성 신호의 잡음 제거 성능을 높이기 위한 연구에서 MMSE 추정 기반의 로그 스펙트럼 필터링 기법을 사용해서 확연한 SNR 성능 향상을 보였다[15].

2.2. Pre-emphasis for High-Frequency Enhancement

음성 신호의 주파수 특성은 일반적으로 저주파 성분이 강하고 고주파 성분이 약한 경향을 보인다. 그러나 사람의 음소 구분과 자음 인식에는 고주파 대역이 중요한 역할을 한다. 따라서 고주파 성능을 강조하여 인식 성능을 개선할 필요가 있으며, 주로 저주파와 고주파의 성분 차이를 줄이는 pre-emphasis 필터를 사용한다[16].

pre-emphasis를 적용하는 과정으로는 입력된 음성 신호에 식 1과 같이 1차 차분 필터를 사용하여 고주파 성분을 강조한다.

$$y[n] = x[n] - \alpha \cdot x[n-1] \quad (1)$$

이러한 전처리를 통해 고주파에서 간섭이 강한 환경에서도 잡음의 영향을 줄일 수 있으며 음향 특징 추출 단계에서 유의미한 효과가 있다. Iván López-Espejo의 연구에서는 음성 신호 처리에 있어 pre-emphasis 필터링이 단순하면서도 계산 비용이 낮은 방식으로 음성 향상 성능을 개선할 수 있음을 제안하였다[16]. Ertan Loweimi의 연구에서는 phase 기반 음성인식 특징에 고주파증폭과 윈도잉을 적용했을 때 Aurora2 잡음 환경에서 WER이 최대 15%까지 감소한 결과를 보였다[17]. 이를 통해 간단한 전처리만으로도 음성인식성능이 효과적으로 개선됨을 알 수 있다.

2.3. Speech Recognition

본 연구에서는 STT 모델 중 하나인 Whisper를 사용하며, Whisper를 이용한 대학 강의 영상 자막 자동 생성에 대한 연구 사례가 존재한다[18, 19]. 이 연구에선 높은 WER 문제를 후처리로 보완하였지만, 강의에 특화된 전문용어를 처리하는 기술은 여전히 부족하다. 이를 통해 강의 분야에서도 Whisper가 높은 인식 성능을 보이지만, 전문용어에 대해서는 추가적인 보정이 필요함을 확인할 수 있다.

2.4. LLM based Post-processing

최근 STT 모델은 높은 성능을 보이지만, 여전히 고유명사, 외래어, 도메인 특화 단어, 긴 문장 구조 등에서 오류가 발생하는 경우가 많다[20]. 특히 Whisper와 같은 모델은 전사 정확도는 높지만 높은 WER을 보이며, 실사용을 위해서는 후처리 과정이 필요하다[21]. 이러한 한계를 보완하기 위해, LLM(Large Language Model)을 활용한 후처리 방법이 연구되고 있다. LLM 기반의 후처리는 STT 결과를 문맥적으로 재해석하여 잘못 인식된 단어를 보정하거나 문장을 더 자연스럽게 구성하도록 도와준다.

BERT는 Google이 2018년에 제안한 자연어 처리 모델로, 트랜스포머 아키텍처 기반의 인코더 구조를 사용하며, 문장의 앞뒤 맥락을 동시에 고려하여 단어의 의미를 학습하는 양방향 학습 구조를 채택하고 있다[22]. 이를 통해 단어 간 의미 관계 및 문장 내 문법적 정합성을 보다 정밀하게 파악할 수 있다.

최근에는 한국어의 형태적, 음운적 특성을 반영할 수 있는 사전학습 언어모델이 활발하게 개발되고 있다. KoBERT를 시작으로 다양한 한국어 특화 BERT 계열 모델이 등장하였으며, 예를 들어 KRongBERT는 형태소 기반 임베딩 방식을 채택하여 문맥 표현력을 크게 향상시킨 바 있다[23]. 또한 KoGEC는 한국어 번역 기반 사전학습 모델을 활용하여 조사, 어미, 띄어쓰기 등 한국어 고유 문법을 정교하게 반영한 자동 문법 교정 프레임워크를 제안하였다[24]. 이와 함께 KcBERT 기반 연구에서는 MLM 기반 예측 기능을 이용해 문맥적으로 자연스러운 교정어를 생성함으로써 언어 정제 성능을 개선하였다[25].

이러한 선행연구들은 STT(음성인식) 결과의 후처리 단계에서 문맥 기반 교정 정확도를 높이기 위한 유용한 한국어 맞춤 접근법으로, 문맥 평가, 교정 기법의 적절성을 검증했다. 다만, 기존 BERT 기반 교정 방식은 문맥 예측에 의존하여 실제 발화되지 않은 단어가 새롭게 삽입되는 문제가 발생할 수 있어, 음성 기반 교정에 그대로 적용하기에는 한계가 존재한다.

본 연구에서는 BERT를 통해 앞뒤 맥락을 고려하여 STT에서 출력한 단어를 맥락에 맞게 수정한 뒤, 번역을 위해 DeltaLM의 입력값으로 사용한다[26]. DeltaLM은 트랜스포머 구조를 기반으로 하여, 다양한 언어 간 문맥 정보를 효과적으로 반영할 수 있는 번역 모델이다. 이 모델은 입력된 문장을 특정 언어로 번역하며, 문장의 의미 보존과 문맥 흐름 유지에 적합한 구조로 설계되어 있다. DeltaLM에 추가적으로 전문용어들을 학습시켜 기존 연구들에서 취약점인 전문용어에 대한 인식 및 번역 성능을 보완하여 학습자가 대학 강의를 효과적으로 학습할 수 있도록 돕는다.

