pISSN 2005-8063
eISSN 2586-5854
2025. 9. 30.
Vol.17 No.3
pp. 71-82

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.3.071

Automated English speaking assessment with large language models: A framework for multi-dimensional scoring and feedback generation

Jong In Kim · Hyung-Bae Jeon · Jeon Gue Park*

AI Lab, Tutorus Labs Inc. Daejeon, Korea

Abstract

Traditional automated English speaking assessment systems are limited in their ability to provide meaningful improvement guidance to learners, as they typically focus solely on generating overall scores. To address this limitation, this study proposes an integrated LLM-based assessment model that simultaneously performs quantitative score prediction and qualitative feedback generation for comprehensive English speaking evaluation. The model combines Whisper, BEATs, and QFormer for multidimensional audio feature extraction, utilizes ChatGPT-generated training data for Llama-based instruction tuning, and employs large language models to predict scores across four domains (task completion, delivery, accuracy, appropriateness) while generating specific feedback and corrections. Experimental results demonstrate reasonable correlations with human evaluators (Pearson correlation coefficients ranging from 0.730 to 0.789) and feedback quality with average scores above 4.0 points in all evaluation categories as validated by LLM-as-a-Judge methodology.

Keywords: Automated Speaking Assessment (ASA), Feedback Generation, Multitask-based speaking assessment, LLM-based speaking assessment

1. 서론

최근 교육 현장에서는 자동 말하기 평가 시스템의 도입 필요성이 제기되고 있다. 기존의 인간 평가 방식은 평가자 간 주관적 판단의 차이로 인해 신뢰성과 객관성 확보에 어려움이 있다(Isaacs, 2017). 동일한 발화에 대해서도 평가자의 경험과 전문성에 따라 상이한 점수가 부여되는 경우가 발생하며, 평가자 효과, 평가 척도의 해석 등 다양한 요인이 평가 결과에 영향을 미치는 것으로 보고되고 있다(Fan & Yan, 2020).

대규모 언어 능력 평가에서는 이러한 문제와 함께 실용적인 한계도 나타난다. 충분한 전문 평가 인력의 확보와 관리에는 상당한 비용과 시간이 소요되며, 평가자의 피로도와 개인적 편향이 결과에 미치는 영향을 완전히 배제하기 어렵다. 또한 전통적인 평가 방식은 실시간 채점과 즉각적인 피드백 제공에 제약이 있어, 학습자의 신속한 학습 개선을 지원하는 데 한계를 보인다. 이러한 배경에서 자동 말하기 평가 시스템은 일관된 평가 기준 적용과 효율적인 대규모 평가 처리, 실시간 피드백 제공의 가능성을 제시하는 대안으로 관심을 받고 있다.

* jgpark@tutoruslabs.com, Corresponding author, Corresponding author
Received 4 August 2025; Revised 9 September 2025 Accepted 9 September, 2025

이러한 한계를 보완하기 위해 자동 말하기 평가 시스템이 효과적인 대안으로 주목받고 있다. 자동화된 평가 시스템은 사전에 정의된 평가 기준과 알고리즘에 따라 평가 과정을 표준화한다. 이를 통해 평가자 간 편차를 최소화하고 평가의 일관성을 확보할 수 있다. 또한, 대규모 응시자 처리 측면에서 자동 평가는 효율성을 제공한다. 인력 비용을 절감하고 평가 시간을 단축할 수 있을 뿐만 아니라, 동시다발적인 평가 실시가 가능하다.

자동 말하기 평가에 관한 연구는 1990년대 초반 자동 음성 인식 기술의 본격적인 도입을 계기로 점차 발전하기 시작하였다. 초기 연구들은 주로 음소 길이, 음향 스펙트럼 등의 저차원 음향적 특징을 분석하여 발음 정확도와 유창성을 자동 평가하는 시스템을 중심으로 이루어졌다. Eskenazi(1998)의 CMU FLUENCY, Franco et al.(2000)의 Eduspeak는 읽기 및 발음 평가 영역에서 음성 인식을 보다 광범위하게 적용하여, 다양한 학습자 집단을 대상으로 자동 평가의 범위를 넓혔다. 또한 PhonePass(Bernstein et al., 1999)는 발음, 읽기 유창성, 리듬, 어휘 등 언어적 요소를 평가 항목에 포함하여 의미 기반 평가의 가능성을 모색하였다. Witt & Young(1997)이 제안한 GOP(Goodness of Pronunciation) 기법은 GMM(Gaussian Mixture Model)을 활용해 음소 단위 발음 정확도를 정량화함으로써, 발음 평가의 정밀도를 높이는 데 기여하였다. 2000년대 중반 이후에는 대규모 시험 데이터의 축적, 상용 시스템의 개발, 실시간 피드백에 대한 수요 증가 등이 맞물리며, 자동 말하기 평가는 본격적인 교육 분야에 적용되었다. 예를 들어, Zechner et al.(2007)의 SpeechRater™ v1.0은 TOEFL iBT Practice Online에서 운영된 자유 발화 자동 채점 시스템으로, 주로 유창성(fluency) 중심의 약 40개의 피처를 활용해 자동 채점한다.

최근에는 인공지능 기술의 발전에 따라, 딥러닝 기반의 자동 말하기 평가 연구가 활발히 진행되고 있다. 딥러닝, 시퀀스 투 시퀀스, 어텐션 메커니즘 등 신경망 계열 모델이 도입되면서, DNN-HMM 기반의 발음 오류 탐지, 음향-언어 통합 모델, 다차원 자연어 처리(NLP) 기반 평가 모델, 다국어 음향 모델, End-to-End 평가 모델 등이 연구되고 있다. 예컨대 Chen et al.(2018), Fu et al.(2020), Metallinou & Cheng(2014), Yu et al.(2015) 등의 연구는 실제 데이터를 기반으로 딥러닝 기반 자동 평가 시스템의 정밀도를 검증하였다. 또한 Bannò & Matassoniet al.(2023)은 자기 지도 학습 기반의 말하기 평가 모델을 제시하여, Wav2vec 모델을 활용하여 평가 모델을 구축하였다.

그럼에도 불구하고 기존 자동 말하기 평가 시스템은 점수 평가와 피드백 제공의 통합적 접근에서 근본적인 한계를 보인다. 현재 대부분의 시스템은 발음, 유창성, 문법 등에 대한 점수 산출에는 높은 정확도를 보이지만, 점수 근거를 바탕으로 한 구체적 피드백 제공에는 제한적이다. 점수 산출 모델과 피드백 생성 모델이 독립적으로 구성되어, 평가 과정에서 추출된 언어학적 특징이나 오류 패턴이 피드백에 효과적으로 활용되지 못한다. 이로 인해 학습자는 자신이 받은 점수의 근거를 명확히 이해하기 어렵고, 점수 향상을 위한 구체적인 개선 방향을 파악하기 힘들다. 예를 들어 발음 점수가 낮을 때 어떤 음소에서 문제가 발생했는지, 어떻게 개선해야 하는지에 대한 세부 정보가 부족하다.

본 연구는 기존의 영어 말하기 자동평가 시스템이 단순한 점수 산출에 그치는 구조적 한계를 극복하고, 평가와 동시에 교육적 피드백까지 제공하는 통합형 자동평가 모델을 개발하는 것을 목적으로 한다. 구체적으로, 본 연구는 음성 기반 대형 언어 모델(LLM)을 활용하여 기존의 일방향적 평가 시스템을 평가-피드백 통합 시스템으로 전환하는 것을 목표로 한다. 이 모델은 4가지 평가 영역[TC(task completion), DL(delivery), AC(accuracy), AP(appropriateness)]에 대한 정량적 점수 산출과 더불어, 각 영역별 구체적인 개선점 및 학습 방향을 제시하는 맞춤형 텍스트 피드백, 첨삭을 동시에 제공한다. 즉 단순히 점수를 산출하는 데 그치지 않고, 학습자의 말하기 능력 향상을 실질적으로 지원하기 위해 정량적 평가와 정성적 피드백을 동시에 제공하는 말하기 평가 시스템을 구현하고자 하였다.

