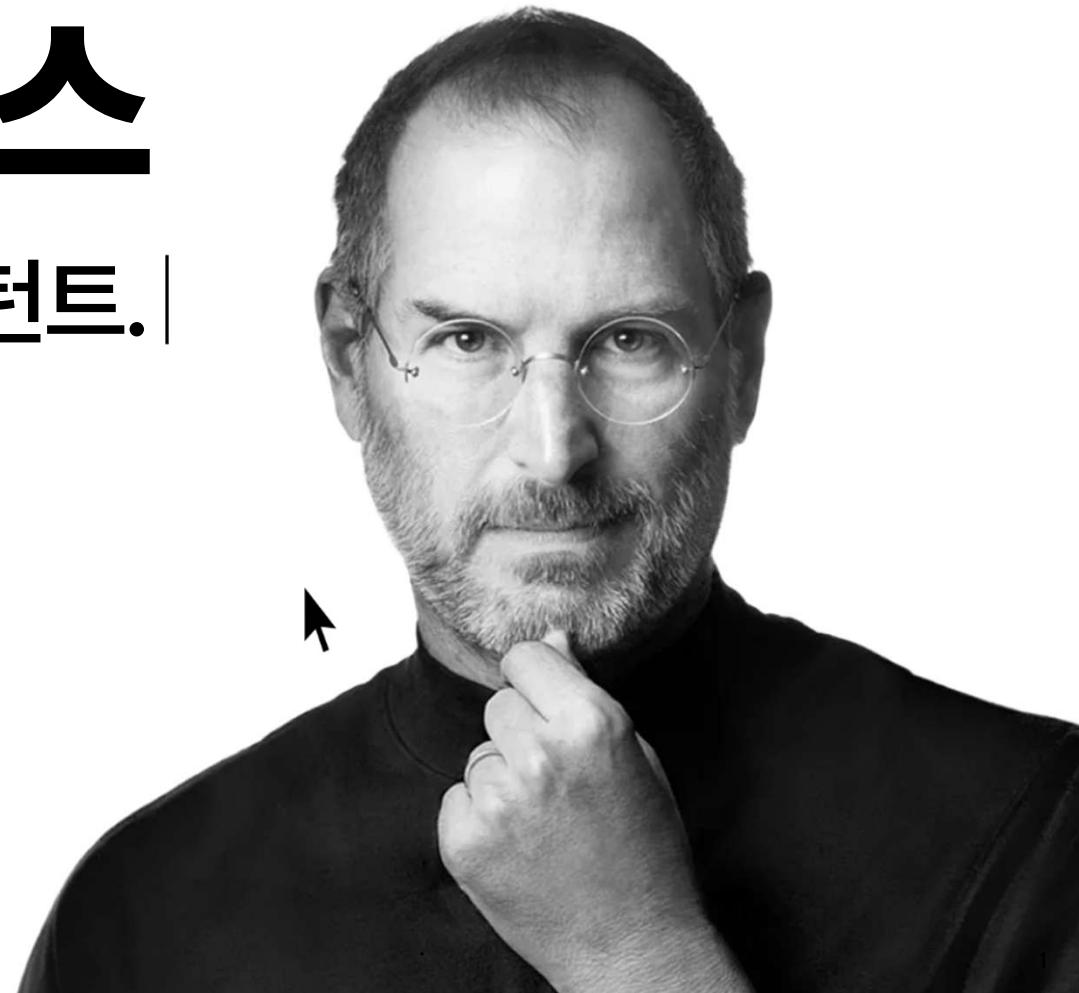


스티브 따라잡스

: 최고의 발표를 위한 AI 컨설턴트.



CONTENTS.