III. The Proposed Scheme

본 연구에서는 앞서 제기된 문제들을 해결하기 위해, 고등교육 강의 환경에 최적화된 실시간 다국어 번역 자막 시스템을 제안한다. 제안된 시스템은 강의자의 음성을 실시간으로 수신하고, 이를 고정밀 음성인식, 문맥 기반 다국어 번역, 자막 후처리, 영상 동기화 및 자막이 포함된 강의 스트리밍까지 하나의 파이프라인으로 통합하여 제공하는 형태로 구성된다. 각 구성 모듈은 고등교육 환경의 특수성을 반영하여, 잡음 환경에 강인한 음성인식, 전문용어 및 외래어 대응, 다국어 번역 정확도 향상, 실시간 자막 동기화 등의 기능을 수행한다.

그림 1은 제안하는 시스템의 전체 구조를 나타낸다. 강의자의 음성은 마이크를 통해 입력되며, 노이즈 제거 및 음성 강조를 위한 전처리 과정을 거친다. 정제된 음성은 Whisper 모델을 통해 텍스트로 변환되며, 오인식 단어를

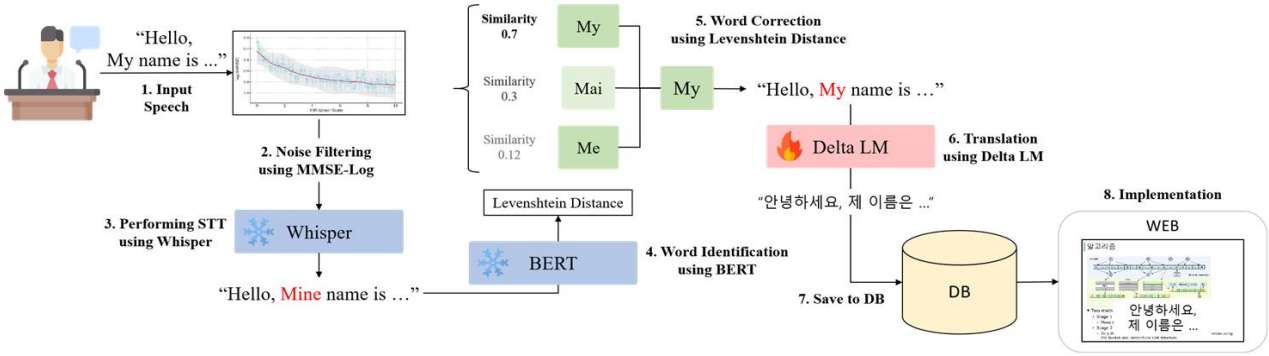

Fig. 1. Overall Architecture of The Proposed Terminology-Aware Multilingual Lecture Script Generation System

BERT를 통해 추출한 후 레벤슈타인 거리(Levenshtein Distance) 를 활용하여 철자 및 문맥 기반 오류 보정이 수행된다[27]. 보정된 텍스트는 DeltaLM 기반 번역 모델을 통해 주요 언어로 실시간 번역되고, 최종적으로 자막과 강의 영상이 통합되어 웹 서비스에 실시간으로 송출된다.

3.1. Pre-processing Speech Data for Efficient Speech Recognition

3.1.1. Noise Filtering

제안 시스템에서 음성의 입력은 강의자의 마이크를 통해 입력된다. 입력된 음성은 STT를 통해 자막으로 송출하기 위해 사용되지만, 잡음이 강한 환경에서 이 음성 신호는 활용하기 어렵다. 따라서 본 절에서는 STT를 효율적으로 수행하기 위한 음성신호 정제 과정을 설명한다.

먼저 수신된 음성 신호는 노이즈를 제거하고 강의자 음성을 강조하는 전처리 과정을 거친다. 정제된 음성은 Whisper를 통해 텍스트로 변환되며, 번역 정확도 향상을 위해 문맥 기반의 후처리 기법이 적용된다. 최종적으로, 이러한 텍스트는 다국어 번역 모델에 입력되어 다양한 언어로 변환된다.

수신한 음성에서 노이즈를 필터링 하기 위해, 사람이 소리를 인식하는 특성을 반영할 수 있는 MMSE 로그 스펙트럼 추정법을 이용한 노이즈 필터를 사용한다. 이 기법은 입력 신호를 짧은 프레임 단위로 분할 한 후, 각 프레임에 일정 윈도우 구간으로 나눠 푸리에 변환의 입력값으로 전환한다. 각 프레임을 윈도우 크기에 따라 나누고 Overlap을 적용한 후, STFT를 적용하여 시간 도메인에서 주파수 도메인으로 전환하여 시점별로 주파수 성분을 파악할 수 있도록 한다. 또한 음성이 시작되기 전 몇 프레임을 이용하여 노이즈의 파워 스펙트럼을 추정하고, STFT를 통해 얻은 주파수 성분에 로그를 취하여 로그 스펙트럼을 계산해 필터 설계에 이용한다. 이를 바탕으로 MMSE 로그 스펙트럼 추정기를 설계하여 적용하고 Inverse STFT를 통해 주파수 도메인 신호를 시간 도메인으로 복원한다. 이후 Overlap-Add 기법을 통해서 시간 신호를 부드럽게 재구성한다.

3.1.2 Noise Reduction and Speech Enhancement

노이즈가 제거된 음성 신호에서 인식 성능을 개선하기 위해, 노이즈 제거 단계에서 MMSE 로그 스펙트럼 추정 기법을 적용하고, 음성 강조 단계에서는 pre-emphasis를 적용한다. 인간의 발성 기관과 소리의 특성에 의해 낮은 주파수 대역에 비해 높은 주파수 대역의 진폭이 낮은 것을 강조함으로써 주파수 스펙트럼 간의 균형을 맞춘다. 스펙트럼 기반 특징을 보다 정확하게 추출하고, 음성인식 모델이 신호를 안정적으로 처리할 수 있다.

3.2. Speech to Text

전처리된 음성신호를 통해 정확하고 문맥에 적합한 자막을 생성하기 위해서는, 해당 음성을 실시간으로 문자 데이터로 변환하는 과정이 필요하다. 제안 시스템은 Whisper 기반의 사전학습 STT 모델을 기반으로 하되, 고등교육 강의 환경에 특화된 조건을 반영하기 위해 추가적인 학습 및 전처리 보정 단계를 도입한다.