이러한 통합적 구조는 기존 시스템의 정량적 평가와 정성적 피드백 간의 분리 문제를 해결하고, 학습자의 학습 개선을 지원하는 실용적 도구로서 기능한다. 본 연구는 말하기 평가와 피드백 제공을 단일 프로세스로 통합함으로써, 자동 말하기 평가 분야에서 교육 현장 활용 가능성을 높이는 데 기여한다.

2. 음성 기반 LLM을 활용한 말하기 평가-피드백 통합 시스템

2.1. 문제 정의

본 연구는 한국인 학습자의 영어 발화를 평가하기 위한 멀티태스크 학습 프레임워크를 설계한다. 모델은 학습자의 음성 신호, 평가 프롬프트, 그리고 Whisper를 통해 생성된 STT 전사 결과를 입력으로 받아, 실제 영어 교사의 평가 방식에 따라 총 9개의 평가 태스크를 동시에 수행한다.

이 때, 학습자의 음성 신호는 Whisper와 BEAT 인코더를 거쳐 발음 특성, 억양이 반영된 벡터로 변환된다. 프롬프트 지시문은 텍스트 형태로 입력되어 LLM 토크나이저와 LLM 인코더를 통해 임베딩된다. Whisper 기반 STT 전사 결과는 프롬프트의 세부 항목으로 포함된다. 최종적으로 음성 벡터와 언어 벡터는 결합되어 한국인의 영어 발화 모델 훈련에 활용된다.

모델이 수행하는 태스크는 점수 산출과 피드백 생성으로 구분된다. 점수 산출 태스크는 과제 완수(TC), 전달력(DL), 정확성(AC), 적합성(AP)의 4개 영역에 대해 1–5점 척도의 점수를 예측하는 회귀 태스크이다. 피드백 생성 태스크는 동일한 4개 평가 영역에 대해 각 발화의 강점이나 개선점을 자연어로 설명하는 문장을 생성한다. 추가로 학습자 발화의 문법적·표현상 오류를 수정한 문장을 생성하는 발화 오류 교정 태스크가 포함된다. 따라서 전체 모델은 하나의 입력으로부터 발화 오류 교정 1개, 점수 예측 4개, 피드백 생성 4개로 총 9개 태스크를 멀티태스크 방식으로 동시에 처리한다. 각 태스크의 출력은 회귀값 또는 자연어 문장 형태로 제공되어, 학습자의 말하기 수행에 대한 종합적인 평가와 구체적인 개선 방향을 제시하는 데 활용된다.

2.2. 데이터 증강

영어 말하기 평가 모델 학습을 위한 충분한 데이터 확보는 중요한 과제이다. 기존에는 학습자의 발화에 대한 피드백 데이터가 거의 존재하지 않거나 체계적으로 구축되지 않은 상황이었다. 이러한 데이터 부족으로 인해 모델이 다양한 오류 상황에서 적절한 피드백을 생성하도록 학습하기 어려웠다. 이 문제를 해결하기 위해 본 연구에서는 그림 1과 같이 ChatGPT를 활용한 데이터 증강 기법을 적용하였다.

첫째, 첨삭 데이터 증강이다. 실제 한국인 학습자의 말하기 데이터에 포함된 오류 문장을 기반으로, 기존에 수동으로 전사된 텍스트를 활용하여 ChatGPT 4.1을 통해 수정 문장을 생성하였다. 이로써 오류 문장과 수정 문장의 쌍을 구성하여, 모델이 문장 수준의 오류와 그에 대한 교정 사례를 입력-출력 형태로 학습할 수 있도록 하였다.

둘째, 피드백 데이터 증강이다. 말하기 평가의 네 가지 영역인 과제 완수도(TC), 전달력(DL), 정확성(AC), 적절성(AP)에 대해 다양한 오류 유형을 반영한 발화를 구성하고, 각 발화에 대해 ChatGPT를 활용하여 영역별 맞춤형 피드백을 생성하였다. 예를 들어 정확성 영역에서는 문법, 어휘 등의 각종 오류를, 전달력 영역에서는 유창성 문제에 대해 피드백하는 발화를 구성하였다. 이렇게 구축된 증강 데이터는 모델이 각 평가 영역의 특성을 이해하고 영역별로 차별화된 피드백 양식을 학습하는 데 활용되었다.

셋째는 STT(Speech-to-Text) 기반의 전사 데이터 증강이다. 실제 음성 데이터에는 발화자의 개인정보가 포함될 수 있으므로, 데이터 활용에 앞서 비식별화(de-identification) 과정이 요구된다. 기존 데이터셋에서는 개인정보 보호를 위해 전사 데이터 내 개인을 특정할 수 있는 요소를 제거하거나 변환하여 비식별화를 수행하였다. 다만 이 과정에서 고유명사나 맥락 단서가 일부 누락될 수 있으며, 이는 후속 분석에서 데이터의 완전성에 영향을 미칠 수 있다.이러한 한계를 보완하고 데이터의 활용 범위를 확장하기 위해 전사 데이터 증강을 적용하였다.

2.3. 모델 구조

2.3.1. 전체 시스템 개요

본 연구는 영어 말하기 자동 평가와 교육적 피드백 생성을 동시에 수행할 수 있는 멀티 테스크 기반 심층 신경망 기반 모델을 제안한다. 제안된 시스템은 기존의 SpeechLLM 기반 멀티모달 프레임워크인 SALMONN의 아키텍처를 기반으로 하며, 이를 영어 말하기 평가 도메인에 적용하였다(Tang et al., 2023).

본 시스템은 음성의 언어적 내용과 음향적 특성을 동시에 분석하여 종합적인 평가와 각 항목별 구체적인 피드백을 제공하는 것을 목표로 한다. 전체 시스템(그림 2)은 크게 네 가지 주요 모듈로 구성된다:

(1) 듀얼 음성 인코딩 모듈, (2) Q-former 기반의 크로스모달 정렬 모듈, (3) 다중 태스크 말하기 평가 모듈 (4) LLM 기반 피드백 및 첨삭 생성 모듈이다. 모든 모듈은 end-to-end 방식으로 학습되며, 멀티태스크 학습을 통해 점수 예측과 텍스트 피드백 생성을 동시에 수행한다.