01 프로젝트 개요

- 추진 배경
- 기업 분석
- 전 기수 분석

02 프로젝트 소개

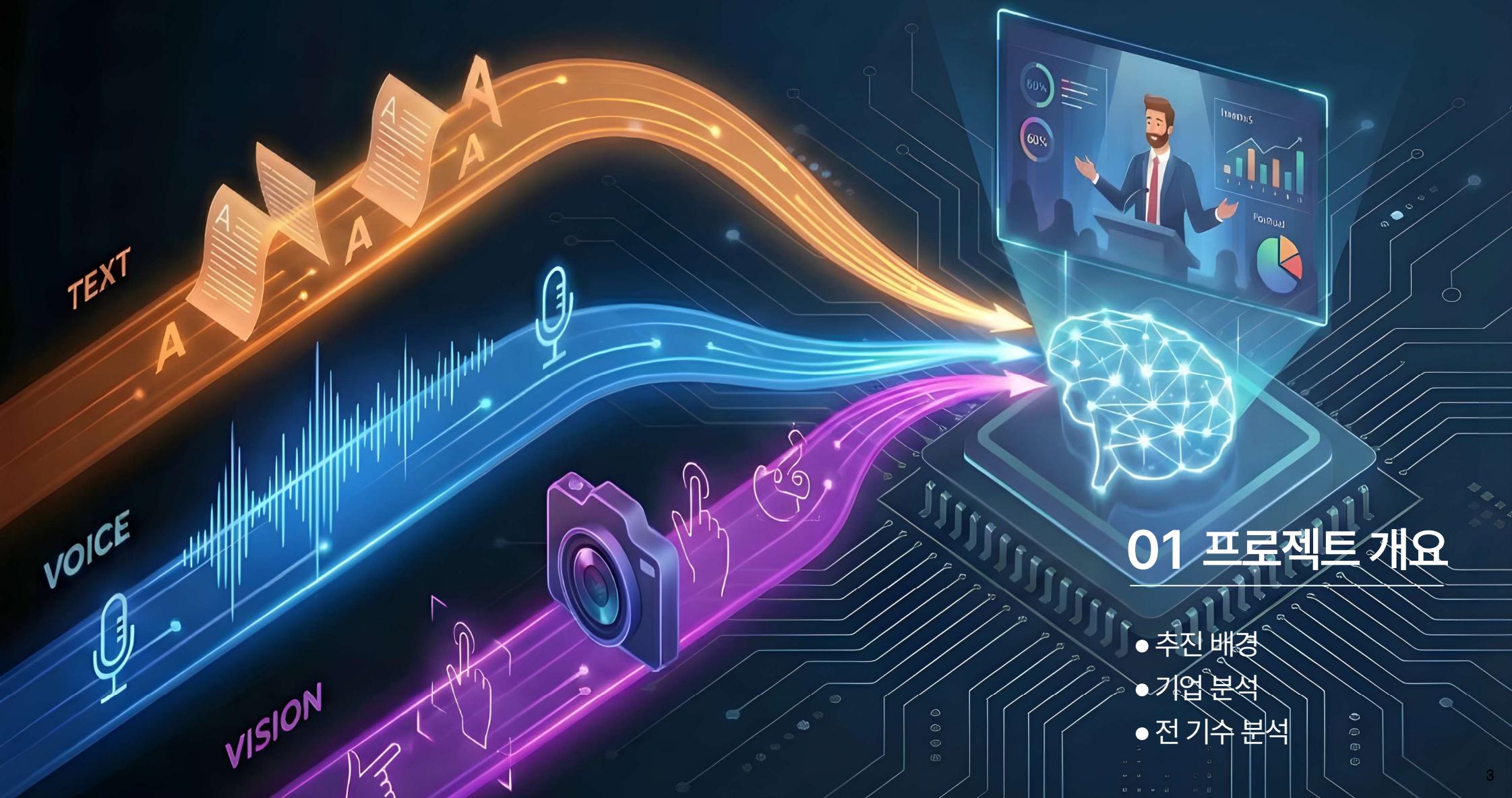
- Flow Chart

03 기술 구현

- Text
- Voice
- Vision
- 멀티모달
- 부가지표
- LLM
- 점수화

04 결론

- 시연 영상
- 최종 Output
- 기대 효과
- 한계점 및 개선기회
- Q&A



자기PR시대…‘스피치 교육’ 열풍

“예전에는 큰 목소리와 자신감을 키우려는 목적이 대부분...
지금은 실무형 소통 능력을 배우려는 직장인이 전체의 70% 이상”

“자녀 나이에 '0'붙이면 월 학원비”…사교육비 부담 '곡소리'

강남의 한 스피치학원은 1대1 수업 총 8회 (90분) 수강 시 160만 원이 든다고 안내했다.

“대화·전화 응대 어려워”…사교육 받는 2030 직장인들



현대 사회에서 **스피치 능력의 중요도는 계속 상승** 중이나,
발표 공포심·불안감으로 어려움을 겪는 이들이 다수





1-1 | 추진 배경



자기PR시대…‘스피치 교육’ 열풍

“예전에는 큰 목소리와 자신감을 키우려는 목적이 대부분...
지금은 실무형 소통 능력을 배우려는 직장인이 전체의 70% 이상”

“자녀 나이에 '0'붙이면 월 학원비”…사교육비 부담 '곡소리'

강남의 한 스피치학원은 1대1 수업 총 8회 (90분) 수강 시 160만 원이 든다고 안내했다.