3.2.1 Zero-Shot Inference Using Whisper

Whisper는 다국어 음성인식과 번역 기능을 지원하는 트랜스포머 기반의 사전학습 STT 모델로, 다양한 발화 환경에서도 견고한 인식 성능을 보인다. 그러나 해당 모델은 주로 대화체 데이터를 기반으로 학습되었기 때문에, 고등교육 강의 환경과 같이 학문 분야별 전문용어가 자주 등장하거나, 외래어의 음차 표현, 문어체에 가까운 발화 스타일, 그리고 고정 마이크가 아닌 환경 잡음이 포함된 녹음과 같은 특수 조건에서는 인식 오류가 빈번하게 발생한다.

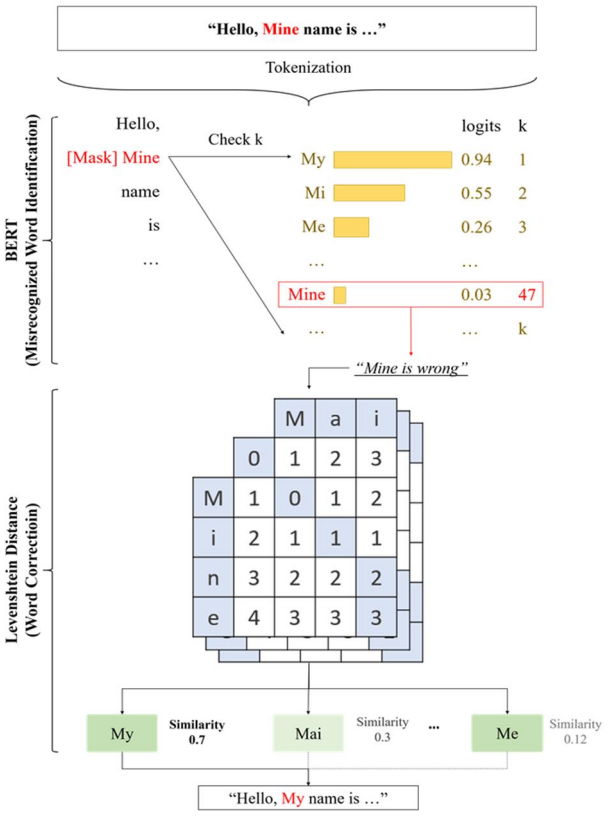

Fig. 2. Word Identification and Correction Using BERT and Levenshtein Distance

따라서 제안 시스템은 사전학습된 Whisper 모델을 Zero-Shot 방식으로 적용한다. Whisper는 대규모 다국어 데이터 및 다양한 도메인 음성 데이터로 사전학습되어 있어, 학습하지 않은 도메인에서도 고수준의 음성인식 및 번역 성능을 유지한다. 이 모델을 이용하여 STT를 수행하기 위해, 제안 시스템은 실제 대학 강의 음성(약 12시간)을 데이터로 사용하였으며, 모델의 파라미터를 수정하지 않고 프롬프트 및 입력 전처리 조정을 통해 강의 도메인에 최적화된 결과를 얻었다. 특히 다양한 말투, 억양, 환경 잡음을 포함한 실제 강의 음성을 활용함으로써 Whisper의 잡음 내성과 문맥 일반화 성능을 검증하였다.

3.2.2 Checking Spell and Correcting Words

STT 모델을 통해 출력된 초기 텍스트는 실시간성 및 인식 한계로 인해 오탈자, 철자 오류 또는 의미 왜곡이 포함될 가능성이 높다. 특히 고등교육 강의의 경우, 자막 정확도는 학습자의 개념 이해 및 몰입도에 직접적인 영향을 미치므로, 자막 품질 보정을 위한 후처리 단계가 필수적으로 요구된다.

제안 시스템은 철자 오류 및 문맥적 부정확성을 보정하기 위해 두 단계로 구성된 텍스트 정제 절차를 적용한다. 해당 절차의 전체 흐름과 적용 예시는 그림 2와 같다. 먼저 오인식 단어를 찾는 과정은 다음과 같다. 그림 2와 같이 BERT의 마스킹 언어 모델(Masked Language Model, MLM) 방식을 활용하여 [MASK]로 처리한 문장에서 [MASK] 단어가 오인식 단어인지 아닌지를 평가한다. [MASK] 단어가 Top-30의 후보 단어인 경우, BERT는 Softmax 기반 확률 분포를 출력하며, 이 중 가장 높은 확률을 보이는 단어가 최종적으로 선택된다. 하지만 [MASK] 단어가 Top-30의 후보 단어로 오지 않는다면 이는 오인식단어로 판단한다.

예를 들어, "Hello, Mine name is …" 라는 STT 결과에 대해, 오인식 단어를 추론하는 단계에서 "Hello, [MASK] name is …"로 구성한 후, 각 마스킹 위치에 원래 단어인 "Mine"이 Top-30의 후보 단어로 오지 않는다면 이는 오인식 단어로 판단한다. 이러한 문맥기반 오류검출 과정을 통해 조사 누락, 어순 오류, 자연스럽지 않은 어휘 선택 등 문법적으로 다양한 오인식 단어를 추론할 수 있다.

오인식 단어를 추론한 후, 검출된 오인식 단어를 교정하기 위해 편집 거리 기반의 교정 기법을 적용한다. 편집 거리는 두 문자열 간의 최소 편집 횟수를 계산하는 방식으로, 본 연구에서는 레벤슈타인 거리를 편집 거리로 사용한다. 레벤슈타인 거리는 삽입, 삭제, 교체 연산을 통해 한 단어를 다른 단어로 변환하는 데 필요한 최소 연산 횟수를 정의하며, 오인식 단어와 Top-30의 후보 단어들 간의 거리를 계산하여 유사 단어 후보군을 생성한다. 이후 미리 정의된 임계값 이하의 단어들을 선별하여 교정 후보로 구성한다. 예를 들어, 입력이 "Mine"인 경우 레벤슈타인 거리 계산을 통해 "Mai", "My", "Me" 등 철자 유사도가 높은 단어들이 후보로 선택된다. 이후, 유사도가 가장 높은 "My"가 교정 단어로 선택된다.