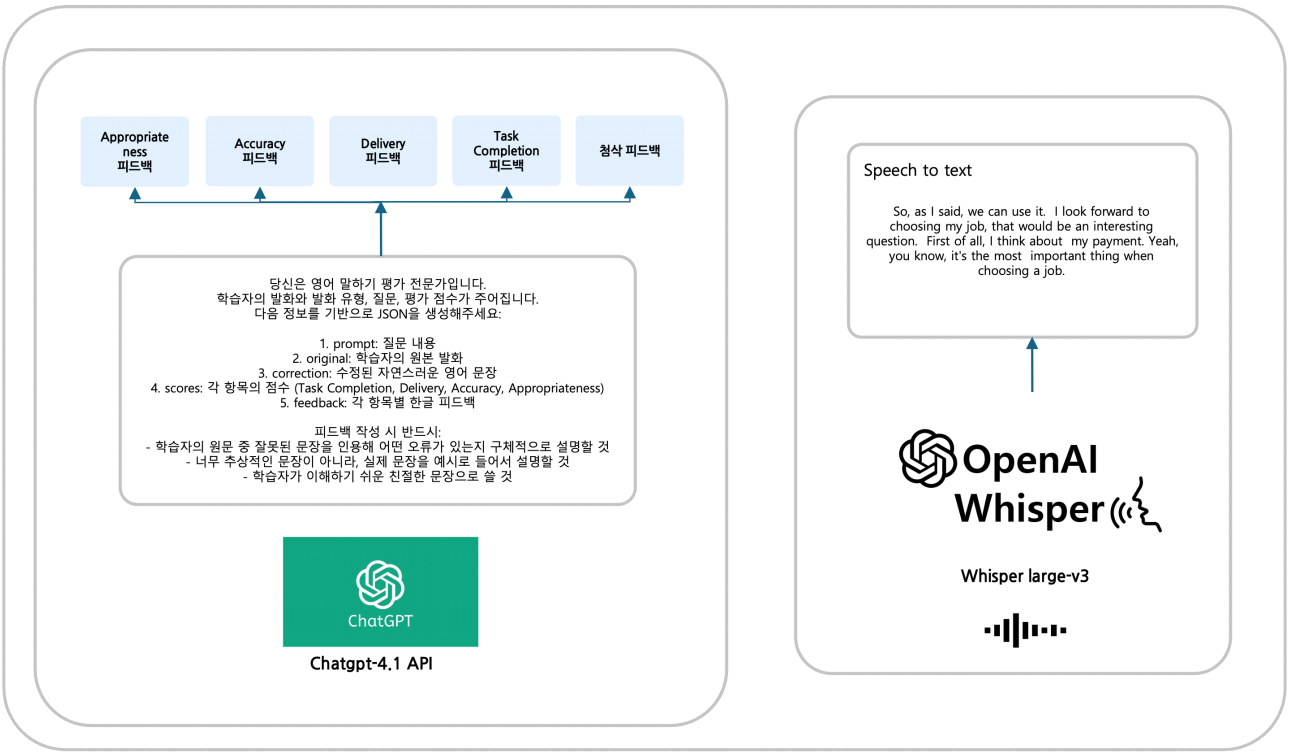

그림 1. 영어 말하기 자동 평가 데이터에 적용된 데이터 증강 방법
Figure 1. Workflow of data augmentation methods applied to automatic english speaking assessment data

2.3.2. 입력 데이터 및 전처리

모델의 입력은 학습자의 원시 음성 신호, 질문 프롬프트, 그리고 STT 결과로 구성된다. 각 인코더에 맞는 특화된 전처리 과정을 거치는데, Whisper 인코더의 경우 음성 신호를 log-mel 스펙트로그램으로 변환한다. 이를 위해 25 ms 윈도우와 10 ms 홉 사이즈로 STFT를 적용한 후, 80개의 mel-scale 필터뱅크를 거쳐 로그 스케일로 정규화한다. BEATs 인코더는 원시 오디오 신호를 직접 입력으로 받아 처리한다.

2.3.3. 모듈 1. 듀얼 음성 인코딩 모듈

본 모델은 Whisper와 BEATs 두 인코더를 병렬로 사용하여 음성의 다양한 측면을 포착하는 것을 목적으로 한다. 두 인코더는 서로 다른 음성 정보를 추출하도록 설계되었다. Whisper는 언어적 내용에 대한 의미론적 이해를 제공하는 것을 목표로 하는 반면, BEATs는 음색, 리듬, 피치 변화 등의 음향적 특성과 초분절적 요소을 포함한 음향에 대한 표현을 제공하는 것을 목적으로 한다. 음성의 언어적 내용 추출을 위해 사전학습된 Whisper-Large-v3 turbo 모델의 인코더 부분을 활용한다.

입력된 80차원 log-mel 스펙트로그램은 먼저 1D 컨볼루션 레이어를 통과하여 초기 특징을 추출한다. 컨볼루션 레이어를 거친 특징 벡터에는 사인-코사인 기반의 Positional Encoding이 시간적 순서 정보를 모델에 제공한다. 이후 트랜스포머 인코더 레이어를 순차적으로 통과하며, 각 레이어는 멀티헤드 어텐션 메커니즘과 피드포워드 네트워크(FFN)로 구성된다. 트랜스포머 레이어에서는 잔차 연결(residual connection)과 레이어 정규화(layer normalization)를 적용하여 학습 안정성을 확보하고, 드롭아웃을 통해 과적합을 방지한다.

본 연구에서는 음성의 음향적 특성 포착을 위해 사전학습된 BEATs(Bidirectional Encoder representation from Audio Transformers) 인코더를 병렬로 적용하였다. 원시 오디오는 BEATs 인코더에 직접 입력되어 처리된다. 입력된 원시 오디오 신호는 고정된 길이의 패치로 분할되고, 각 패치는 패치 임베딩 레이어를 통해 고차원 벡터로 변환된다. 변환된 패치 임베딩에는 위치 정보를 제공하는 위치 임베딩이 추가된다. 임베딩된 패치 시퀀스는 여

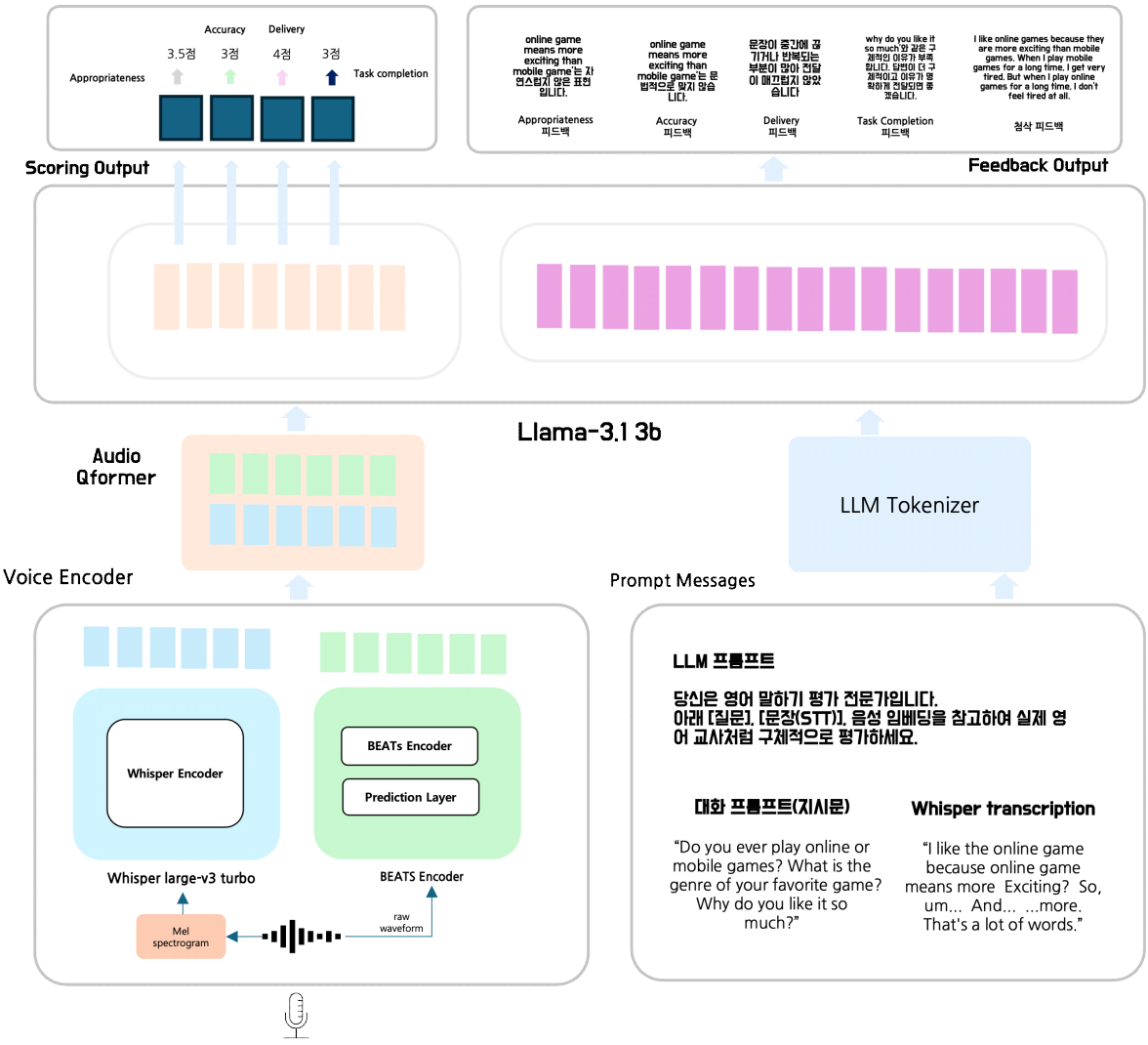

그림 2. 영어 말하기 자동 평가를 위한 LLM 기반 점수 예측 및 피드백 생성 통합 모델
Figure 2. LLM-based integrated score prediction and feedback generation model for automated english speaking assessment

러 층의 트랜스포머 인코더 레이어를 순차적으로 통과한다. 각 트랜스포머 레이어에서는 멀티헤드 셀프 어텐션을 통해 패치 간의 관계를 모델링하고, 피드포워드 네트워크를 통해 비선형 변환을 수행한다. 각 레이어는 잔차 연결과 레이어 정규화를 적용하여 안정적인 학습을 보장한다. BEATs 인코더는 최종적으로 시간 축을 따라 오디오의 음향적 특성을 인코딩한 특징 벡터 시퀀스를 출력한다.