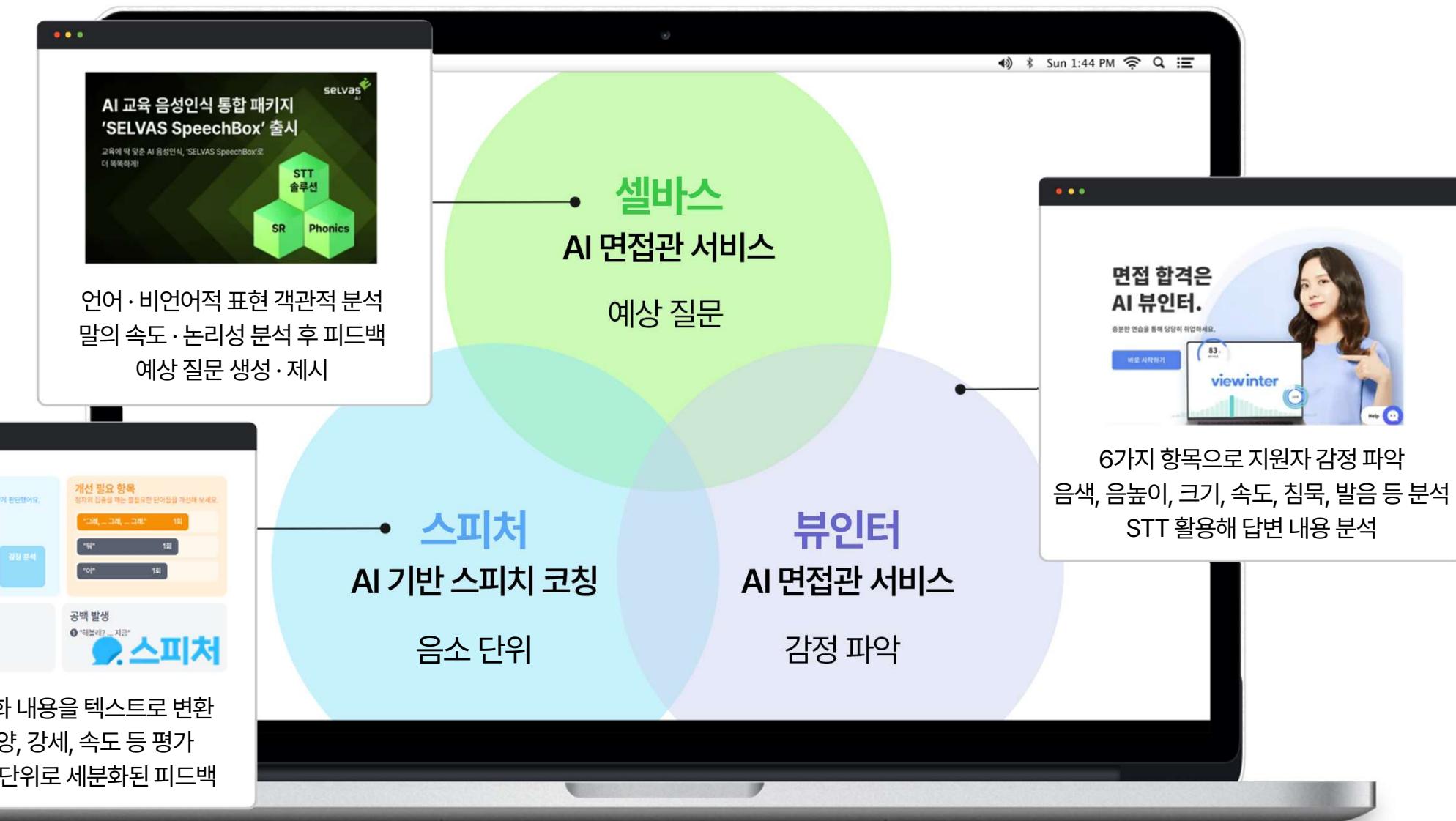


상세한 AI 피드백으로 시간·공간·비용의 제약으로부터 자유로운
'스티브 따라잡스'





1-2 | 기업 분석





1-2 | 기업 분석

Sun 1:44 PM

AI 교육 음성인식
'SELVAS Speech'

언어 · 비언어적
말의 속도 · 논리
예상 질문

주요 키워드
제작자에게 인증된 AI 모델
시가판
설정 조작

말의 빠르기
평균 속도
202 wpm

학습자 발화 내용을 텍스트로 변환
발음, 억양, 강세, 속도 등 평가
음소 · 음가 단위로 세분화된 피드백

현재 시스템은 면접 위주일 뿐이며,
발표에서 필수적인 강조 포인트 캐치가 불가합니다.