이러한 철자 기반 교정 단계는 앞서 수행된 문맥 기반 탐지 결과를 보완하여, 최종적으로 의미적 일관성과 정확도가 높은 단어를 확정하는 역할을 수행한다. 이 두 단계로 구성된 텍스트 정제 절차는 특히 유사 발음 또는 부분 음소 손실로 인한 철자 오류를 효과적으로 완화하며, 문맥 기반 오인식 단어를 추론하는 단계와의 연계를 통해 인식 결과의 품질을 향상시킬 수 있다.

3.3. Multilingual Translation System

본 절에서는 전문용어를 포함하며 한국어, 영어, 중국어, 일본어, 베트남어 처리가 가능한 다국어 번역 시스템 구축 과정을 설명한다. 제안 시스템은 문장 전처리, 다국어 번역, 자막 표출의 3단계 파이프라인으로 구성되며,

Table 1. Examples of Multilingual Training Dataset

Korean	English	Chinese	Japanese	Vietnamese
스프링(Spring) 프레임워크는 자바 기반...	The Spring Fr...	Spring框架广...	Springフレーム...	Spring Framewor...
Apache Spark는 대규모 데이터 분석과...	Apache Spark...	Apache Spa...	Apache Spa...	Apache Spark...
문자열(String)은 문자의 시퀀스로, 파이...	A string is a s...	字符串是字符串...	文字列は文字...	Chuỗi là một dã...
세그먼트(Segment) 메모리는 프로그램의...	Segent memory...	段式内存分为...	セグメントメモ...	Bộ nhớ phân đo...
스레드(Thread)는 하나의 프로세스 내에...	A thread is the...	线程是进程中...	スレッドは、プ...	Luồng là đơn vị...
네트워크 노드(Node)는 데이터를 송수신...	A network nod...	网络节点是指...	ネットワーク...	Nút mạng là...
클러스터(Cluster)는 여러 서버를 하나의...	A cluster is a t...	集群是一种使多...	クラスタは...	Cụm máy chủ l...
파이프라인(Pipeline)은 데이터의 흐름...	A pipeline is a...	流水线是按阶...	パイプライン...	Pipeline là cấu trú...
게이트웨이(Gateway)는 내부 네트워크와...	A gateway is a...	网关是连接内部...	ゲートウェイは..	Gateway là thiết...
셸(Shell)은 사용자가 운영체제와 상호...	The shell is an...	Shell是允许用户...	シェルは、ユー...	Shell là giao diệ...
포크(Fork)명령은 기존 프로세스를 복...	The fork comm...	fork命令复制现...	forkコマンドは...	Lệnh fork sao c...

DeltaLM 기반 번역 모델을 하이퍼파라미터 최적화와 파인튜닝을 통해 학습한다.

표 1은 다국어 번역 시스템 구성을 위해 컴퓨터공학 분야의 전문용어를 포함한 다국어 병렬 데이터를 수집 및 정제하여 모델 학습용 데이터셋이다. 데이터셋은 한국어, 영어, 중국어, 일본어, 베트남어 등 5개 주요 언어로 구성되며, 각 언어 쌍은 문장 단위로 정렬되어 번역 모델의 학습 효율을 극대화한다. 데이터 수집은 대학의 컴퓨터공학 전공 교재, 강의자료, 오픈소스 기술 문서, 개발자 커뮤니티, 기술 블로그 등에서 병렬 문장을 확보하는 방식으로 수행되었다.

이후 수집된 데이터는 각 언어별 형태소 분석 및 구문 구조 검사를 통해 비문법 문장, 불완전 번역, 중복 문장을 자동 검출 후 제거하였으며, 언어별 길이 비율 기반 필터링을 적용하여 정렬 품질을 보정하였다. 이를 통해 약 10만 건 이상의 고품질 병렬 번역 문장 쌍을 확보하였다.

특히 컴퓨터공학 도메인에서 일반 단어와 중의성을 가지는 전문용어를 집중적으로 포함시켰다. 예를 들어 “스프링(Spring)”, “포크(Fork)”, “스트링(String)” 등의 단어는 일상 언어에서도 사용되지만, 소프트웨어 개발 문맥에서는 각각 웹 프레임워크, 프로세스 복제, 문자열 자료형을 의미한다. 이러한 도메인 특화 어휘를 포함함으로써 모델이 일반 언어 문맥과 기술 문맥을 구분하여 번역할 수 있도록 학습 기반을 강화하였다. 구축된 데이터셋으로 사전학습된 DeltaLM 모델을 도메인에 특화되도록 파인튜닝하였으며, 주요 하이퍼파라미터 설정은 표 2에 제시하였다.

학습 안정성과 메모리 효율, 도메인 적합성을 고려하여 Batch Size 128과 Tokens per Batch 1,024를 적용하였으며, 과적합을 방지하고 문맥 일반화 성능을 높이기 위해 Label Smoothing을 적용한 Cross-entropy Loss와 Adam($\beta_1 = 0.9, \beta_2 = 0.98$) Optimizer를 사용하였다. 또한 DeltaLM의 사전학습 지식을 유지하면서 도메인 특화 데이터에 부드럽게 적응하도록 Learning Rate를 1e-4로 설정하였고, 강의 대본 특성상 문장 길이가 일정하지 않다는 점을 고려하여 Max Sequence Length와 Max Target Length는 512로 지정하여 충분한 문맥 처리 공간을 확보하였다.

