POSCO 청년 AI·Big Data 아카데미 31기
AI 프로젝트 보고

# 스티브 따라잡스

## : 최고의 발표를 위한 AI 컨설턴트


**프로젝트명 : 스티브 따라잡스**
**기수 및 조 : 31기 A반 3조**
**제 출 자 : 홍 정 택**

**(조장)홍정택 유창우 이민성 윤영채 정주영 허유진**

# 목 차

# Ⅰ. 프로젝트 개요

## 1.1. 추진 배경

최근 한국 사회에서는 발표·면접 상황에서의 커뮤니케이션 능력, 그중에서도 비언어적 전달력의 중요성이 크게 부각 되는 추세다. 직장인은 물론 취업 준비생까지 자신의 생각을 명확하고 설득력 있게 전달해야 하는 요구가 확대되고 있다. 목소리 톤, 발화 안정성, 제스처와 같은 표현 방식은 청중의 이해도와 신뢰 형성에 직접적인 영향을 준다.

비언어적 커뮤니케이션 연구의 세계적 권위자인 UCLA의 심리학자 Albert Mehrabian(1969)는 설득 상황에서 메시지의 영향력이 내용 자체보다 비언어적 요인(목소리·표정·태도)에 의해 결정되는 비중이 훨씬 크다는 연구 결과를 제시하며, 음성 표현과 시각적 신호가 의사소통 효과의 중심임을 강조했다. 해당 연구에 따르면 목소리(38%)와 표정(35%)이 전달의 핵심이며, 메시지 내용이 차지하는 비중은 7%에 불과하다.

MEAN VALUES OF NONVERBAL BEHAVIORS AND PERCEIVED PERSUASIVENESS FOR THREE DEGREES OF INTENDED PERSUASIVENESS

| Dependent variable | Reliability | *F* | *MS*$_e$ | Means for three degrees of intended persuasiveness | | |
|---|---|---|---|---|---|---|
| | | | | None | Moderate | High |
| Position cues | | | | | | |
| Distance | .95 | | | | | |
| Eye contact | .50 | 8.6 | 284 | 40% | 44% | 51% |
| Shoulder orientation | .96 | | | | | |
| Posture Cues | | | | | | |
| Arm position openness | .93 | | | | | |
| Arm position symmetry | .87 | | | | | |
| Leg position symmetry | .96 | | | | | |
| Reclining angle | .95 | 7.6 | 121 | 13.8° | 13.4° | 7.4° |
| Sideways lean | .63 | | | | | |
| Movement cues (number/minute) | | | | | | |
| Trunk swivel | .98 | | | | | |
| Rocking | .98 | | | | | |
| Head nodding | .97 | 5.2 | 8.3 | 6.1 | 6.4 | 7.5 |
| Gesticulation | .99 | 18.5 | 41 | 6.9 | 9.1 | 13.3 |
| Leg movement | .97 | | | | | |
| Foot movement | .96 | | | | | |
| Self-manipulation | .93 | | | | | |
| Facial cues | | | | | | |
| Pleasantness | .79 | | | | | |
| Activity | .49 | 9.6 | .35 | .70 | .92 | 1.13 |
| Verbal cues | | | | | | |
| Duration | .91 | | | | | |
| Rate | .77 | 27 | .42 | 2.16 | 2.35 | 2.93 |
| Volume | .88 | 29 | .47 | 1.84 | 1.98 | 2.65 |
| Intonation | .44 | 38 | .45 | 1.99 | 2.40 | 2.97 |
| Unhalting quality | .70 | 8.1 | .44 | 1.95 | 2.14 | 2.38 |
| Perceived persuasiveness | | | | | | |
| *E* viewing through one-way mirror | | 33 | 1.7 | 2.4 | 3.2 | 4.2 |
| *Es* viewing video-recording | | 41 | 1.03 | 2.4 | 3.1 | 4.0 |
| *Ss* viewing video-recording | | 29 | .29 | 1.6 | 1.8 | 2.3 |

▲ Mehrabian, A., & Williams, M. (1969). Nonverbal Concomitants of Perceived and Intended Persuasiveness. Journal of Personality and Social Psychology, 42.

발표·스피치 교육 시장은 활발하다. 실제로 스피치·발표 코칭 학원의 수강료와 수요는 꾸준히 상승하고 있으며, 기업들 또한 구성원들의 커뮤니케이션 역량 강화를 위해 발표 교육을 정규 프로그램으로 도입하고 있다. 직무 특성에 따라 프레젠테이션, 보고, 고객 응대 등 다양한 상황에서 비언어적 전달력의 차이가 성과와 직결되는 만큼, 발표 역량은 더 이상 선택이 아닌 필수 역량으로 자리 잡고 있다.

그러나 이러한 사회적 요구와 시장의 확대에도 불구하고, 많은 2030 세대는 여전히 발표 불안, 목소리 떨림, 속도 조절의 어려움, 시각적 표현의 미숙함 등 비언어적 요소에서 큰 부담을 느끼고 있다. 기존의 전문 교육은 높은 비용과 시간적 제약으로 인해 지속적인 훈련이 어렵고, 1:1 코칭 역시 접근성이 떨어진다는 한계가 존재한다. 특히 발표에서의 '표현 방식'은 개인의 습관과 불안 요인까지 영향을 받기 때문에, 일회성 학습만으로는 개선이 쉽지 않은 영역이다.

이러한 배경 속에서, 목소리·제스처 등 비언어적 발표 역량을 개인에 맞게 개선할 수 있는 AI 기반 피드백 서비스의 필요성은 더욱 커지고 있다. 사용자는 별도의 교육장이나 고가의 코칭 없이도 발표 영상만으로 전문 분석과 개선 방향을 바로 제공받을 수 있다. 시간과 비용의 제약을 낮추면서도 전문성과 개인화 수준을 높일 수 있다는 점에서, AI 기반 피드백이 효과적인 대안으로 주목받고 있다.

이에 본 프로젝트는 발표 역량의 중요성에 따른 2030 세대의 실질적인 어려움을 해결하기 위해 기획됐다. 특히 음성·제스처·강조 포인트 등 비언어적 요소 중심의 데이터 분석을 통해 보다 실용적인 피드백을 제공하는 것을 목표로 한다. 이는 누구나 쉽게 접근할 수 있는 발표 역량 개선 도구의 필요성을 충족함과 동시에, 발표 교육 시장에서의 새로운 접근법을 제시한다.

## 1.2. AI 서비스 분석

㈜제네시스랩의 '뷰인터'는 AI 기반 영상면접 코칭 서비스로, 사용자의 발표·면접 영상을 분석해 비언어적·언어적 요소에 대한 객관적 피드백을 제공한다. 감정·표정·시선 처리, 음색·음높이·발음 등 다양한 비언어적 특성을 정교하게 평가하고, 자체 STT 기술을 활용해 답변 내용과의 일관성까지 분석한다. 반복 연습 및 비교 분석 기능으로 높은 접근성을 확보하며, 고비용 오프라인 코칭의 대안을 제공하고 있다.

그러나 주로 **'면접 상황'에 특화**되어 있어, **일반 발표/프레젠테이션 맥락에서의 강조 포인트, 제스처 흐름 분석** 등은 상대적으로 **제한적**이다.

셀바스AI의 SpeechBox는 발음·억양·강세·발화 속도 등 음성 요소를 문장 단위로 정밀 진단하는 말하기 학습 솔루션이다. 음소 단위까지 세분화된 피드백, 텍스트 변환 정확도, 다양한 환경에서의 모듈형 제공 방식 등은 전문 학습 솔루션으로서 강점이 크다.

다만 발화 자체의 교정 기능은 탁월하지만, 발표 맥락에서 중요한 **'시각적 표현(제스처·시선·자세)'**나 **'맥락상 강조 구조'** 분석에 따른 **기능은 없다.**

스피처는 발표·스피치 영상을 기반으로 비언어적·언어적 요소를 동시에 분석하여 피드백을 제공하는 서비스다. 말의 속도, 논리 흐름, 제스처와 시선 처리, 강조 포인트 등 발표 상황 전체를 다각도로 분석한다는 점에서 개인 발표 역량 향상에 직접적이다.

그러나 주어진 결과가 **정적 리포트 중심**이라, '어떤 타이밍에 어떻게 고쳐야 하는지'와 같은 **실시간·구간별 가이드**가 **제한적**이다.