2.3.4. 모듈 2. Q-former 기반의 크로스모달 정렬 모듈

본 연구에 적용한 Q-Former는 BERT 기반 구조를 사용하여 학습 가능한 쿼리 토큰을 통해 음성 특징에서 필요한 정보만을 선택적으로 추출한다. Q-Former 초기화 시 BERT-base-uncased 설정을 기반으로 하되, 교차 어텐션 레이어를 모든 블록에 추가하고 음성 입력 차원을 인코더 폭으로 설정하였다. 학습 가능한 쿼리 토큰은 가우시안 분포로 초기화되며, 고정된 개수로 생성하였다. 입력 처리 과정에서 Whisper와 BEATs의 연결된 음성 임베딩에 대해 레이어 정규화를 적용한 후, 어텐션 마스크를 생성하였다. Window-level Q-Former가 활성화된 경우, 음성 임베딩을 지정된 윈도우 크기와 스트라이드로 분할하였다. 구체적으로 설정된 윈도우 시간과 스트라이드 비율에 따라 계산된 크기로 겹치는 윈도우를 생성하였다. Q-Former의 핵심 연산에서는 학습 가능한 쿼리 토큰을 배치 크기만큼 확장하고, BERT 모델의 교차 어텐션을 통해 음성 임베딩과 상호작용시켰다. 이 과정에서 쿼리 토큰이 음성 임베딩으로부터 중요한 정보를 선택적으로 추출하며, 어텐션 마스크를 통해 유효한 음성 구간만을 처리하였다. Q-Former 출력은 선형 투영 레이어를 통해 LLaMA 모델의 임베딩 차원으로 변환하였다. Window-level 처리가 적용된 경우, 분할된 윈도우들을 다시 배치 차원으로 재구성하였다. 최종적으로 변환된 음성 임베딩과 해당 어텐션 마스크를 생성하여 후속 언어 모델 처리에 사용하였다. Whisper와 BEATs 인코더의 출력을 결합한 멀티모달 표현은 Q-Former를 통해 다음과 같이 변환된다.

2.3.5. 모듈 3. 다중 태스크 말하기 평가 모듈

본 연구에서는 음성 임베딩으로부터 4개 영역의 말하기 점수를 예측하는 보조 태스크를 구성하였다. Q-Former에서 생성된 음성 임베딩에 어텐션 마스크를 적용하여 유효한 음성 구간만을 선택한 후, 시간 축을 따라 평균 풀링을 수행하여 고정 크기의 특징 벡터를 생성하였다. 생성된 특징 벡터는 과제 완성도 TC, DL, AC, AP를 위한 4개의 독립적인 선형 회귀 헤드에 각각 입력되어 개별 점수를 예측하도록 설계하였다. 각 헤드는 단일 뉴런으로 구성되어 연속적인 점수를 출력하며, 예측된 점수와 실제 점수 간의 평균 제곱 오차 손실을 통해 학습하였다. 각 태스크 i ∈ {TC , DL , AC , AP} 에 대해, 공유된 인코딩 벡터 z 를 기반으로 한 선형 분류기를 통해 다음과 같이 계산한다.

$$\forall i \in TC, DL, AC, AP,\ \hat{y}_i = W_i z_{pooled} + b_i \qquad (1)$$

2.3.6. 모듈 4. LLM 기반 피드백 및 첨삭 모듈

본 연구에서는 피드백 생성을 위해 LLaMA 기반의 Llama-3.2-3B-Instruct 모델을 활용하였다. 이 모델은 3B 파라미터를 가진 경량화된 언어모델이다. 모델의 입력은 멀티모달 정보로 구성된다. Q-Former에서 추출된 오디오 토큰, 학습자에게 제시된 질문 프롬프트, STT 전사 결과, 그리고 피드백 유형을 지정하는 구조화된 평가 태그가 포함된다.

태그는 <feedback_task_completion>, <feedback_pronunciation>, <feedback_fluency>, <feedback_grammar>, <correction> 태그를 정의하여 문법 교정, 과제 완성도, 발음, 유창성, 문법 피드백을 그림 3과 같이 각각 생성하도록 구조화하였다.

피드백 생성을 위해 SFT(Supervised Fine-Tuning) 방식의 instruction tuning을 수행하였다. 기존에 부족했던 4개 영역의 피드백 데이터와 첨삭 데이터를 새롭게 생성하여 학습 데이터로 활용하였다. 이를 통해 기존의 4개 말하기 평가 점수와 함께 총 9개의 출력을 생성할 수 있는 통합 모델을 구축하였다.

LLM 모델 학습 시에는 Cross-Entropy Loss를 적용하여 모델이 올바른 피드백 생성 패턴을 학습하도록 하였다. 모델은 미리 정의된 태그 구조에 따라 평가 결과를 생성하도록 설계하였다. 각 태그는 특정 평가 측면을 담당하며, 모든 출력은 한글로 생성되도록 제약을 설정하였다.

대규모 언어모델의 효율적인 파인튜닝을 위해 LoRA(Low-Rank Adaptation) 기법을 적용하였다. 전체 모델 파라미터를 모두 업데이트하는 대신, 트랜스포머의 각 레이어에 존재하는 어텐션 모듈에 저차원 어댑터를 삽입하여 파라미터 수를 효과적으로 줄이고 학습 효율을 높였다. 구체적으로는 멀티헤드 어텐션 구조의 Query, Key, Value 투영 행렬과 Output 투영 행렬에 LoRA 어댑터를 추가하여, 경량화된 학습이 가능하도록 하였다.

2.3.7. 멀티 테스크 학습의 손실 함수

본 연구는 다중 태스크 학습의 효율성을 높이기 위해 MultiTaskUncertaintyLoss를 활용하였다. 이 방법은 각 태스크의 불확실성(uncertainty)을 학습 가능한 파라미터로 도입하여, 태스크별 손실 가중치를 자동으로 조정할 수 있도록 한다. 전체 K개의 태스크에 대해, 각각의 태스크 k에는 손실 함수 L_k와 불확실성 파라미터 σ_k 가정의된다. 이때 σ_k 는 해당 태스크의 상대적인 난이도와 중요도를 반영하며, 계산의 안정성을 위해 $\log\sigma_k$형태로 모델에 적용된다. 이러한 불확실성 파라미터는 학습 과정에서 동적으로 최적화되며, 상대적으로 어려운 태스크에는 손실 기여도를 감소시키고, 쉬운 태스크에는 상대적으로 높은 가중치를 부여하도록 설계되었다. 이를 통해 모델은 다중 태스크 간의 상대적 중요도를 자동으로 조정함으로써, 보다 균형 있고 효율적인 학습이 가능하도록 하였다. 식 (2)은 본 연구에서 채택한 전체 손실 함수 L_{total} 은 식 (2)와 같다.

$$L_{total} = \exp(-\log\sigma_{main})\cdot L_{main} + \log\sigma_{main} + \exp(-\log\sigma_{tc})\cdot L_{tc} + \log\sigma_{tc} + \exp(-\log\sigma_{dl})\cdot L_{dl} + \log\sigma_{dl} + \exp(-\log\sigma_{ac})\cdot L_{ac} + \log\sigma_{ac} + \exp(-\log\sigma_{ap})\cdot L_{ap} + \log\sigma_{ap} \quad (2)$$

3. 실험

3.1. 데이터셋

본 연구에서는 한국지능정보사회진흥원(NIA)에서 2022년 구축한 '한국인의 주제적응형 영어말하기 평가데이터'(AI Hub, 2022)를 활용하였다. 이 데이터셋은 한국인 영어 학습자의 말하기 능력 평가를 위한 AI 자동평가 시스템 개발 목적으로 구축된 대규모 음성 코퍼스이다. 데이터셋은 총 2,640명의 한국인 화자로부터 수집된 1,052시간 분량의 영어 발화 음성데이터로 구성되어 있다.