제안된 다국어 번역 시스템은 입력 문장에서 자막 생성까지의 과정을 하나의 통합 파이프라인 형태로 설계하였으며, 문장 전처리-다국어 번역-자막 표출의 세 단계로 진행된다. 먼저, 문장구조 전처리 단계에서는 음성인식 결과를 기반으로 문법적 구조를 보정하여 번역 모델 입력에 최적화된 형태로 변환한다.

이 과정에서 구두점과 접속어를 기준으로 문장을 분리하여, 긴 복합문을 단문 단위로 세분화함으로써 번역의 안정성과 일관성을 높인다. 또한, 음성인식 과정에서 발생할 수 있는 특수문자, 중복 공백, 이모지, HTML 태그, 비완성 괄호 등 불필요한 요소를 제거한다. 이러한 전처리 과정을 통해 입력 문장은 언어적 일관성을 확보한 표준 번역 포맷으로 변환된다.

다국어 번역 단계에서는 전처리된 문장이 DeltaLM 기반의 다국어 번역 엔진을 통해 5가지 언어로 실시간 번역된다. DeltaLM 모델은 문맥을 고려하여 각 단어 간의 의미적 관계를 해석하고, 전문용어 사전과 연동되어 도메인 표준 번역어를 자동으로 적용한다.

Table 2. Hyperparameter Settings for DeltaLM Fine Tuning

Hyperparameter	Value
Batch Size	128
Token per Batch	1,024
Loss Function	label_smoothed_cross_entropy
Optimizer	adam($\beta_1 = 0.9, \beta_2 = 0.98$)
Learning Rate	1e-4
Learning Rate Scheduler	inverse_sqrt
Max Sequence Length	512
Max Target Length	512

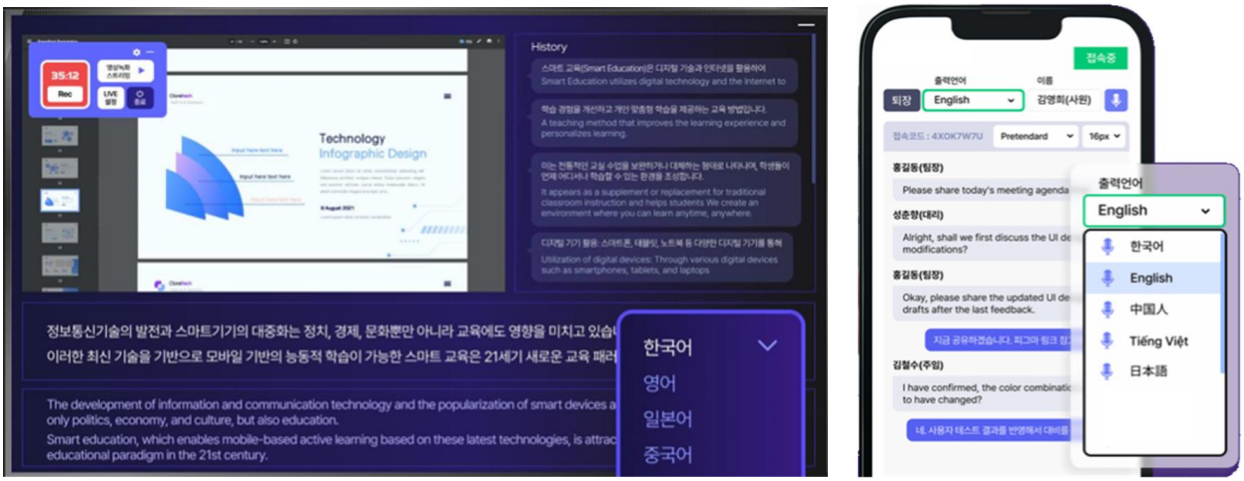

Fig. 3. Implemented Real-Time Multilingual Translation System (Windows and Mobile apps)

마지막으로, 자막 표출 단계에서는 번역 결과가 실시간 자막 형태로 시각화된다. 그림 3과 같이 자막은 강의 화면 하단에 시각적으로 안정된 형태로 표시되며, 언어별로 색상, 위치, 출력 타이밍을 차별화하여 학습자의 이해를 돕는다. 이와 같은 구조를 통해 제안된 시스템은 실시간 번역과 자막 생성을 동시에 수행하면서, 강의 콘텐츠의 언어 접근성과 학습 효율을 크게 향상시킬 수 있다.

IV. Experiments

본 장에서는 제안된 시스템의 성능을 정량적으로 검증하기 위해, Whisper만을 사용한 Baseline, Whisper 출력에 BERT 기반 교정 모듈을 추가한 최신 연구 기법인 KcBERT 기반 모델, 그리고 제안된 음성 교정 알고리즘을 적용한 제안 기법을 대상으로 실험을 수행하였다. 각 모델에 대한 정의는 표 3에 요약하였다. 또한 실험에 사용한 컴퓨팅 장치의 사양을 표 4에 요약하였다.

Table 3. Model Definition for WER Comparison

Type	Method
Baseline	Whisper(Large-V3)
KcBERT-based	Whisper + KcBERT(Top-30)
Proposed	Whisper + MMSE + Pre-emphasis + BERT(Top-30) & LD Correction

Table 4. Device Specification for Experiment

Device	Specification
CPU	Intel(R) Core(TM) i7-10700, 2.90GHz
Memory	16GB
GPU	NVIDIA RTX 3060

첫 번째 실험은 세 모델에 대해 동일한 음성 입력을 사용하여 WER을 비교함으로써, 교정 알고리즘이 오인식 단어를 얼마나 효과적으로 탐지하고 수정하는지 평가한다. 또한, 실시간 처리에 요구되는 성능을 확인하기 위해 문장 단위 평균 처리시간을 측정하여 모델별 연산 비용을 함께 분석하였다. 두 번째 실험으로 제안 모델이 문장 내 의미적 일관성과 번역 품질을 향상시키는지를 검증하기 위해, 세 모델에서 생성된 번역문을 대상으로 BLEU 점수 기반 번역 품질 비교 평가를 수행하였다.