## 1.3. 전기수 분석

### (1) 11기 A3조 – POresentation: 초심자를 위한 발표 도우미 AI

11기 A3조의 POresentation은 발표 경험이 적은 사용자를 대상으로 발표 과정에서 나타나는 비언어적·언어적 특징을 다각도로 분석하고, 이를 정량화된 피드백으로 제공하는 발표 보조 AI 서비스다.

비언어적 요소 측면에서는 FaceNet 기반 Emotion Recognition 기능을 통해 발표자의 표정을 인식하고 시계열로 변화 패턴을 시각화해, 우수 발표자의 표정 패턴과 비교 분석이 가능하도록 구현했다. 또한 Gaze Tracking 기술을 활용해 발표자 시선의 움직임을 추적하고, 일정 기준을 벗어나는 시선 회피 횟수를 측정함으로써 발표 집중도와 안정성을 판단할 수 있다.

언어적 요소 분석에서는 Google Speech API 기반 STT 기술과 KoSpacing, KSS를 적용하여 발표 음성을 정확한 문장 형태로 정제하고, mecab 형태소 분석 + Doc2Vec 모델을 통해 실제 대본과 발화 대본 간의 유사도를 산출했다. 이를 통해 사용자가 대본을 어느 정도 숙지하고 있는지, 발표 중 어떤 부분에서 누락·이탈이 발생했는지를 정량적으로 확인할 수 있다.

POresentation은 이들 기술을 종합해 총 발표 시간, 발화 속도, 담화표지 사용 빈도, 시선 처리 횟수 등을 제공함으로써, 초심자가 스스로 발표 습관을 파악하고 교정할 수 있는 자기 진단형 훈련 솔루션으로 기능한다.

다만, 비언어적 요소가 '표정·시선' 중심으로 제한돼 있고, 제스처 분석이나 발표 맥락 기반의 강조 포인트 탐지 기능은 포함되지 않았다는 점에서 발전 여지가 있다.

### (2) 15기 A2조 – PO-inter: 비대면 AI 면접 연습 도우미

15기 A2조의 PO-inter는 비대면 AI 면접 환경을 중심으로 설계된 면접 대비 솔루션으로, 면접자의 시각적 행동 패턴과 발화 특징을 자동 분석해 면접 상황에서 개선이 필요한 요소를 제시하는 AI 서비스다. 비언어적 분석에서는 OpenCV 기반 Face Detection & Analysis를 적용하여 눈 깜빡임 빈도, 시선 및 고개 움직임을 카운팅하고, 선행 연구에서 제시된 기준값을 활용해 과도하거나 불안정한 행동을 탐지하도록 구성했다.

언어적 요소 분석에서는 CNN 기반 추임새(Filler) 분류 모델을 구축해 '음성 특징 벡터 → CNN → 추임새/침묵/기타'로 구분하는 구조를 사용했다.

또한 STT 변환 결과를 기반으로 발화 흐름·빠르기 등을 해석하여 면접자의 언어적 안정성을 평가하였다.

한편, PO-inter의 특징적인 요소는 KoGPT2-FineTuning 기반 답변 제안 기능이다. 약 13,000건의 합격 자기소개서를 학습시켜 사용자가 답변 아이디어가 떠오르지 않을 때 참고할 수 있는 문장을 생성하는 방식이지만, 면접 답변과 자기소개서 문체의 차이, 상황 맥락 고려 부족 등으로 실제 면접용 답변 생성에는 한계가 존재한다.

전체적으로 PO-inter는 AI 면접 환경에서 필요한 행동 패턴 분석 + 추임새 탐지 + 단순 답변 제안 기능까지 포괄하고 있으나, 분석의 정밀도와 실제 면접 맥락 반영 측면에서는 개선 가능성이 있다.

(3) 전기수 분석의 시사점

11기 A3조와 15기 A2조의 공통점은 발표/면접 수행 과정에서의 비언어적·언어적 요소를 정량화해 피드백하는 시스템을 구축했다는 점이다. 두 프로젝트 모두 시선·표정·발화 흐름 등 핵심 요소를 탐지하는 데 집중해, 사용자 스스로 개선 지점을 확인할 수 있다는 장점이 있다.

그러나 두 시스템 모두 제스처(손 움직임·몸 전체의 표현) 분석, 문맥 기반 강조 포인트 탐지, 실제 발표 맥락에 맞춘 구간별·상황별 코칭 기능 미흡이라는 공백을 남긴다.

따라서 이러한 한계를 보완하기 위해 음성-제스처-텍스트를 통합한 멀티모달 발표 분석 기능이 새로운 기회로 도출될 수 있을 것으로 판단된다.

# Ⅱ. 프로젝트 구조

## 2.1. 프로젝트 소개

본 프로젝트는 발표 영상 데이터를 기반으로 발표 피드백을 원하는 사용자의 비언어적·언어적 표현을 분석하고, 문맥에 따른 강조 구간을 자동으로 도출하는 발표 보조 AI 시스템을 구현하는 것을 목표로 한다. 발표 상황에서 실제 전달력에 영향을 미치는 요소들을 정량적으로 평가하고, 이를 종합적인 피드백 형태로 제공하는 데 목적이 있다.

프로젝트는 크게 맥락 중요도 분석(STT·NLP), 음성 분석(Voice), 컴퓨터 비전(CV) 기술을 결합한 멀티모달 구조로 설계되었다.

먼저 Whisper Large v3-Turbo를 활용해 발표 영상의 음성을 텍스트로 변환하며, 백색소음 주입을 통해 음성 인식 과정에서 발생할 수 있는 할루시네이션을 방지한 후 정확한 전체 발화를 추출한다. 이후 정제된 텍스트는 문장 단위로 분할되어 자연어 처리 모델을 통해 문맥 분석 및 강조 가능성 평가에 활용된다.

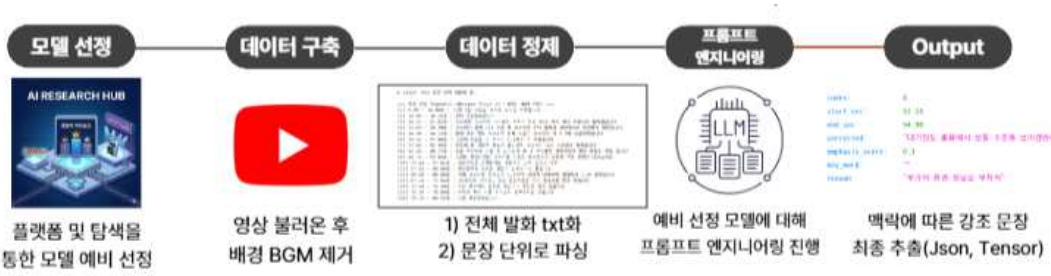
문맥 분석 단계에서는 LLM 기반 모델을 활용해 문장별 중요도를 산출하고, 강조 점수(emphasis score)와 강조 이유(reason)를 함께 생성함으로써 모델 판단의 해석 가능성을 높인다. 동시에 NLP 기반 후처리 모델을 적용해 강조 구간을 정량적으로 보정하고, 발표 흐름에 부합하는 결과를 도출한다.
음성 분석(Voice) 단계에서는 입력 영상으로부터 오디오 신호만을 추출한 뒤, 음성 성분을 분리하고 Librosa 라이브러리를 활용해 Mel-Spectrogram 기반 음성 특징을 추출한다. 개인마다 상이한 음높이(Pitch), 발화 강세 등의 차이를 보정하기 위해 정규화 과정을 수행하며, 이후 강조 발화 여부를 라벨링한 데이터를 CSV 형태로 구성한다. 해당 음성 특징 데이터는 Bi-LSTM 모델의 입력으로 사용되어 시간 흐름에 따른 음성 패턴과 강조 특성을 학습한다. 모델의 출력은 멀티모달 결합을 위한 특징 벡터(Tensor)와, 발화 시점 정보 및 강조 점수가 포함된 JSON 파일 형태로 생성되어 후속 분석 단계에 활용된다.

컴퓨터 비전 영역에서는 얼굴 표정, 시선, 제스처 등 비언어적 요소를 분석하여 사용자의 발표 습관과 전달 방식을 평가한다. 이러한 비언어적 분석 결과는 음성 및 텍스트 기반 분석 결과와 결합돼, 사용자에게 보다 입체적인 피드백을 제공한다.