본 연구에서 사용된 말하기 평가 데이터셋은 web-based workbench를 통해 전문 평가자들에 의해 채점되었으며, 모든 평가자는 말하기 인증 시험 평가 경험을 가진 인원으로 연구 목적에 맞추어 선발되었다. 실제 평가에 앞서 평가자들은 연구진이 제공한 평가 기준 교육을 이수하고 calibration 절차를 거쳐 채점 성향을 표준화하였으며, 연구진이 사전에 기채점한 표준 발화(reference)를 루브릭에 따라 채점한 결과가 기준 점수와 평균 2점 이상 차이를 보이지 않을 때만 본 평가에 참여할 수 있었다. 이후 각 발화는 2인 이상의 전문가가 과제 완수도(TC), 전달력(DL), 정확성(AC), 적합성(AP)의 네 가지 영역에 대해 1–5점 척도로 평가하였다.

실험에서 데이터셋은 전체 34,325개 발화 중 30,780개(약 89.4%)를 훈련용으로, 3,545개(약 10.6%)를 테스트용으로 사용하였다.

3.2. 실험

본 연구에서는 LLM 모델은 Llama-3.2-3B-Instruct(https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct)를 사용하였으며, 음성 인코더는 whisper-large-v3-turbo(https://github.com/ openai/whisper)를 적용하고, 음향 정보 추출에는 BEATs_iter3_ finetuned_on_AS2M_cpt2(https://github.com/microsoft/unilm/tree/master/beats) 사전 훈련된 모델을 활용하였다.

본 실험에서는 음성과 텍스트 모달리티 간의 정렬 성능을 향상시키기 위해 윈도우 기반 Q-Former(https://github.com/salesforce/LAVIS/blob/main/lavis/models/blip2_models/Qformer.py)를 도입하였다. 학습 과정에서 Q-Former를 업데이트 대상에 포함시켜 음성-텍스트 정렬 능력을 강화하였으며, 파라미터 효율성을 고려하여 LoRA 기법을 적용하였다(rank=4, alpha=8, dropout=0.1). 특히 음성 쿼리 토큰은 1개로 설정하였는데, 이는 발화 전체를 대표하는 단일 임베딩을 효과적으로 학습하기 위함이다. 다수의 쿼리 토큰을 사용할 경우 정렬 과정에서 불필요한 노이즈와 중복 표현이 발생하여 발화 단위의 평가와 피드백 생성에 불안정성을 초래할 수 있다. 반면 단일 쿼리 토큰은 텍스트 표현과 직접적으로 1:1 정렬되어 alignment가 단순하고 안정적이며, 발화 전체의 특성을 요약적으로 포착하기 위함이다.

Whisper 및 BEATs 모델은 파라미터를 동결하여 사전 훈련된 인코더의 안정적인 특징 추출 능력을 보장하였다. 텍스트 처리에서는 최대 길이를 512 토큰으로 제한하였으며, 종료 심볼로 "</s>"를 사용하고 보조 임베딩 차원은 3072차원으로 설정하였다.

음성 정보와 텍스트 정보를 모두 활용하는 것은 말하기 능력은 단순히 텍스트의 언어적 내용만으로 평가될 수 없기 때문이다. 발음, 억양, 리듬, 유창성, 발화 속도, 휴지(pause)와 같은 음향적 특성이 핵심적인 평가 기준을 이룬다. 따라서 텍스트 전사만을 활

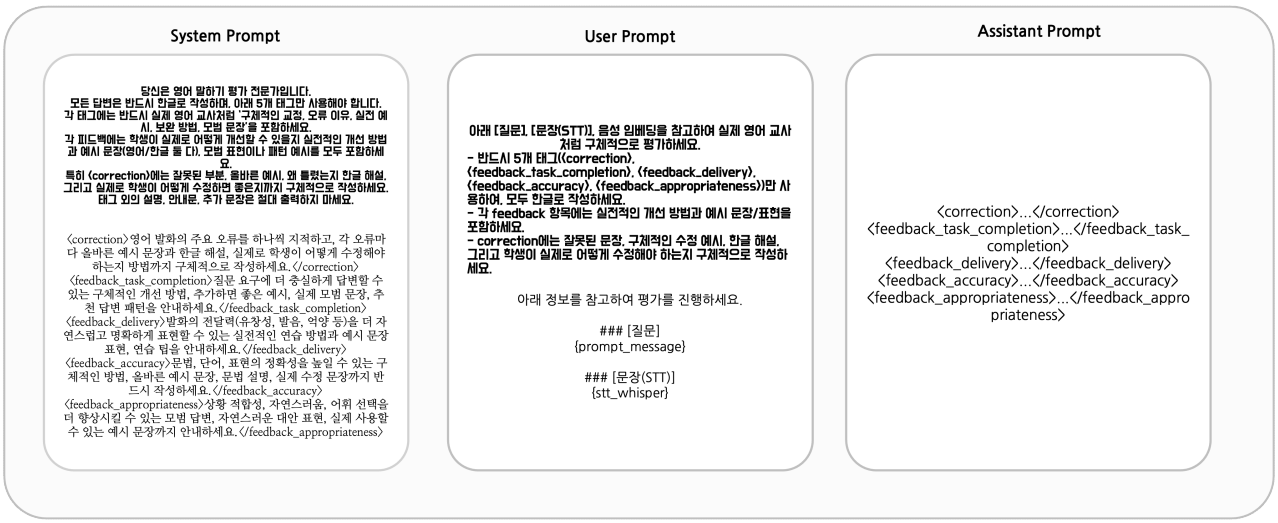

그림 3. LLM 기반 영어 말하기 평가 및 피드백 생성 프롬프트 체계

Figure 3. LLM prompt for English speaking assessment system

용하는 접근은 말하기 능력의 상당 부분을 간과하게 되며, 본 연구는 이를 보완하기 위해 Whisper 기반 음성 인코더에서 추출된 음향 표현과 LLM 텍스트 인코더에서 추출된 언어적 의미 표현을 동시에 활용하였다.

훈련 과정에서는 AdamW 옵티마이저를 사용하여 최대 2 에포크 동안 학습을 수행하였다. 초기 학습률은 3e-5로 설정하였으며, 웜업 스텝 400회 동안 웜업 시작 학습률 1e-6에서 점진적으로 증가시켰다. 최소 학습률은 1e-5, 가중치 감쇠는 0.01, beta2는 0.999로 구성하였다. 메모리 효율성을 위해 훈련 및 평가 배치 크기를 모두 1로 설정하였으며, 그래디언트 누적을 16회 반복마다 수행하여 배치 크기를 확보하였다. 에포크당 반복 수는 32,000회로 설정하였으며, AMP를 활성화하여 메모리 사용량을 최적화하였다. 분산 학습 환경에서는 8개의 GPU(NVIDIA RTX A6000 48GB)를 사용하여 병렬 처리를 수행하였다. 이 때 음성은 음성은 GPU Out-of-Memory 이슈를 고려하여, 60초만 학습하였다.