4.1. WER and Processing Time Evaluation

제안된 음성 정제 및 교정 알고리즘의 효과를 검증하기 위해 Whisper만을 사용한 Baseline, Whisper 출력에 BERT 기반 교정 모듈을 적용한 KcBERT 기반 모델, 그리고 제안 알고리즘을 통합한 제안 기법의 WER 성능과 처리속도를 비교하였으며 그 결과를 표 5에 요약하였다. 비교분석을 위해 이를 위해 캐글(Kaggle)의 한국어 데이터셋(Korean Single Speaker Speech)로부터, 약 7~9개의 어절로 구성된 2~3초 분량의 한국어 문장 1,000건을 임의로 선택하여 평가용 대본을 준비하였다[28]. 그리고 대본의 내용을 음성인식 모델에 음성을 저자들이 직접 음성으로 녹음하여 각 모델의 입력값으로 사용하였다.

Table 5. Comparison of WER and Processing Time

Type	WER	Process Time
Baseline	9.2% (0.092)	0.24 sec
KcBERT-based	6.2% (0.062)	0.61 sec
Proposed	4.7% (0.0047)	0.70 sec

실험결과, Whisper 단독 사용 시 WER은 9.2%였으며, BERT 기반 정제 과정을 포함한 KcBERT 기반 모델은 오인식 단어의 문맥적 교정으로 WER이 6.2%로 감소하였다. 또한 MMSE 기반 잡음 제거와 pre-emphasis 고주파 강조, BERT 및 LD 기반 단어 단위 교정의 최적화를 포함한 제안 모델(제안 기법)을 적용한 경우 WER은 4.7%로 가장 낮게 측정되었으며, 이는 Baseline 대비 약 49%의 WER 감소를 의미한다. 이러한 결과는 제안한 정제, 교정 알고리즘이 발음 혼동 및 단어 오인식을 효과적으로 줄여, 문장 내 단어 인식 정확도를 크게 향상시켰음을 보여준다.

한편, 모델별 처리속도 비교 결과, Baseline은 평균 0.24초로 가장 빠른 응답 시간을 보였고 KcBERT 기반 모델은 평균 0.61초로 측정되었다. KcBERT 기반 모델은 문장 단위 확률 분석 및 교정 과정이 추가되면서 처리시간이 약 2.5배 증가한 것으로 판단된다. 제안 기법은 모델들 중 가장 연산량이 많아 평균 0.70초로 가장 높은 처리시간을 기록하였으나, 그럼에도 불구하고 제안 기법은 WER을 크게 개선하였고 처리시간이 실시간 스트리밍 환경에서도 허용 가능한 수준의 지연이라는 점에서 실용성을 갖는 결과로 판단된다.

4.2. BLEU Evaluation for Terminology-Aware Multilingual Translation

본 절에서는 제안된 전문용어 기반 파인튜닝이 다국어 번역 품질에 미치는 영향을 평가하기 위해, 파인튜닝이 적용되지 않은 기본 DeltaLM 모델과, 전문용어 데이터셋으로 파인튜닝된 제안 모델의 번역 품질을 비교하였다. 표 6에 번역에 사용될 모델을 나타내었다. 비교 실험을 위해 컴퓨터공학 핵심 교과목(알고리즘, 자료구조, 운영체제, 컴퓨터구조, 데이터베이스)에서 자주 사용되는 전문 용어가 포함된 강의 대본 데이터셋을 구축하였다. 이 데이터셋은 각 과목 담당 교수로부터 사용 허가를 받은 후, 녹화된 강의 영상으로부터 대본을 발췌, 정제하여 약 3초 분량의 문장으로 총 1,000건으로 구성하였다.

Table 6. Model Definition for BLEU Evaluation

Type	Training Method
Baseline	Pre-trained DeltaLM Model
Proposed	DeltaLM Fine-tuned with Domain Terminology Dataset

실험은 5개 주요 언어(한국어, 영어, 중국어, 일본어, 베트남어)에 대해 입력-출력 언어 쌍을 구성하고, 총 1,000개의 문장 쌍의 검증 데이터셋을 대상으로 수행하였다. 두 모델 모두 동일한 학습 파라미터를 사용하였으며, 제안 모델은 전문용어 사전 기반 데이터로 파인튜닝된 모델이다.

표 7은 Baseline과 제안 모델의 다국어 번역 성능을 5개 주요 언어를 입력으로 하여 서로 다른 4개 언어로 번역한 결과를 나타낸다. 모든 언어 쌍에서 제안 모델이 Baseline 대비 BLEU 점수 향상을 보였으며, 평균적으로 약 3.9의 성능 향상이 확인되었고 이는 6.9%의 수준의 성능 향상이다.

이 결과는 제안 모델이 특정 언어나 데이터셋에 의존하지 않고, 언어 간 문장 구조 차이, 어순 변화, 표현 다양성 등 일반적 번역 변동 요인에 대해 보다 강한 일반화 성능을 나타낸다. 즉, 제안된 학습 방식은 전문 도메인 용어뿐 아니라 일반 문장에서도 문맥 일관성을 향상시켰다. 또한, BLEU 점수 향상은 모든 입력 언어에 대해 고르게 나타나 특정 언어 쌍에 편향되지 않았으며, 이는 제안된 모델이 다국어 번역 품질의 균형적인 개선을 달성했음을 보여준다. 결과적으로 제안 모델은 기존 사전학습 모델이 가진 도메인 및 언어 편향을 완화하고, 다양한 언어 간의 의미적 일관성과 표현 정확도를 향상시켰다.