최종적으로 본 프로젝트는 음성·텍스트·영상 정보를 시간 축 기준으로 통합 분석하여 사용자가 스스로 발표의 문제점을 인식하고 개선할 수 있도록 돕는 데이터 기반 발표 피드백 시스템을 제안한다. 이를 통해 기존의 주관적이거나 비용 부담이 큰 발표 코칭 방식을 보완하고, 누구나 접근 가능한 발표 역량 향상 도구로 활용 가능성을 제시한다.

## 2.2 모델 구현 및 결과

### 2.2.1. 맥락 분석 모델




▲ 맥락 분석 모델 파이프 라인

본 프로젝트의 맥락 분석 모델은 발표 영상으로부터 텍스트 기반 맥락 정보를 추출하고, 문장 단위의 의미적 중요도를 평가해 강조 구간을 자동 산출하는 것을 목표로 한다. 전체 파이프라인은 모델 선정 → 데이터 구축 → 데이터 정제 → 프롬프트 엔지니어링 → 결과 출력의 다섯 단계로 구성된다.


▲플랫폼 및 벤치마크 Searching

첫째, 모델 선정 단계에서는 플랫폼 및 벤치마크 결과를 고려해 모델을 결정했다. 모델별 Output을 바탕으로 한국어 문맥 이해, 추론 능력을 종합적으로 검토한 뒤 프로젝트 목적에 부합하는 모델을 최종 선정했다.

둘째, 데이터 구축 단계에서 발표 영상을 수집하고, Whisper Large v3 Turbo를 활용해 음성 인식을 수행했다. 영상 내 BGM을 약화하고 음성만을 정제한 후, Whisper의 환각 방지를 위해 백색소음 첨부 후, 인식된 텍스트를 토대로 전체 발화 내용을 시간 정보와 함께 구조화했다.

셋째, 데이터 정제 단계에서는 STT 결과에 포함된 오인식·띄어쓰기 오류 등을 KoSpacing과 KSS를 통해 교정한다. 이후 전체 문장을 txt 형태로 저장하고, 분석 단위를 통일하기 위해 문장 단위로 다시 분할한다. 이러한 전처리 과정을 통해 전체 발표 스크립트와 문장 데이터를 구축했다.

넷째, 프롬프트 엔지니어링 단계에서는 선정된 LLM을 기반으로 문장별 강조 가능성을 예측했다. 시스템 프롬프트 내에는 프로젝트 목적에 맞춘 강조 점수(emphasis_score) 산출 기준을 포함시키고, 모델이 각 문장 중요도를 판단할 수 있도록 맥락 정보를 제공했다. 또한 모델이 강조해야 하는 이유(reason)를 자동 생성하도록 설계해 분석의 해석 가능성을 높였다.

마지막으로, Output 단계에서는 모델이 산출한 강조 점수 및 강조 이유를 JSON 구조로 저장했으며, 후속 모델(KLUE-RoBERTa·Tensor 기반 scoring 모듈)을 통해 점수를 보정하여 최종 결과를 도출하였다. 생성된 결과는 문장별 강조 구간 탐지, 발표 지도용 시각화, 후속 멀티모달 분석과 연동되는 자료로 활용된다.

### 2.2.2 음성 분석 모델

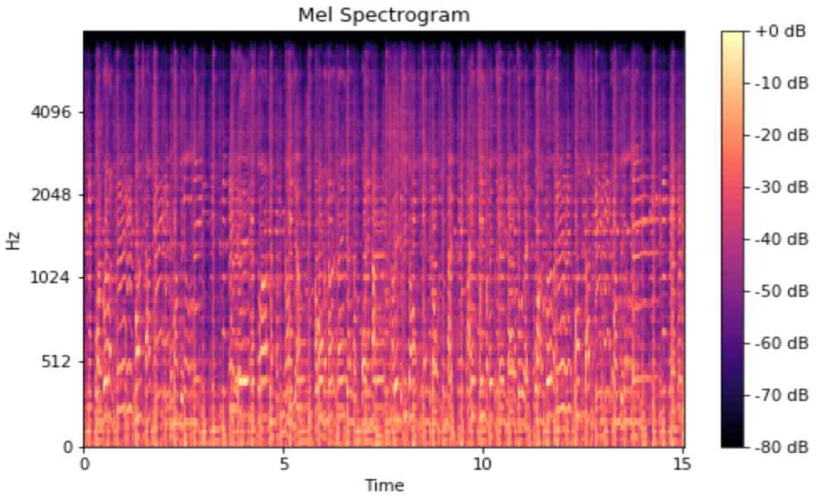



그림 1〈 Mel-Spectrogram 〉

음성의 특징을 분석할 때, 대부분의 논문에서 Mel-Spectrogram 또는 MFCC(Mel Frequency Cepstral Coefficients) 기술을 활용하는 것을 확인할 수 있었다.

Mel-Spectrogram은 원본 소리에 대해 STFT(Short Time Fourier Transform)를 진행한 뒤 Mel-Scale Filter를 거친다. 추출 과정이 단순하기 때문에 원본 음성의 많은 정보를 보존하고 있는 것이 특징이다. 따라서 선형 관계에 특화된 머신 러닝 모델보다는 비선형 관계에 강점을 보이는 딥러닝 모델에서 더 좋은 성능을 보이는 것이 일반적인 연구 결과이다.

MFCC는 위 방식에 더해 log를 취한 뒤, DCT(Discrete Cosine Transform) 과정을 진행하기 때문에, 모든 특징들 중에서도 핵심만 남기는 기술이다. 따라서 Mel-Spectrogram에 비하면 보다 선형 관계에 특화되어, 머신 러닝 모델에서 우수한 성능을 보인다.

| Comb | | At maximum recall (Validation set) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Validation set | | | | | Training set | | |
| F | Models | Recall | Acc. | Loss | Epoch | Time | Recall | Acc. | Loss |
| MFCCs 3s | CNN | 40.50% | 45.41% | 3.040 | 47 | 57m | 98.82% | 99.13% | 0.0787 |
| | LSTM | 32.44% | 38.01% | 3.778 | 47 | 91m | 77.86% | 84.65% | 0.6491 |
| | GRU | 35.50% | 39.53% | 3.979 | 46 | 107m | 85.69% | 89.77% | 0.4622 |
| | CRNN LSTM | **44.76%** | 46.10% | 4.233 | 42 | 74m | 96.17% | 97.01% | 0.1556 |
| | CRNN GRU | 44.40% | 47.81% | 3.470 | 41 | 88m | 90.80% | 92.79% | 0.3498 |
| MFCCs 1.5s | CNN | 38.82% | 44.44% | 3.134 | 45 | 52m | 98.19% | 98.76% | 0.1116 |
| | LSTM | 27.27% | 33.65% | 3.909 | 45 | 89m | 71.99% | 80.79% | 0.7931 |
| | GRU | 31.36% | 36.23% | 4.097 | 46 | 111m | 79.56% | 85.26% | 0.6354 |
| | CRNN LSTM | 42.36% | 43.95% | 4.404 | 43 | 76m | 91.97% | 93.81% | 0.2796 |
| | CRNN GRU | **43.35%** | 46.47% | 3.539 | 40 | 82m | 90.82% | 92.64% | 0.3476 |
| Mel Spectrogram 3s | CNN | 47.51% | 43.05% | 2.574 | 40 | 68m | 99.05% | 99.37% | 0.0761 |
| | LSTM | 26.27% | 33.32% | 3.832 | 44 | 95m | 65.85% | 75.80% | 0.9608 |
| | GRU | 28.23% | 33.48% | 4.271 | 43 | 110m | 75.96% | 83.36% | 0.7034 |
| | CRNN LSTM | 47.95% | 50.30% | 3.475 | 48 | 93m | 94.40% | 95.60% | 0.2140 |
| | CRNN GRU | **50.17%** | 53.13% | 3.003 | 45 | 98m | 90.64% | 92.70% | 0.3659 |
| Mel Spectrogram 1.5s | CNN | 44.21% | 49.28% | 2.957 | 41 | 68m | 98.76% | 99.14% | 0.0813 |
| | LSTM | 24.62% | 32.51% | 3.790 | 47 | 101m | 62.68% | 73.29% | 1.051 |
| | GRU | 27.56% | 32.51% | 4.071 | 46 | 115m | 79.56% | 85.46% | 0.6354 |
| | CRNN LSTM | 46.46% | 48.48% | 3.763 | 40 | 78m | 91.30% | 93.39% | 0.3109 |
| | CRNN GRU | **47.92%** | 50.71% | 3.286 | 46 | 102m | 89.74% | 91.95% | 0.3935 |

**[표 1] 〈 Feature Extraction Method와 DeepLeraning Model의 조합 성능 비교**

전반적으로 MFCC보다 Mel-Spectrogram을 활용할 때 더 좋은 성능을 보였는데, 이는 해당 방식이 고차원적 특징 학습에 유리해 딥러닝 모델에

적합하기 때문이라 할 수 있다. 그 중에서도 CNN GRU, CNN LSTM, CNN 순으로 성능이 높은 것을 확인할 수 있다. 해당 모델들의 학습 파라미터 수는 각각 150,65,915개, 8,981,307개, 9,33,211개로, 본 프로젝트의 제한된 기한을 고려하여 모델 선정 과정에서는 단순 성능뿐만 아니라 파라미터 효율성까지 참고했다.