모델의 추론 과정에서는 최대 512개의 새로운 토큰을 생성하도록 설정하였으며, 4개의 빔을 사용한 빔 서치를 적용하고 샘플링을 비활성화하여 결정론적 생성을 보장하였다. 디코딩 파라미터로는 온도 0.5, top-p 0.9, 반복 페널티 1.0, 길이 페널티 1.0을 사용하였다.

본 연구에서는 제안한 모델의 성능을 검증하기 위해 기존의 전통적인 기계학습 및 딥러닝 기반 모델들과 비교 실험을 수행하였다. 비교 대상으로는 CNN(Convolutional Neural Network), LSTM(Long Short-Term Memory), 그리고 Random Forest 등을 선정하였다.

CNN은 총 3개의 합성곱 층으로 구성되며, 첫 번째 층은 Conv1d(입력 39채널 → 64채널, 커널 크기 5), 두 번째 층은 Conv1d(64→128, 커널 크기 5), 세 번째 층은 Conv1d(128→256, 커널 크기 3)으로 설계하였다. 각 층마다 ReLU 활성화 함수를 적용하였고, 마지막으로 전역 평균 풀링(Global Average Pooling)을 거쳐 256-128-64-출력으로 으로 회귀를 수행하였다. LSTM 모델은 은닉 차원 128, 층 수 2의 LSTM으로 처리하였으며, 각 층 사이에 드롭아웃 0.3을 적용하였다. 최종 은닉 상태는 64차원은 출력 4차원으로 매핑된다. Random Forest모델은 실험에서는 n_estimators=100, max_depth=15, random_state=42를 하이퍼파라미터로 설정하였다.

추가적으로, 한국인 영어 학습자의 발화 데이터에서 오류 문장을 분석하였다. 분석은 정확성(AC), 과제 완수(TC), 적합성(AP), 전달력(DL)의 네 가지 평가 관점에서 이루어졌다.

3.3. 평가지표

본 연구에서는 사람과 모델의 평가 점수의 상관관계를 측정하는 6가지 지표와 피드백의 품질을 측정하는 LLM 기반 평가(LLM-as-a-Judge)를 주요 지표로 사용하였다. 모델의 상관관계를 측정하는 Pearson 상관계수(PCC)를 통해 인간 평가자의 채점과 모델 예측 간의 일치 정도를 평가하였다. PCC는 −1과 +1 사이에서 분포하고 절댓값이 클수록 높은 상관관계를 의미한다. 결정계수(R^2)를 활용하여 모델이 실제 채점 결과의 변동성을 포착하는 정도를 분석하였으며, 0과 1 사이의 범위에서 높은 값일수록 우수한 예측 능력을 의미한다. 순서가 있는 평가 척도의 특성을 고려하여 QWK(Quadratic Weighted Kappa)를 적용하였는데, 이는 평가자 간 합의 수준을 측정하면서 오차의 크기에 비례하여 가중치를 적용하는 특징을 갖는다. 예측 정확성 측정을 위해 MAE(Mean Absolute Error)를 사용하여 예측치와 참값 간의 평균 절대 편차를 계산하였으며, 낮은 수치일수록 정밀한 예측을 나타낸다. RMSE(Root Mean Square Error)를 통해서는 예측 오류의 제곱평균제곱근을 구하여 특히 큰 편차에 대한 민감도를 확인하였다. 마지막으로 ±1점 정확도는 예측치가 기준값 대비 1점 범위 내에서 일치하는 사례의 비율을 산출하여 현실적 활용 관점에서의 모델 신뢰성을 검증하였다

LLM-as-a-Judge는 언어 모델 자체를 평가 도구로 사용하는 방식으로, 명시적인 평가 기준을 포함한 프롬프트를 기반으로 대상 텍스트에 대한 점수를 산출하도록 구성하였다. 프롬프트는 총 5가지 항목에 대해서 적절한지 여부를 평가하도록 설정되었으며, 각 항목에 대해 1점에서 5점 사이의 점수를 부여하도록 하였다. 평가자는 고정된 LLM으로 설정하고, 동일한 평가 기준과 입력 형식을 유지하여 응답 간 일관성을 확보하였다. 평가에는 ChatGPT 4.1 API를 활용하였다

4. 실험 결과

4.1. 한국인의 영어 말하기 데이터의 오류 문장 분석

한국인의 주제적응형 데이터셋에서 관찰한 한국인 영어 학습자의 발화는 말하기 평가의 오류 네 가지 핵심 기준인 정확성(AC), 전달력(DL), 적절성(AP), 과업 수행(TC)의 측면에서 분석하였다. 정확성 측면에서는 한국어 문법 체계의 전이로 인해 어순 오류(예:“I my age is twenty three”), 시제 오류(예:“I go school ”), 관사 생략(예:“I have cat”), 전치사 오용(예:“I’m attending in Suwon Science College”) 등이 빈번하게 나타난다. 또한 직역적 표현(예:“my dad better me better than me”)이 나타난다. 전달력 측면에서는 특정 발음의 부정확성과 한국어식 평탄한 억양이 두드러지며, 필러(“uhm”, “sorry”)의 과다 사용과 단어 반복(예:“after after after come back in my home”)은 유창성을 떨어뜨린다. 한편, 적절성 측면에서는 대화 맥락 유지가 어려워 질문에 단편적이거나 회피적인 응답(예:“sorry I can’t answer this question”)을 보이는 경우가 많지만, 동시에 “I use Kakao Talk app”, “I finished my military service”와 같이 한국적 사회·문화 맥락을 반영한 발화가 관찰된다. 과업 수행 측면에서는 질문에 충분히 답변하지 못하거나 발화를 중단하는 경우(예:“it is too hard for me”)가 흔히 나타나지만, 일부 발화에서는 불완전한 문장 구조에도 불구하고 핵심 요소를 포함하며 응답을 완성하려는 경향성(예:“when i choosing a job i consider three things first it's a working time”)이 관찰된다.

4.2. 말하기 평가 모델 성능 평가

본 연구에서 개발한 LLM 기반 영어 말하기 평가 모델의 성능을 4개 평가 영역(TC, DL, AC, AP)에 대해 표 1과 같이 분석하였다.

모델과 인간 평가자 간의 Pearson 상관계수(PCC) 분석 결과, DL 영역에서 0.7893으로 가장 높은 상관관계를 보였으며, AC 0.7771, TC 0.7521, AP 0.7301 순으로 나타났다. 모든 영역에서 0.73 이상의 상관관계를 달성하였으며, DL와 AP 영역 간 0.059의 차이를 보였다. ±1점 오차 범위 내 정확도는 DL 84.85%, AC 83.84%, TC 78.14%, AP 77.38%로 측정되었다. DL와 AC 영역에서는 80% 이상의 정확도를 보인 반면, TC과 AP 영역에서는 78% 내외의 정확도를 나타냈다. 영역 간 최대 7.47%포인트의 차이가 관찰되었다. 평가자 간 일치도를 나타내는 QWK(Quadratic Weighted Kappa) 값은 DL 0.6991, AC 0.6778, TC 0.6430, AP 0.6227로 산출되었다. 모든 영역에서 0.6 이상의 값을 기록하였으며, DL 영역에서 가장 높은 일치도를, AP 영역에서 가장 낮은 일치도를 보였다.