V. Conclusions

본 연구에서는 고등교육 강의 환경에서 외국인 학습자의 언어 장벽을 완화하고, 전문용어가 포함된 강의 내용을 다국어로 정확히 전달하기 위한 실시간 다국어 번역 자막

Table 7. BLEU Performance Evaluation and Comparison

Source \ Target	Korean	English	Chinese	Japanese	Vietnamese
Korean	-	58.4 / **65.8**	55.1 / **58.9**	55.7 / **59.7**	54.2 / **56.2**
English	57.2 / **61.0**	-	56.1 / **59.1**	56.4 / **60.4**	55.2 / **57.8**
Chinese	57.8 / **62.4**	58.1 / **63.2**	-	55.5 / **58.7**	54.9 / **56.9**
Japanese	56.5 / **60.3**	57.9 / **62.1**	55.0 / **58.3**	-	53.8 / **55.7**
Vietnamese	55.3 / **58.6**	57.2 / **61.4**	54.6 / **56.8**	53.5 / **55.9**	-

Bold font: our proposed system

시스템을 제안하였다. 제안 시스템은 MMSE 기반 잡음 제거와 pre-emphasis, BERT 및 레벤슈타인 거리 기반 문맥 교정, 전문용어 병렬 데이터로 파인튜닝된 DeltaLM 번역 모델을 통합하여, 음성인식부터 번역 및 자막 출력까지 하나의 End-to-End 구조로 구성하였다.

실험 결과, 1,000건의 문장에 대한 음성인식 정확도 평가에서 제안 시스템의 WER은 단순히 Whisper 기법을 이용한 성능인 9.2%에서 4.7%로 감소시켰다. 또한 컴퓨터공학 교과목의 전문용어가 포함된 1,000건의 문장번역 성능 평가에서 5개 언어(한국어, 영어, 중국어, 일본어, 베트남어)에 대한 BLEU 점수는 평균 3.9 향상되었다. 이는 제안 모델이 잡음 환경, 전문용어 인식, 문맥 일관성, 다국어 번역 품질 등에서 기존 시스템의 한계를 개선했음을 입증한다. 결과적으로 본 연구는 도입부에서 제시한 주요 문제인 잡음 환경에서의 인식 취약성, 전문용어 인식 오류, 문장 경계 불명확, 다국어 자막의 부재, 그리고 모듈 간 통합성 부족을 실질적으로 해결하였다.

다만, 제안 기법은 제한된 소음 조건과 컴퓨터공학 중심 교과목을 대상으로 평가가 이루어졌기 때문에, 다양한 소음 유형과 더 폭넓은 학문 분야에 대한 번역 성능을 추가적으로 검증할 필요가 있다. 따라서 이러한 한계점을 극복하고 실시간 스트리밍 기반 구조를 구현하여 지연 시간을 최소화하여 실제 강의 환경에서 안정적으로 동작하는 실시간 다국어 자막 시스템으로 확장하는 것이 목표이며 이를 향후 연구로 남긴다.

ACKNOWLEDGEMENT

This work was supported by the Digital Innovation Hub project supervised by the Daegu Digital Innovation Promotion Agency(DIP) grant funded by the Korea government(MSIT and Daegu Metropolitan City) in 2025(No.225C000366, 2025년 R&BD협업 프로젝트(상용화 지원)).

REFERENCES

[1] Korea Education and Research Information Service, University Distance Education Status and Satisfaction Survey (2022-2 Semester Report), Korea Education and Research Information Service, 2023 [cited May 19, 2025]. [Online]. Available: https://keris.or.kr/main/ad/pblcte/selectPblcteRRInfo.do?mi=1138&pblcteSeq=13733

[2] J. Won, “A study on foreign students’ perceptions of English-medium instruction in Korean universities,” Learner-Centered Curriculum and Instruction Research, vol. 19, no. 21, pp. 377–406, 2019, doi: 10.22251/jlcci.2019.19.21.377.

[3] Ministry of Education, Status of Foreign Students in Domestic Higher Education Institutions (1999–2024), Korea Education Statistics Service, 2024.

[4] X. Li, B. Curle, and Y. Zhan, “Towards deeper learning in EMI lectures: The role of cognitive load,” J. Further Higher Educ., vol. 48, no. 2, pp. 205–221, May 2023, doi: 10.1080/01434632.2023.2248078.

[5] S. Liao, J.-L. Kruger, and S. Doherty, “The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension,” J. Specialised Translation, no. 33, pp. 70–98, Jan. 2020, doi: 10.26034/cm.jostrans.2020.549.

[6] E. N. Casey, K. M. Kalmon, J. S. Schultz, and C. G. Bryan, “Descriptive analysis of interpreter service mode costs and usage in Northwestern Wisconsin pre- and peri-COVID-19,” BMC Health Serv. Res., vol. 25, no. 1, pp. 1–9, Jan. 2025, doi: 10.1186/s12913-025-12248-0.

[7] K. Kuhn, V. Kersken, B. Reuter, N. Egger, and G. Zimmermann, “Measuring the accuracy of automatic speech recognition solutions,” ACM Trans. Accessible Comput., vol. 16, no. 4, pp. 1–23, 2024.

[8] N. Sethiya and C. K. Maurya, “End-to-end speech-to-text translation: A survey,” Comput. Speech Lang., vol. 90, p. 101751, 2025.

[9] A. Rafiei Oskooei, E. Caglar, İ. Şahin, A. Kayabay, and M. S. Aktas, “Whisper, translate, speak, sync: Video translation for multilingual video conferencing using generative AI,” in Proc. Int. Conf. Comput. Sci. Its Appl., Cham, Switzerland: Springer Nature, Jun. 2025, pp. 217–234.

[10] A. A. Attia, D. Demszky, T. Ogunremi, J. Liu, and C. Espy-Wilson, “CPT-boosted Wav2vec 2.0: Towards noise-robust speech recognition for classroom environments,” arXiv preprint arXiv:2409.14494, Sep. 2024. [Online]. Available: https://arxiv.org/abs/2409.14494

[11] J. Wang, J. Kim, S. Kim, and Y. Lee, “Exploring lexicon-free modeling units for end-to-end Korean and Korean-English code-switching speech recognition,” arXiv preprint arXiv:1910.11590, Oct. 2019. [Online]. Available: https://arxiv.org/abs/1910.11590

[12] E. Matusov, A. Mauser, and H. Ney, “Automatic sentence segmentation and punctuation prediction for spoken language translation,” in Proc. IEEE ICASSP, Apr. 2008, pp. 353–356, doi: 10.1109/ICASSP.2008.4518807.