MFCC는 Mel-Spectrogram에서 Log 변환 후, DCT(Discreat Cosine Transform)과정을 진행해 핵심적인 특징만 남긴다. Mel-Spectrogram에 비해 상대적으로 적은 정보만 가지므로, 선형관계에 특화된 머신러닝 모델에서 우수한 성능을 가진다.

| Networks | Iteration | Accuracy(%) | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| CNN | 700 | 96.4 | 0.954 | 0.952 | 0.953 |
| | 1400 | 97.00 | 0.962 | 0.964 | 0.963 |
| | 2810 | 98.4 | 0.972 | 0.971 | 0.971 |
| LSTM | 700 | 95.23 | 0.942 | 0.948 | 0.945 |
| | 1400 | 96.9 | 0.952 | 0.956 | 0.954 |
| | 2810 | 97.02 | 0.965 | 0.964 | 0.964 |
| Bi-LSTM | 700 | 96.6 | 0.950 | 0.959 | 0.959 |
| | 1400 | 96.6 | 0.949 | 0.940 | 0.940 |
| | 2810 | 97.7 | 0.962 | 0.968 | 0.965 |
| GRU | 700 | 94.4 | 0.942 | 0.945 | 0.943 |
| | 1400 | 95.30 | 0.939 | 0.930 | 0.933 |
| | 2810 | 96.10 | 0.971 | 0.968 | 0.969 |
| CONV-LSTM | 700 | 96.6 | 0.956 | 0.950 | 0.953 |
| | 1400 | 97.57 | 0.961 | 0.959 | 0.960 |
| | 2810 | 98.6 | 0.971 | 0.972 | 0.971 |

[표 2] 〈Deep Learning Model 성능 비교〉

선행 연구된 논문에서 주로 사용하는 딥러닝 모델들은 'LSTM', 'CNN', 'CNN+LSTM'등이 있다. 이 외에도 다양한 모델을 사용한 연구들이 있지만, 이번 프로젝트는 공통적으로 사용하는 모델들로 진행됐다.

LSTM(Long Short-Term Memory)은 긴 Sequence 데이터에서 과거의 중요한 정보를 기억하고 현재의 정보와 연결해 학습하는 것이다. 해당 프로젝트에서는 강조를 찾는 것이 핵심이다. 따라서 과거 정보만 보기 보다, 역방향(미래 시점의 정보)를 볼 수 있는 Bi-LSTM (Bidirectional-LSTM)을 선정했다.

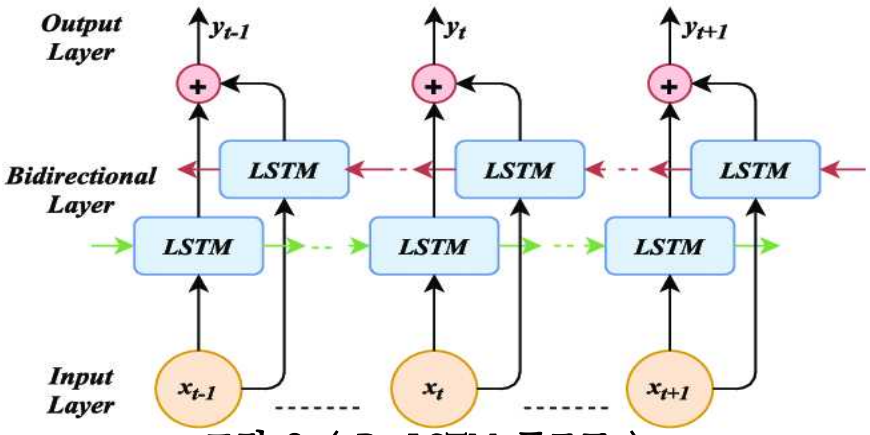



그림 2 〈 Bi-LSTM 구조도 〉

CNN(Convolutional Neural Network)은 필터(Kernel)를 통해 오디오의 지역적 패턴(Local Feature) 추출하고, 급격한 음량 변화나 톤 변화와 같은 순간적 특징 포착에 강점이 있어 선정했다.

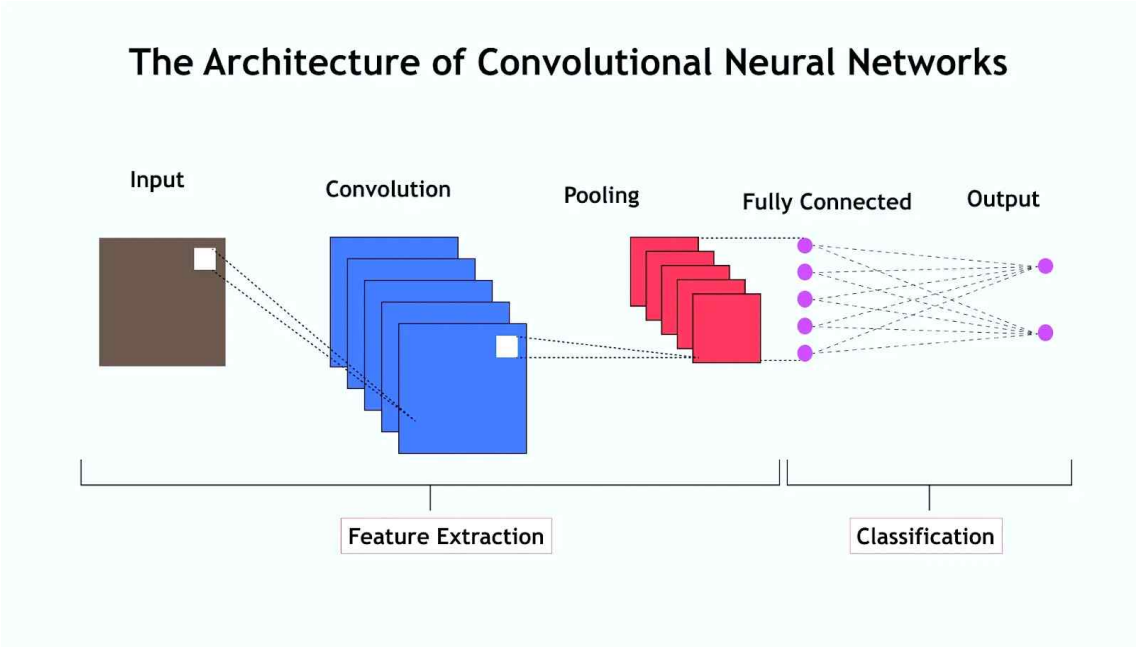



그림 3 〈 CNN 구조도 〉

CNN+LSTM은 앞서 서술한 두 모델을 같이 사용한다. 앞단에서는 CNN, 뒷단에서 LSTM을 이용한다. CNN의 Pooling 과정에서 중요 특징만 남기므로 단독 LSTM보다 Noise에 강점을 가지므로 선정했다.