표 1. 영어 말하기 자동 평가 시스템의 영역별 성능 분석
Table 1. performance analysis of automated english speaking assessment system

지표	Accuracy	Appropriateness	Delivery	Task Completion
PCC	0.7771	0.7301	0.7893	0.7521
R^2	0.4882	0.3927	0.537	0.3729
QWK	0.6778	0.6227	0.6991	0.643
MAE	0.565	0.6566	0.5495	0.6473
RMSE	0.7155	0.8381	0.6974	0.8179
±1점 정확도	0.8384	0.7738	0.8485	0.7814

4.3. 베이스라인 모델과의 성능 비교(PCC)

표 2는 다양한 베이스라인 알고리즘과 제안 모델의 영어 말하기 평가 자동화 실험 결과를 제시한다. 각 모델의 성능은 AC, AP, DL, TC의 4가지 지표를 통해 평가되었으며, Average PCC는 이들 지표의 평균 피어슨 상관계수를 나타낸다. 전통적인 머신러닝 기반 모델인 Random Forest, Extra Trees, XGBoost의 Average PCC는 각각 0.5822, 0.5893, 0.5815로 측정되었다. 딥러닝 기반 모델들은 LSTM, EfficientNet1D, Conformer, CNN 순으로 비교되었으며, 이들 중 CNN이 AC(0.658), AP(0.631), DL(0.678), TC(0.665), Average PCC(0.658)를 기록하였다. LSTM과 Conformer 모델은 유사한 수준의 전반적 성능을 보였으나, 개별 지표에서는 상이한 결과를 나타냈다. 제안된 모델은 모든 평가 지표에서 기존 베이스라인 모델들보다 높은 성능을 기록하였다. 구체적으로 AC(0.7771), AP(0.7301), DL(0.7893), TC(0.7521), Average PCC(0.76215)의 결과를 보였다.

표 2. Pearson 상관계수 기반 영어 말하기 평가 모델의 성능 분석
Table 2. pearson correlation-based performance analysis of English speaking assessment models

Model	Accuracy	Appropriateness	Delivery	Task Completion	Average PCC
CNN	0.658	0.631	0.678	0.665	0.658
LSTM	0.610	0.576	0.625	0.600	0.603
EfficientNet1D	0.639	0.62	0.658	0.649	0.6415
Conformer	0.641	0.6	0.656	0.637	0.6335
Random Forest	0.568	0.564	0.596	0.601	0.5822
Extra Trees	0.574	0.572	0.601	0.61	0.5893
XGBoost	0.569	0.558	0.597	0.602	0.5815
cProposed Model	0.7771	0.7301	0.7893	0.7521	0.76215

4.4. 피드백 평가

LLM-as-a-Judge 구조를 활용하여 모델이 생성한 피드백의 품질을 평가하였다. 평가는 TC, DL, AC, AP, Correction의 5개 영역에 대해 구체성, 정확성, 적합성을 기준으로 수행되었다. 표 3과 같이 모든 영역에서 평균 4.1점 이상의 피드백 품질 점수를 획득하였다. 표 4는 제안 모델의 피드백 및 첨삭 생성 결과 예시를 나타낸다.

표 3. LLM-as-a-Judge 기반 피드백 평가
Table 3. LLM-as-a-Judge Based Feedback Assessment.

평가 항목	평균 점수
Accuracy	4.35
Appropriateness	4.13
Delivery	4.42
Task Completion	4.32
Correction	4.52

5. 논의

5.1. 모델 성능 분석

본 연구에서 개발한 LLM 기반 모델은 모든 평가 영역에서 베이스라인 모델들보다 높은 성능을 보였다. 기존 자동화 말하기 평가 시스템들과의 비교에서 본 연구의 성능 개선이 확인되었다. ETS의 SpeechRater가 보고한 0.73 수준의 상관계수와 비교할 때, 본 연구는 DL 영역에서 0.789, AC 영역에서 0.777로 개선된 성능을 보였다(Zechner et al., 2014). 또한 대부분의 딥러닝 기반 말하기 평가 연구들이 데이터셋에 0.6–0.8 범위의 상관계수를 보고한 것과 비교하여 본 연구는 모든 영역에서 0.730 이상의 상관계수를 달성하여 자동화 평가 시스템으로서 합리적인 성능 수준을 보여준다. 영역별 성능 분석에서는 차이가 관찰되었다. DL 영역에서 0.789로 가장 높은 성능을 보인 것은 시스템이 전달력 영역에서 효과적으로 분석할 수 있음을 시사한다. 반면 AP 영역에서 0.730으로 상대적으로 낮은 성능을 보인 것은 화용

론적 적절성 판단의 복잡성을 반영된 것으로 추측된다.

5.2. 선행 연구와의 차별성

기존 연구들은 주로 점수 예측 또는 피드백 생성 중 하나에만 집중해왔다. SpeechRater와 같은 상용 시스템들은 주로 전체 점수나 특정 영역 점수 산출에 초점을 맞추었으며, 대부분의 학술 연구들도 인간 평가자와의 상관관계 향상에만 집중하였다. 반면 피드백 생성 연구들은 주로 작문 평가 영역에 제한되어 있거나, 말하기 평가에서도 일반적인 개선 제안 수준에 머물렀다. 본 연구는 단일 모델로 점수 예측과 구체적 피드백 생성을 동시에 수행하는 통합적 접근을 제시하였으며, 이는 기존 연구들에서 분리되어 다뤄졌던 평가와 진단 기능을 하나의 시스템으로 구현한 차별적 접근이다.

전통적인 자동화 평가 시스템들은 전체 점수 산출이나 발음, 유창성 등 1–2개의 제한된 영역에만 집중하는 경향을 보였다. 예를 들어, 많은 기존 연구들이 음성학적 특징에 기반한 발음 평가나 단일 차원의 유창성 평가에 초점을 맞추었다. 이와 달리 본 연구는 TC, DL, AC, AP의 4개 영역에 대해 각각 독립적인 평가와 피드백을 제공한다. 이는 학습자에게 발음이나 문법뿐만 아니라 담화 수준의 세분화된 진단 정보를 제공할 수 있다는 교육적 장점을 갖는다.

기존 연구들이 피드백 생성 시스템의 품질을 체계적으로 평가하지 않았던 것과 달리, 본 연구는 LLM-as-a-Judge 방법론을 활용하여 생성된 피드백의 품질을 정량적으로 평가하였다. 대부분의 기존 연구들은 피드백이 생성되었다는 사실 자체에 초점을 맞추거나 소수의 전문가가 주관적으로 평가하는 수준에 머물렀다. 본 연구에서는 LLM을 이용한 평가를 통해 모든 영역에서 평균 4점 이상의 품질 점수를 달성하였다.

6. 결론

본 연구는 LLM을 활용하여 영어 말하기 평가에서 점수 예측과 피드백 생성을 동시에 수행하는 통합적 평가 시스템을 개발하였다. TC, DL, AC, AP의 4개 영역에 대해 각각 점수를 예측하고 개선 방향을 제시하는 피드백을 생성하는 모델을 구축하였으며, 기존의 CNN, EfficientNet1D, Conformer 등 베이스라인 모델들과 성능을 비교 분석하였다.

본 연구의 한계점은 다음과 같다. 첫째, 피드백의 구체성과

표 4. 말하기 평가 및 피드백 결과 사례
Table 4. Illustrative Results of Speaking Assessment and Feedback