[13] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,”

IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443–445, 2003.

[14] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-domain neural speech enhancement with very low algorithmic latency," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, 2022. [Online]. Available: https://zqwang7.github.io/publications/TASLP2022_STFTlowlat.pdf

[15] R. J. Nasir and H. A. Abdulmohsin, "A hybrid method for speech noise reduction using log-MMSE," Iraqi J. Sci., pp. 860–875, 2025.

[16] I. López-Espejo, A. Joglekar, A. M. Peinado, and J. Jensen, "On speech pre-emphasis as a simple and inexpensive method to boost speech enhancement," arXiv preprint arXiv:2401.09315, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.09315

[17] E. Loweimi, S. M. Ahadi, T. Drugman, and S. Loveymi, "On the importance of pre-emphasis and window shape in phase-based speech recognition," in Advances in Nonlinear Speech Process., LNCS vol. 7911, Springer, 2013, pp. 160–167, doi: 10.1007/978-3-642-38847-7_21.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. 40th Int. Conf. Mach. Learn., PMLR vol. 202, pp. 28492–28518, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2212.04356

[19] A. Rao, "Transcribing educational videos using Whisper: A preliminary study on using AI for transcribing educational videos," arXiv preprint arXiv:2307.03200, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.03200

[20] Anonymous, "Improving domain-specific ASR with LLM-generated contextual descriptions," arXiv preprint arXiv:2407.17874, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.17874

[21] M. Barański, J. Jasiński, J. Bartolewska, S. Kacprzak, M. Witkowski, and K. Kowalczyk, "Investigation of Whisper ASR hallucinations induced by non-speech audio," arXiv preprint arXiv:2501.11378, Jan. 2025.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[23] H. Yu, Y. Cho, G. Park, and M. Kim, "Krongbert: Enhanced factorization based morphological approach for the korean pretrained language model," Information Processing & Management, vol. 62, no. 3, p. 104072, 2025. DOI: 10.1016/j.ipm.2025.104072

[24] T. Kim, S. Jeong, and Y. Song, "Kogec: Korean grammatical error correction with pre-trained translation models," arXiv preprint arXiv:2506.11432, 2025. DOI: 10.48550/arXiv.2506.11432

[25] D. Min, S. Nam, and D. Choi, "A Study on Improving the Accuracy of Korean Speech Recognition Texts Using KcBERT," Journal of KIISE, vol. 51, no. 12, pp. 1115–1124, 2024. DOI: 10.5626/JOK.2024.51.12.1115

[26] S. Ma, L. Dong, S. Huang, D. Zhang, A. Muzio, S. Singhal, H. H. Awadalla, X. Song, and F. Wei, "DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders," arXiv preprint arXiv:2106.13736, Jun. 2021, doi: 10.48550/arXiv.2106.13736.

[27] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Sov. Phys. Dokl., vol. 10, pp. 707–710, 1966.

[28] "Korean Single Speaker Speech Dataset," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasets/bryanpark/korean-single-speaker-speech-dataset

Authors

DongHyeon Shin received his three-year B.S. degree from Yeungnam University College, Korea, in 2024, and completed an advanced major program in Computer Engineering at Yeungnam University in 2025. He was admitted to the M.S. program in the Department of Computer Engineering at Yeungnam University in 2025, where he is currently pursuing his master's degree. His current research interests include large language models (LLMs), retrieval-augmented generation (RAG), and deep learning.

Oh-Gyu Kwon received his B.S. degree in Computer Engineering from Yeungnam University, Gyeongsan, Korea, in 2002, and his M.S. degree in 2009. He is currently pursuing a Ph.D. in Computer Engineering while serving as the CEO of Mobifin. His recent research interests focus on AI-based application software.

ManKi Min received B.S. degrees in the Department of Computer Engineering from Yeungnam University, Korea, in 2025. He is currently a M.S. student in the Department Computer Engineering at Yeungnam University. His current research interests include IEEE 802.11 MAC protocol, semantic communication and network optimization.

Minseo Yoon received B.S. degrees in the Department of Computer Engineering from Yeungnam University, Korea, in 2025. He is currently a M.S. student in the Department of Computer Engineering at Yeungnam University. His current research interests include IEEE 802.11 based pose estimation, full duplex.

DoHyun Oh received the B.S. degree in Computer Engineering from Yeungnam University College, Daegu, Republic of Korea, in 2023, and the M.S. degree in Computer Engineering from Yeungnam University, Gyeongsan, Korea, in 2025. He is currently a Ph.D. student in the Department of Computer Engineering at Yeungnam University. His current research interests include large language model, natural language processing and visual document understanding.

Jae-Woo Ryu received the B.S. degree in Computer Engineering from Yeungnam University College, Daegu, Republic of Korea, in 2014. He has been pursuing an M.S. degree in Computer Engineering at Yeungnam University, Gyeongsan, Korea, since 2025. He is currently a Senior Researcher at the Corporate R&D Center of Nearnetworks. His research interests include AI-based application software, intelligent systems, and data-driven artificial intelligence technologies.

Young Deok Park received his B.S. degree in Computer Engineering from Sungkyunkwan University, Suwon, South Korea in 2012, and the M.S. and Ph.D. degrees in Computer Science and Engineering from Pohang University of Science and Technology (POSTECH), Pohang, South Korea in 2014 and 2019, respectively. He worked as a PostDoctoral Researcher with the Department of Computer Science and Engineering, POSTECH until August 2019, and worked at Samsung Electronics Co., Ltd., as a Staff Engineer until February 2021. Since March 2021 he has been an Assistant Professor with the Department of Computer Engineering, Yeungnam University, Gyeongsan, South Korea. His current research interests include IEEE 802.11 PHY/MAC protocol and LTE/NR system design. He is a member of the IEEE.