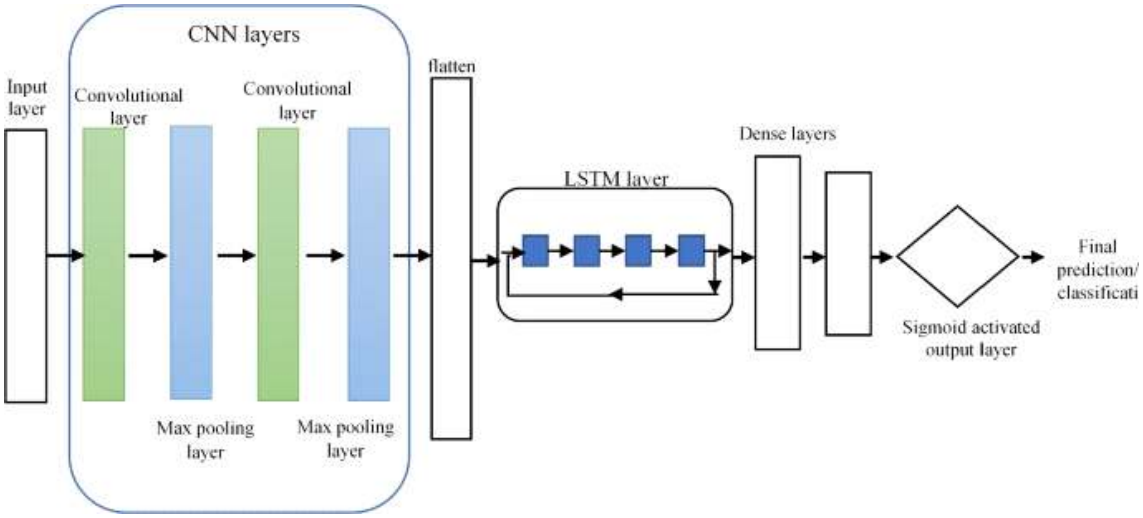



그림 4 〈 CNN + LSTM 구조도 〉

논문마다 성능이 좋은 모델이 다르고, 데이터에 따른 전처리 방법이 결과에 많은 영향을 미친다. 따라서 음성 특징 2가지, 모델 3개의 경우의 수 6개를 모두 학습시켜 성능을 비교했다. 결과는 그림 5,6의 그래프와 같다.


그림 5 〈 음성 특징이 MFCC일 때 모델들의 성능지표 〉

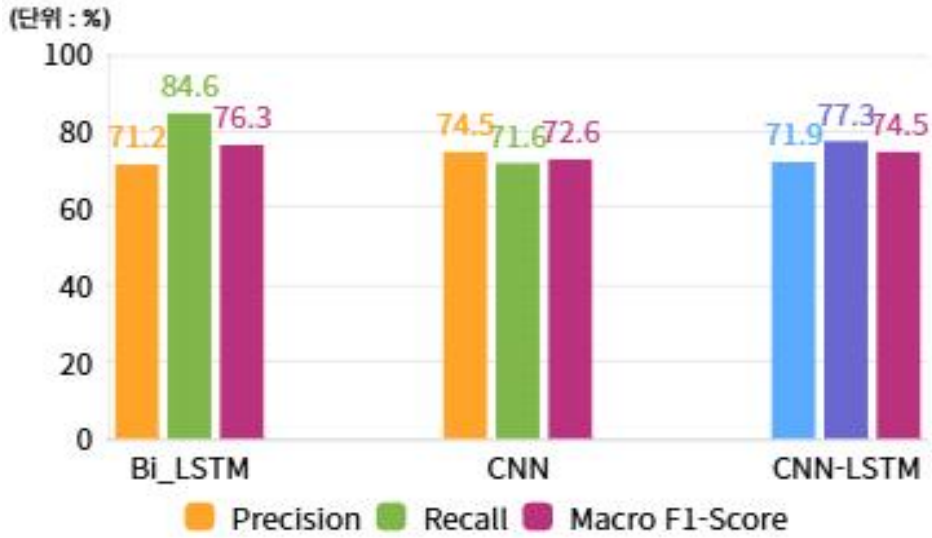



**그림 6 〈 음성특징이 Mel-Spectrogram일 때 모델들의 성능지표 〉**

본 프로젝트는 강조인 부분을 놓치지 않고 강조로 잘 파악하는 것이 중요하다. 따라서 Recall 값과 Macro F1-Score가 각각 84.6%, 76.3%로 가장 높은 Mel-Spectrogram + Bi-LSTM을 최종 모델로 선정했다.

### 2.2.2.3 음성 모델 파이프라인

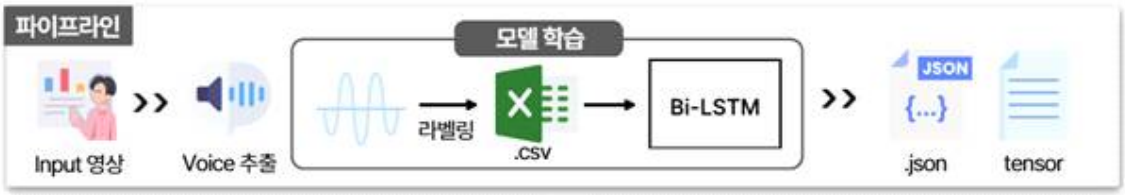



**그림 7 〈 음성 모델 파이프라인 〉**

먼저, Input으로 들어오는 비디오에서 오디오만 추출한다. 여기서 사람의 음성만 분리해낸 뒤, Librosa 라이브러리를 이용하여 Mel-Spectrogram을 뽑아낸다. 이때, 사람마다 목소리의 높낮이나 볼륨이 모두 다르기 때문에 정규화 과정을 진행한다. 이 데이터는 라벨링된 영상으로 학습된 Bi-LSTM의 Input으로 들어가서, 멀티모달 모델에 들어가기 위한 특징 벡터(Tensor)와 강조 시작 - 끝 시간 정보가 담긴 Json 파일로 Output을 내게 된다.

### 2.2.3 행동 분석 모델

본 프로젝트는 발표 영상 분석을 통한 구체적인 피드백을 제공하여 사용자의 스피치 역량을 강화하는 것을 목표로 한다. 이를 위해 강조 표현 시 발생하는 맥락 정보와 음성 신호, 손 제스처에 관한 기존 연구 사례를 분석하여 서비스 구현에 필요한 기술적 기반을 확립하고자 한다.

손 제스처 인식은 정지 이미지 기반의 객체 인식과 달리, 시간적 흐름과 공간적 움직임을 동시에 해석해야 하는 복합적인 과제이다. 기존의 Convolutional Neural Network(이하 CNN)와 Recurrent Neural Network(이하 RNN)를 결합한 구조는 동작의 연속적인 흐름을 온전히 포착하는 데 한계가 있다.

이러한 이유로 제스처 인식 분야에서는 시공간 정보를 통합 처리할 수 있는 3D-CNN이 핵심 기술로 자리 잡았다. 3D-CNN은 프레임 간 시간 축까지 함께 학습하여 짧은 동작이나 순간적인 제스처 변화를 인식하는 데 매우 효과적인 것이 입증됐다. (이진원(2017). 3D-CNN을 이용한 효율적인 손 제스처 인식 방법에 관한 연구. 서울대학교 대학원, 전기·컴퓨터공학부 공학석사 학위논문, p9-73.)

해당 논문은 3D-CNN의 강점에도 불구하고 발생하는 높은 연산량, 방대한 파라미터 수, 실시간 처리의 한계를 해결하는 데 초점을 둔다. 특히 영상 데이터 기반의 시스템은 메모리 비용이 크기 때문에, 자원이 제한된 임베디드 및 모바일 환경에서도 원활히 구동될 수 있도록 다음과 같은 세 가지 구조적 개선 방안을 제안한다.

1) 입력 영상으로 차영상을 이용하는 방법
입력 영상으로 두 개의 서로 다른 영상을 픽셀별로 빼서(감산하여) 두 영상 간의 차이를 시각적으로 나타낸 영상인 차영상(差映像, Difference Image)을 활용하여 배경 정보를 배제하고 동작 정보만을 강조한다. 이는 신경망의 불필요한 연산 부하를 줄여 학습 효율성을 극대화하는 전략이다.

2) 3차원 분해 인셉션 구조
높은 성능을 검증받은 인셉션 모듈을 3D로 확장하되, 이를 작은 필터 단위로 분해하여 계산량을 최적화한다. 연산의 90% 이상이 발생하는 컨볼루션 층을 병렬 구조로 재구성함으로써 파라미터를 대폭 절감해 연산 비용 감소와 메모리 절감 결과를 이끌어낸다.

3) 3차원 글로벌 평균 풀링
파라미터가 집중되는 전결합층을 GAP으로 대체해 파라미터 수를 감소시킨다. 시공간 축의 평균 특징을 요약하여 모델의 메모리 사용량을 줄이고 과적합 위험을 방지해 제스처 인식에 필요한 특징만 남긴다.

이와 같이 연산 효율과 정확도를 동시에 확보한 해당 연구의 성과는 발표자의 제스처를 분석하여 강조 구간을 탐지하고자 하는 본 프로젝트의 핵심 알고리즘 설계에 유용한 지표가 된다.