항목	ESPEAK_3053_TL_STUD_F_SG_06	ESPEAK_2986_TM_WORK_F_CC_03
정답 점수(TC/DL/AC/AP)	1.50/1.50/1.50/1.50	3.50/3.00/3.00/3.50
모델 점수(TC/DL/AC/AP)	1.59/1.54/1.57/1.44	3.59/2.97/3.04/3.38
피드백(Task Completion)	"where you and your family had to fix something in your house has to skip a ten"라고 하셨는데, 질문에서 요구한 상황 설명과 가족의 역할, 결과 등 구체적인 내용이 부족합니다. 예를 들어, "My family and I had to fix a leaky faucet in our kitchen"처럼 구체적인 상황을 설명하면 더 좋습니다.	"the most difficult project that i have ever been on in the world then was um project a big it was a big project"와 같이 질문에서 요구한 프로젝트의 내용, 어려움, 경험을 모두 언급하셨습니다. 다만, 프로젝트의 구체적인 내용을 좀 더 자세히 설명하면 더 좋겠습니다.
피드백(Delivery)	"where you and your family had to fix something in your house has to skip a ten"처럼 문장이 자연스럽지 않고 중간에 'uh', 'um'과 같은 망설임이 많아 전달력이 떨어집니다. 천천히 생각을 정리한 후 말하면 더 명확하게 전달할 수 있습니다.	"the project was about topic celebrity event and um i've find it difficult because um i was new new worker so i i have no experience of doing build a plan and and and make a progress"와 같이 문장이 길고 중간에 반복이 많아 전달이 매끄럽지 않았습니다. 문장을 짧게 나누고, 불필요한 반복을 줄이면 더 자연스럽게 들릴 수 있습니다.
피드백(Accuracy)	"where you and your family had to fix something in your house has to skip a ten"에서 'has to skip a ten'은 문법적으로 올바르지 않습니다. 올바른 표현은'fix something in your house' 또는 'fix a leaky faucet in our kitchen'입니다.	"i've find it difficult", "i have no experience of doing build a plan and and and make a progress"와 같이 문법적인 오류가 있습니다. 예를 들어, "i've find it difficult"는 "I found it difficult"로, "i have no experience of doing build a plan and and and make a progress"는 "I didn't have any experience with building a plan or making progress"로 수정해야 자연스럽습니다.
피드백(Appropriateness)	"where you and your family had to fix something in your house has to skip a ten"은 질문에 맞는 답변이 아닙니다. 질문에서 요구한 상황, 가족의 역할, 결과를 구체적으로설명해야 합니다. 예를 들어, "My family and I fixed a leaky faucet in our kitchen"처럼 적절한 답변이 필요합니다.	"i ask help to my coworkers and thanks to my coworkers i can finish that project"에서 "ask help"는 "ask for help"로, "thanks to my coworkers"는 "thanks to my coworkers"가 더 자연스럽습니다. 또, "i can finish that project"보다는 "I was able to finish the project"가 더 적절합니다.

개인화 측면에서 개선 여지가 있다. 현재는 일반적인 피드백을 제공하지만, 학습자의 수준이나 학습 목표에 따른 맞춤형 피드백 생성 방안이 필요하다. 둘째, 피드백 품질 평가에서 LLM으로 생성한 답변을 다른 LLM으로 평가하는 방식을 사용하였다는 점이다. 이는 LLM 간의 편향이나 일관성 문제를 완전히 배제할 수 없으며, 인간 전문가에 의한 독립적 평가가 추가로 필요하다. 셋째, 기술적 측면에서 LLM 생성 과정에서 답변이 누락되는 경우가 발생하였으며, GPU 메모리 제약으로 인해 다양한 모델 구조나 하이퍼파라미터 실험에 제한이 있었다. 이러한 기술적 한계로 인해 본 연구는 LLM을 활용한 말하기 평가 연구의 초기 단계로서의 성격을 갖는다.

이러한 한계점에도 불구하고, 본 연구는 LLM을 활용한 영어 말하기 평가에서 점수 예측과 피드백 생성을 통합한 새로운 접근법을 제시하였다. 기존 모델들과 비교하여 합리적인 성능을 달성하였으며, 특히 통합적 평가 시스템으로서의 가능성을 확인하였다. 향후 화용론적 평가 정확성 향상과 개인화된 피드백 생성, 그리고 기술적 안정성 개선에 대한 지속적인 연구가 필요하다.

References

AI Hub(2022). Korean topic-adaptive speaking assessment data. Retrieved from https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71418

Bannò, S., & Matassoni, M. (2023, January). Proficiency assessment of L2 spoken English using wav2vec 2.0. *Proceedings of the 2022 IEEE Spoken Language Technology Workshop* (SLT) (pp. 1088-1095). Doha, Qatar.

Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate Corporation.

Chen, L., Tao, J., Ghaffarzadegan, S., & Qian, Y. (2018, April). End-to-end neural network based automated speech scoring. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6234-6238). Calgary.

Eskenazi, M., & Hansma, S. (1998). The fluency pronunciation trainer. *Proceedings of the STiLL Workshop.* Marholmen, Sweden.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in psychology, 11*, 330.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000). The SRI EduSpeakTM system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTiLL 2000: Intelligent Speech Technology in Language Learning.* Dundee, Scotland.

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication, 116*, 86-97.

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication, 49*(12), 861-873.

Isaacs, T. (2017). Fully automated speaking assessments: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson (Eds.) *The Routledge handbook of contemporary English pronunciation.* London, UK: Routledge.

Metallinou, A., & Cheng, J. (2014, September). Using deep neural networks to improve proficiency assessment for children English language learners. *Proceedings of INTERSPEECH 2014* (pp. 1468-1472). Singapore, Singapore.

Townshend, B., Bernstein, J., Todic, O., & Warren, E. (1998, May). Estimation of spoken language proficiency. *Proceedings of the STiLL Workshop on Speech Technology in Language Learning* (pp. 177-180), Marholmen, Sweden.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., ... Zhang, C. (2023). Salmonn: Towards generic hearing abilities for large language models. arXiv, https://doi.org/10.48550/arXiv.2310.13289.

Witt, S. M. (2000). *Use of speech recognition in computer-assisted language learning* (Doctoral dissertation). University of Cambridge, Cambridge, UK.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication, 30*(2-3), 95-108.

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., ... Qian, Y. (2015, December). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 338-345). Scottsdale, AZ.

Zechner, K., Bejar, I. I., & Hemat, R. (2007). Toward an understanding of the role of speech recognition in nonnative speech assessment (TOEFL iBT Research Report No. 02). Educational Testing Service (ETS), Princeton, NJ.

Zechner, K., Evanini, K., Yoon, S. Y., Davis, L., Wang, X., Chen, L., Lee, C. M., & Leong, C. W. (2014, June). Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language. *Proceedings of the ninth workshop on Innovative Use of NLP for Building Educational Applications* (pp. 134-142). Baltimore, MD.

• **김종인 (Jongin Kim)**
튜터러스랩스
서울 서초구 서초대로 243 4층
Tel: 010-5353-9863
Email: prows12@tutoruslabs.com
관심분야: 음성인식, 음성모델링

• **전형배 (Hyung-Bae Jeon)**
튜터러스랩스
대전 유성구 가정로 218 ETRI 융합기술연구생산센터 708호
Tel: 010-8893-5234
Email: hbjeon@tutoruslabs.com
관심분야: 음성인식, 음성모델링

• **박전규 (Jeon Gue Park)** 교신저자
튜터러스랩스
서울 서초구 서초대로 243 (서초동) 4층
Tel: 010-8466-1117
Email: jgpark@tutoruslabs.com
관심분야: 음성인식, 음성모델링

LLM 기반 영어 말하기 자동 평가: 다차원 채점 및 피드백 생성 프레임워크

김 종 인 · 전 형 배 · 박 전 규

튜터리스랩스

국문초록

기존의 자동화된 영어 말하기 평가 시스템은 단편적인 점수 생성에만 집중하여 학습자에게 의미 있는 개선 지침을 제공하는 능력이 제한적이다. 이러한 한계를 해결하기 위해 본 연구는 정량적 점수 예측과 정성적 피드백 생성을 동시에 수행하는 LLM 기반 통합 평가 모델을 제안한다. 제안 모델은 Whisper, BEATs, QFormer를 결합하여 다차원적 음성 특징을 추출하고, ChatGPT로 생성한 훈련 데이터를 활용하여, Llama 기반 instruction tuning을 통해 4개 영역(과업 완성도, 전달력, 정확성, 적절성)의 점수 예측과 구체적인 피드백 및 첨삭 문장을 생성한다. 실험 결과, 인간 평가자와의 상관관계는 영역별로 Pearson 상관계수 0.730–0.789를 나타냈으며, LLM-as-a-Judge 방법론으로 검증한 피드백 품질은 모든 평가 범주에서 평균 4.0점 이상을 기록하였다.

핵심어: 자동 말하기 평가, 피드백 생성, 멀티태스크 기반 말하기 평가, LLM 기반 말하기 평가

참고문헌

AI Hub(2022). 한국인의 주제적응형 말하기 평가데이터. Retrieved from https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71418