본 프로젝트에서는 발표자의 손 제스처를 정확하게 탐지하고 분석하기 위해 유튜브에서 공개된 강의 및 강연 영상을 활용하여 자체 데이터셋을 구축하였다. 수집된 원본 영상에 대한 데이터셋 구성 절차는 다음과 같다.

1) 영상 선별 및 수집
프로젝트 목적에 부합하도록 발표 상황이 명확하고 발표자의 제스처가 효과적으로 드러나는 영상을 선별하여 수집하였다. 특히 제스처의 미묘한 변화를 일관성 있게 관찰할 수 있도록 발표자 1인이 화면 중앙에 고정되어 있는 영상을 주요 수집 기준으로 설정하였다. 총 100개의 영상을 사용했으며 주요 출처는 다음과 같다.

- 경제 뉴스 채널 '매일경제 굿모닝 외신 스케치, 매일경제 간밤 미국은, 매일경제 머니그램'
- 교육 및 강연 채널 '메가스터디 및 대성마이맥 강사 커리큘럼, 최태성 어쩌다 어른, 스티브 잡스 2007년 PT'

2) 영상 프레임 기반 제스처 구간 라벨링

모델 학습을 위한 GT(ground truth) 데이터 생성을 위해, 영상 프레임 기반으로 행동(제스처) 구간 단위 라벨링을 수행했다.

- 구간 정의: 원본 영상 재생 후, 발표자의 제스처가 명확하게 나타나는 구간을 시작 시간 – 종료 시간 단위로 기록한다.

- 각 구간에 대해 다음과 같은 이진 라벨을 부여

- 0: 제스처가 발생하지 않은 구간
- 1: 명확한 움직임이 관찰되는 강조 제스처 구간

- 라벨링 도구

Python + OpenCV 기반의 구간 단위 마킹 툴(press space to start/end labeling)을 활용하였다.

| video_id | video_path | start_sec | end_sec | label |
|---|---|---|---|---|
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 21.788 | 22.422 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 55.155 | 55.956 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 69.469 | 69.77 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 74.374 | 75.008 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 82.883 | 83.216 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 88.822 | 89.289 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 102.102 | 102.402 | 1 |
| v001 | /home/piai/바탕화면/영상raw/100.mp4 | 109.643 | 110.077 | 1 |

이 과정을 거쳐 각 영상별로 라벨링 결과가 .csv파일이 생성되며, 해당 파일은 위와 같은 구조를 가진다.

3) 학습용 클립 단위 데이터 생성

본 프로젝트에서 활용하는 학습 모델은 3D CNN 기반 구조이므로, 입력 데이터는 C 채널 수, T 시간 길이, H 높이, W 너비로 구성된 3D 텐서 형식을 요구한다.

이에 따라 수집된 각 원본 영상은 다음과 같은 방식으로 슬라이딩 윈도우 기법을 활용하여 학습용 클립 샘플로 변환된다.

- 클립 길이(T, clip_len): 16 프레임
- 슬라이드 간격(stride): 8프레임씩 이동하며 클립 생성
- 라벨링: 생성된 각 클립은 이전 단계에서 정의된 csv 파일의 강조 제스처 구간과 시간축을 매칭하여 이진 라벨을 부여
- 최종 입력 텐서 규격: (3, 16, 112, 112) 형태의 텐서로 모델 입력

이러한 슬라이딩 윈도우 기반 클립 추출 방식은 비디오 기반 행동 인식(Action Recognition) 분야에서 시계열적 특징을 효과적으로 학습시키기 위해 널리 사용되는 표준적인 데이터셋 구성 방법론이다.

제스처 강조 구간 탐지 문제는 짧은 시간 내 발생하는 동작의 변화와 연속적인 움직임을 정밀하게 포착해야 하는 시계열 분석의 성격을 가진다. 이에 따라 본 프로젝트에서는 시공간 정보를 동시에 처리할 수 있는 3D-CNN 구조를 핵심 모델로 선정했다.

선행 연구인 「3D-CNN을 이용한 효율적인 손 제스처 인식 방법에 대한 연구」에서는 제스처 인식과 같은 행동 분석 문제에서 시간적 흐름을 포함한 특징 학습의 중요성을 강조하며, 프레임 단위의 2D-CNN보다 3D-CNN이 짧고 빠른 동작을 인식하는 데 효과적임을 입증하였다. 이는 순간적으로 나타나는 제스처 강조 구간을 탐지하고자 하는 본 프로젝트의 목적과 직접적으로 부합한다.

발표 환경에서의 강조는 정적인 자세보다는 짧은 시간 동안 발생하는 빠른 손동작, 상체의 움직임, 동작의 리듬 변화를 통해 이루어지는 경우가 많다. 따라서 공간 정보(H, W)뿐만 아니라 시간 축(T)을 함께 고려하는 3D-CNN 모델 구조가 필수적이며, 이러한 요구 조건을 충족하는 3D-CNN은 본 프로젝트의 핵심 모델로 가장 적합하다.

대표적인 3D-CNN 계열 모델인 **mc3_18, r3d_18, r2plus1d_18**을 후보 모델로 선정하여 성능 비교 실험을 수행했다.

각 모델의 특징은 다음과 같다.

- **mc3_18**

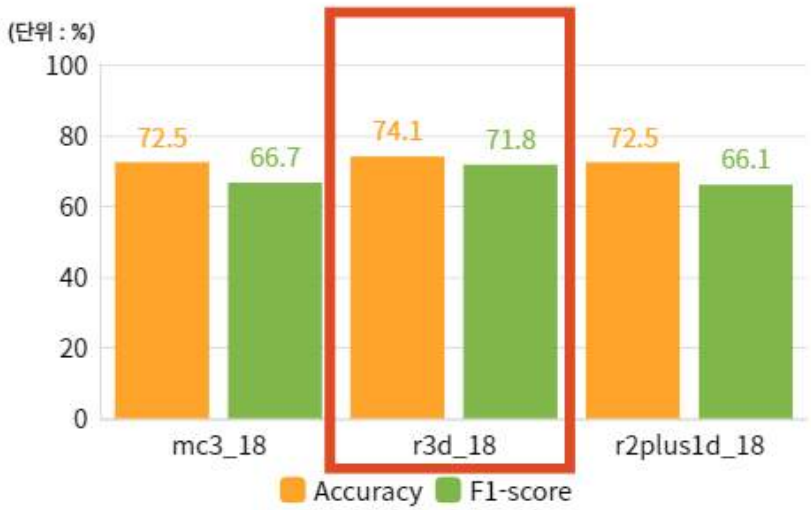
프레임별 자세나 손 형태를 정리하는 2D Convolution과 움직임 여부에 대한 시간 정보를 감지하는 3D Convolution을 혼합화한 경량화 구조로, 공간과 시간을 섞어서 본다.

- **r3d_18**

모든 계층에 3D Convolution을 직접 적용하여 시공간 특징을 통합적으로 학습하며, 강조 제스처 움직임의 변화를 학습한다.

- **r2plus1d_18**

3D Convolution을 2D 공간 필터와 1D 시간 필터로 분리 학습하는 구조를 채택한다. 이를 통해 높은 특징 표현력을 제공하며 복잡한 행동 분석 문제에 효과적이다.




▲ 행동 분석 모델 성능 비교 결과

실험 결과, r3d_18은 다른 후보 모델 대비 F1-score와 Accuracy 지표에서 가장 높은 성능을 나타냈다. r3d_18은 모델 복잡도 대비 안정적인 성능과 더불어, 발표자 제스처 강조 구간 탐지라는 실시간 응용 환경에서도 효율적으로 적용 가능하다고 판단하였다. 이에 따라 본 프로젝트에서는 3D-CNN r3d_18을 최종 제스처 강조 탐지 모델로 선정하였다.

선행 연구에서 제안된 경량화된 3D-CNN 구조의 장점을 활용해, 발표자의 미세하고 빠른 손동작을 실시간에 가깝게 포착하는 것을 목표로 한다. r3d_18의 시공간 특징 추출 능력을 기반으로, 정적인 프레임 분석이 아닌 동작의 연속성가 변화 양상을 고려한 제스처 강조 탐지 시스템을 구현하고자 한다.

발표 상황에서의 제스처 강조 구간을 안정적으로 탐지하기 위해 3D-CNN 기반 비전 모델을 활용하였다. 그러나 3D-CNN 계열 모델은 구조뿐만 아니라 학습 전략과 하이퍼파라미터 설정에 따라 성능 편차가 크게 발생하는 특정을 지닌다. 이에 본 프로젝트에서는 상단에 선택된 모델 구조를 기준으로 다음 세 가지 전략에 따른 성능 차이를 비교·분석하였다.

1) Optimizer 비교: AdamW vs SGD
3D-CNN r3d_18모델에 대해 AdamW와 SGD Optimizer을 동일한 학습 조건 하에서 비교 실험을 수행하였다.

SGD는 학습 초기부터 Train Loss를 빠르게 감소시키며 손실 최적화 측면에서 우수한 수렴 특성을 보였으나, Validation Accuracy 기준으로는 AdamW가 전 Epoch에 걸쳐 일관되게 높은 성능을 기록하였다. 최종 Epoch 기준으로 AdamW는 약 70%의 검증 정확도를 보인 반면, SGD는 약 64% 수준에 머물렀다.

이는 SGD가 학습 데이터에 강하게 적합되는 경향을 보이는 반면에 AdamW는 Adaptive Learning Rate와 Weight Decay 분리를 통해 일반화 성능 측면에서 더 유리한 특징을 가짐을 알 수 있다.

본 프로젝트는 상대적으로 데이터 규모가 제한된 환경에서 수행되므로, 과적합을 완화하고 검증 성능을 안정적으로 유지할 수 있는 AdamW를 채택했다.

2) Pretrained 모델 활용 여부

3D-CNN 계열 모델은 시공간 특징을 동시에 학습하기 때문에 대규모 데이터셋을 필요로 하는 경향이 있다. 이에 Kinetics-400/600과 같은 대규모 비디오 데이터셋으로 사전 학습 된 3D- CNN r3d_18 모델을 활용하여 파인 튜닝을 수행했다.

기존 연구에 따르면, 사전학습된 3D-CNN을 소규모 데이터셋에 적용할 경우 랜덤 초기화 모델 대비 과적합이 완화되고 성능이 향상되는 것으로 보고된다.

실험 결과, 동일한 구조의 모델을 랜덤 초기화로 학습한 경우보다, 사전 학습된 가중치를 기반으로 파인 튜닝한 모델이 보다 안정적인 학습 곡선과 높은 검증 정확도를 보인다.

이에 사전 학습된 3D-CNN r3d_18을 기반으로 발표자 제스처 강조 구간 탐지 작업에 적합하도록 미세 조정을 수행하는 전략을 채택하여 모델의 일반화 성능을 확보했다.

3) 입력 클립 프레임 수 비교

입력 클립의 시간적 길이는 3D-CNN 기반 모델의 성능에 중요한 영향을 미치는 요소이다. 본 연구에서는 입력 클립의 시간적 길이를 16 프레임과 32 프레임으로 설정하여 성능 차이를 비교하였다.

일반적으로 32 프레임과 같이 더 긴 시간 정보를 포함하는 클립은 행동 인식 정확도가 향상되는 경향이 있으나, 연산량과 추론 지연 시간이 증가하여 실시간 적용에 불리해지는 단점이 존재한다.

본 프로젝트는 발표 피드백 시스템이라는 응용 특성을 고려하여, 실시간 처리 가능성을 유지하면서도 제스처의 순간적인 변화를 충분히 포착할 수 있는 프레임 길이를 최종 입력 설정으로 선택하였다.

위 비교 실험 결과를 종합하여, 본 프로젝트에서는 다음과 같은 학습 전략을 최종적으로 적용하였다.

- Backbone 모델: 3D-CNN r3d_18
- 초기 방식: 대규모 비디오 데이터셋 기반 사전 학습 가중치 사용
- Optimizer: 검증 정확도 우위를 보인 AdamW 채택
- 입력 클립 구성: 실시간 응용 환경의 효율성 균형을 고려하여 16프레임의 슬라이딩 인도우 기반 3D 클립 구성

그 결과, 제한된 데이터 환경에서도 발표자의 제스처 강조 구간을 효과적으로 탐지할 수 있는 안정적인 성능을 확보할 수 있었다.

### 2.2.4 멀티 모달 구성

본 프로젝트의 최종 목표는 발표 영상에서 **문맥상 강조가 필요한 구간**(Text)과 실제로 발화에서 나타난 **음성 강조 특성**(Voice), 그리고 발표자의 **제스처·행동 강조**(Vision)를 통합해 사용자가 개선해야 할 발표상의 포인트를 구간 단위로 제시하는 것이다. 이를 위해 텍스트·음성·영상 세 모달리티에서 산출된 특징을 하나의 모델로 결합하는 멀티모달 융합 구조를 설계하였다.

우선 각 모달리티는 동일한 시간 축을 기준으로 정렬된다. 텍스트 모달리티는 STT 기반으로 문장 단위 구간(start-end timestamp)을 생성하고, 문맥 분석 모델이 문장별 강조 점수(emphasis score)와 강조 이유(reason)를 출력한다. 음성 모달리티는 동일 구간에서 Mel-Spectrogram 기반 특징 벡터(Tensor)와 음성 강조 예측 결과를 추출한다. 비전 모달리티는 슬라이딩 윈도우 기반 3D-CNN(r3d_18)을 통해 구간별 제스처 발생 확률 및 행동 특징 벡터를 산출한다. 이렇게 생성된 텍스트/음성/비전의 결과는 구간 단위로 매칭되어 하나의 통합 입력으로 구성된다.

융합 단계에서는 모달리티 간 기여도가 상황에 따라 달라질 수 있다는 점과, 모델 성능을 고려해 Gated Fusion 기반 결합 방식을 적용했다.

Gated Fusion은 각 모달리티 특징에 대해 가중치(게이트)를 학습하여, 특정 구간에서 더 중요한 신호(예: 강한 제스처, 음성 강세 변화, 문맥상 핵심 문장)를 상대적으로 강조하는 방식이다. 이를 통해 텍스트만으로는 포착하기 어려운 실제 전달 방식의 차이를 반영할 수 있으며, 반대로 음성·제스처가 과도하더라도 문맥상 중요도가 낮은 구간은 강조 포인트로 과대 판단되지 않도록 균형을 맞출 수 있다.

최종적으로 멀티모달 모델은 구간별로 (1)최종 강조 점수, (2)강조 누락/과잉 여부, (3)모달리티 별 기여도를 산출한다. 결과는 사용자 피드백 생성 단계에서 활용 가능하도록 JSON 형태로 저장하며, 후속 분석 및 시각화를 위해 구간별 특징 벡터(Tensor)도 함께 출력한다.

이를 통해 “문맥상 강조가 필요한데 목소리·제스처가 약한 구간”, “문맥 중요도는 낮지만 과도한 제스처가 나타난 구간” 등과 같이 실제 발표 코칭에 필요한 형태의 진단을 제공할 수 있다.

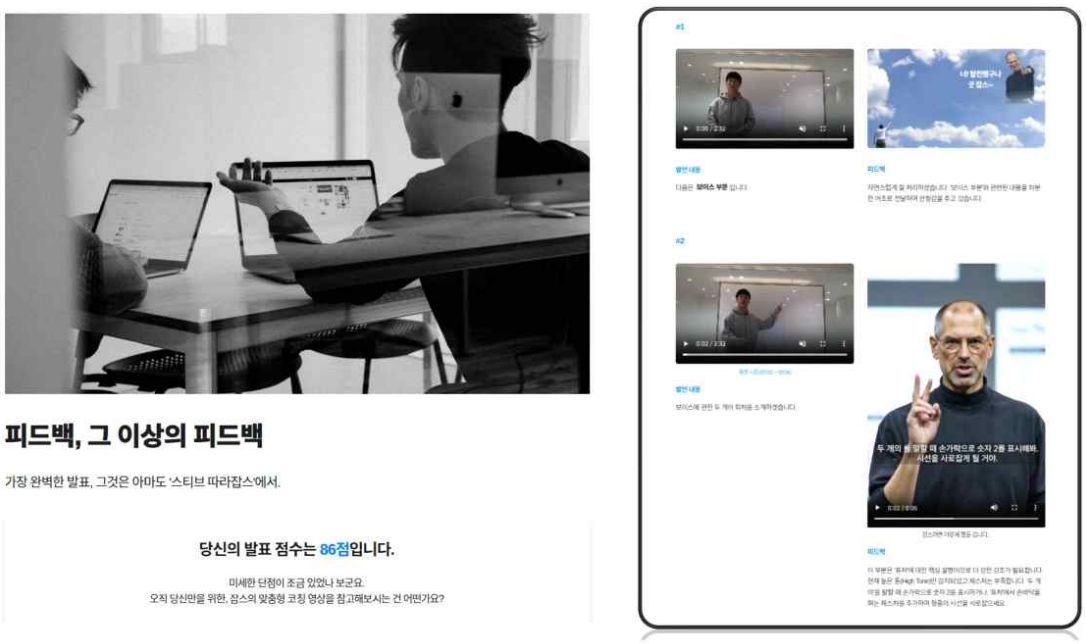
# Ⅲ. 결론

## 3.1. 최종 결과

발표 영상 입력을 기반으로 텍스트·음성·행동 분석 결과를 통합해 사용자 발표를 구간 단위로 평가하고, 이에 대한 정량 점수 및 정성 피드백을 생성하기 위한 멀티모달 발표 분석 시스템을 최종 구현했다. 각 분석 단계에서 산출된 JSON 데이터를 통합해, 사용자의 발표 전달력을 종합적으로 판단할 수 있도록 설계됐다.

멀티모달 결합 단계에서는 세 모달리티에서 생성된 JSON 결과를 기반으로, 구간별 텍스트 중요도, 음성 강조 특성, 제스처 강조 여부를 통합 분석했다. Gated Fusion 기반 결합 방식을 적용함으로써, 구간마다 서로 다른 모달리티의 기여도를 반영하여 최종 강조 점수 및 발표 수행 점수를 도출하였다. 이를 통해 단일 모달 분석에서는 판단이 어려운 발표 표현의 불균형(예: 중요 문장이지만 표현이 약한 구간, 과도한 제스처가 사용된 구간)을 정량적으로 식별할 수 있었다.

최종적으로 시스템은 사용자 발표를 구간 단위로 분해하여 (1) 종합 발표 점수, (2) 구간별 강조 적합도, (3) 문제 유형(강조 부족·과잉·불균형)을 포함한 통합 JSON 결과를 출력한다. 해당 결과는 후속 피드백 생성 단계에서 활용 가능하도록 구조화됐으며, Gemini 기반 생성 모델에 입력되어 구간별 맞춤형 피드백 문장을 자동 생성하는 데 사용된다. 이를 통해 사용자는 “어떤 구간에서, 무엇이 문제였고, 어떻게 개선해야 하는지”를 직관적으로 이해할 수 있는 발표 피드백을 제공받을 수 있다.




▲ 기술 최종 구현 사이트 일부

결과적으로 본 프로젝트는 멀티모달 분석 결과를 실제 코칭 문장으로 연결하는 발표 피드백 파이프라인을 완성했다는 점에서 의미가 있다. 이는 발표자의 표현 방식 전반을 데이터 기반으로 진단하고, 실질적인 개선 행동으로 이어질 수 있는 AI 발표 컨설턴트의 형태를 구현한 것이다.

## 3.2. 기대 효과 및 활용 방안

### 3.2.1. 기대 효과

본 프로젝트의 가장 큰 기대 효과는 텍스트·음성·비전 간의 '정합성(Alignment)' 분석을 통해 발표 전달력의 질적 문제를 구체적으로 진단할 수 있다는 점이다. 기존 발표 평가가 개별 요소의 단순 점검에 머물렀다면, 본 시스템은 문맥상 강조가 필요한 구간과 실제 표현 방식 간의 불일치를 정량적으로 식별한다.

이를 통해 "중요한 문장이지만 목소리가 약한 구간", "문맥상 강조도가 낮음에도 과도한 제스처가 사용된 구간"과 같이 발표 실패의 원인을 구조적으로 설명할 수 있는 피드백 제공이 가능하다. 이는 단순한 잘잘못 평가가 아닌, 발표 전달 구조 전반을 개선하는 데 직접적인 도움을 준다.

또한 본 프로젝트는 시간·공간·비용 제약으로부터 자유로운 발표 피드백 환경을 제공한다. 사용자는 별도의 대면 코칭이나 고가의 스피치 학원에 의존하지 않고, 발표 영상 업로드만으로 언제든지 분석 결과와 피드백을 받을 수 있다. 이는 고비용 중심의 기존 발표 교육 방식에 대한 실질적인 대안으로 작용할 수 있다.

마지막으로, 데이터 기반 피드백을 통해 사용자의 '발표 메타 인지(meta-cognition)'를 강화할 수 있다. 사용자는 자신의 목소리, 제스처, 발화 습관을 주관적 감각이 아닌 정량 지표로 인식함으로써, 반복 학습 과정에서 스스로 문제를 인식하고 개선하는 능력을 기를 수 있다.

### 3.2.2 활용 방안

첫째, 취업 준비생 및 대학생 발표 훈련 도구로 활용 가능하다. PT 면접, 자기소개 발표, 전공 발표 영상 등을 분석하여 강조 부족·과잉 구간을 파악하고, 구간별 개선 피드백을 반복적으로 제공하는 개인 학습 도구로 활용할 수 있다.

둘째, 기업 내 프레젠테이션·보고 역량 강화 도구로 적용할 수 있다. 임직원의 보고 영상이나 발표 자료를 기반으로 전달력 문제를 진단하고, 구성원별 발표 스타일 차이를 데이터로 비교·관리하는 교육 보조 시스템으로 확장 가능하다.

셋째, 온라인 강의·강연 콘텐츠 분석을 통해 콘텐츠 전달 품질 관리 도구로 활용할 수 있다. 강사의 강조 구간, 음성 안정성, 제스처 사용 패턴을 분석하여 강의 몰입도 향상을 위한 피드백 자료로 활용 가능하다.

## 3.3. 개선 방안

다음과 같은 확장 및 개선 여지가 존재한다.

첫째, 언어 다양성 확장 가능성이다. 현재는 한국어 발표 데이터에 특화되어 있으나, 언어별 발화 특성과 강조 방식이 상이하다는 점을 고려할 때 다국어 데이터셋 확장을 통해 글로벌 사용자 대상의 피드백 제공이 가능하다. 이는 언어별 억양, 강세, 강조 패턴을 반영한 고도화된 분석으로 이어질 수 있다.

둘째, 정량적 점수 체계의 구체화가 필요하다. 현재의 발표 점수는 모달리티별 결과를 종합한 상대적 지표 중심이므로, 향후 요소 간 가중치 설정 근거를 명확히 하고 발표 도메인 지식을 반영한 수식 기반 점수 체계로 고도화할 필요가 있다. 이를 통해 평가의 신뢰성과 비교 가능성을 강화할 수 있다.

셋째, 데이터셋 확보를 통한 기타 멀티모달 모델의 활용이다. 본 프로젝트에서는 데이터 규모 제약으로 Gated Fusion 방식을 채택했으나, 향후 데이터셋이 충분히 확보될 경우 Transformer 기반 멀티모달 구조와의 성능 비교 및 적용 가능성에 대한 추가 연구가 가능하다. 이를 통해 모델 선택 기준과 성능 차이를 보다 명확히 분석할 수 있을 것으로 기대된다.

# Ⅳ. 참고문헌


- 이진원 (2017). 3D-CNN을 이용한 효율적인 손 제스처 인식 방법. 서울대학교 전기·컴퓨터공학부 공학석사 학위논문, pp. 9-73.
- 권순복 (2015). 준언어적 구성 요소를 통한 매력적인 목소리 분석과 호감도에 관한 실험 연구.
- Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal Learning with Transformers: A Survey. arXiv preprint arXiv:2206.06488v2. (IEEE TPAMI accepted)
- Liu, R., Li, Y., Liang, D., Tao, L., Hu, S., & Zheng, H.-T. (2021). Towards Attention-Free Visual-Language Modeling with MLPs. arXiv preprint arXiv:2112.04453v1.
- 김은희 (2021). 딥러닝 기반 문장 중요도를 고려한 중심 문장 추출 방법. 조선대학교 산업기술창업대학원 소프트웨어융합공학과 석사학위논문.
- 최지현, 조동욱, 정연만 (2016). 음성 분석을 이용한 청자가 호감을 느끼는 목소리에 대한 규명. 한국통신학회논문지, 41(1), 16-25.
- Zeng, H., Wang, X., Wang, Y., Wu, A., Pong, T. C., & Qu, H. (2022). GestureLens: Visual Analysis of Gestures in Presentation Videos. IEEE Transactions on Visualization and Computer Graphics.