- 발화 내용 교정에만 중점 한계
- 제스처 · 시선 · 자세 분석 X 한계
- 어떤 타이밍에 어떻게 고쳐야 하는지 안내 X 한계
- 맥락 · 음성 · 영상을 종합한 피드백의 필요성 차별화

그만두기

기능 확장하기

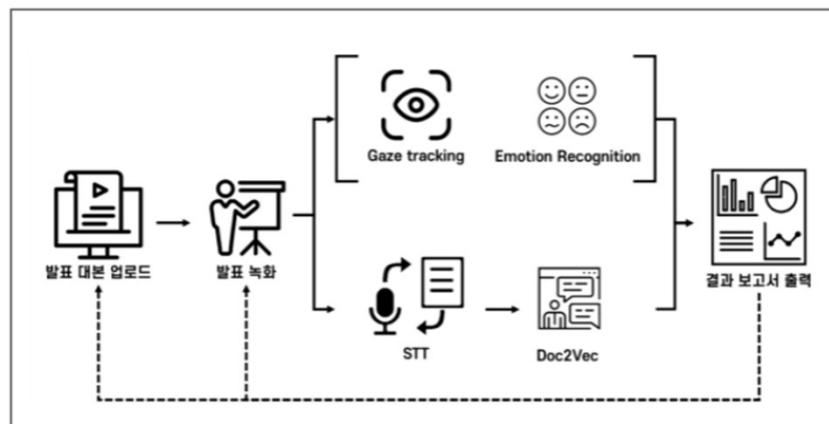
면접 합격은
AI 뷰인터.
종합적인 면접 분석
6가지 항목으로 지원자 감정 파악
음색, 음높이, 크기, 속도, 침묵, 발음 등 분석
STT 활용해 답변 내용 분석

1-3 | 전 기수 분석



[11기 A3조]

POresentation : 초심자를 위한 발표 도우미 AI



- Emotion Recognition: FaceNet
- Gaze tracking
- STT: Google Speech API, KoSpacing, KSS
- Doc2Vec: mecab

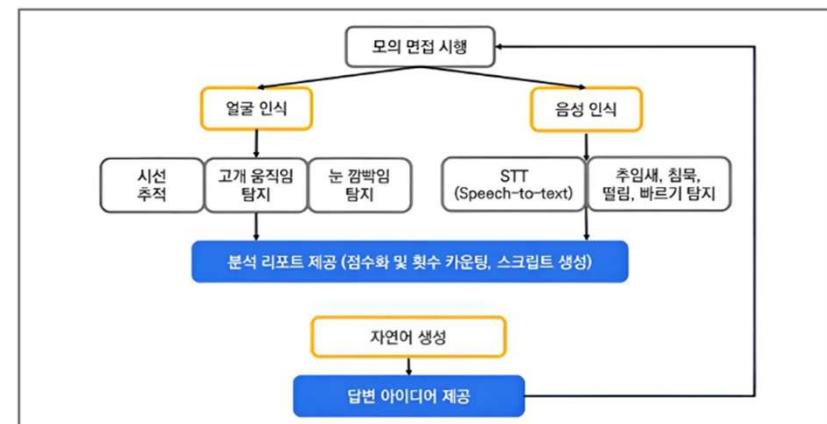
대본과 발화간 유사도 분석

누락된 부분 + 발표 습관(시선, 표정 평균값) 피드백



[15기 A2조]

PO-inter : 셀프 면접 연습 도우미



- Face Detection and Analysis : Open CV
- Filler Detection and Voice to Text : CNN, STT
- Generate Suggestion : KoGPT2-FineTuning

고개 움직임 / 눈 깜빡임 / 추임새 횟수 카운팅,
예시 면접 답변 생성 · 제시



1-3 | 전 기수 분석



[11기 A3조]

POresentation : 초심자를 위한 발표 도우미 AI

- Emotion Recognition: FaceNet
- Gaze tracking
- STT: Google Speech API, KoSpeech
- Doc2Vec: mecab

대본과 발화간 유사도 분석
누락된 부분 + 발표 습관(시선, 표정 평균값) 피드백

[15기 A2조]

PO-inter : 셀프 면접 연습 도우미

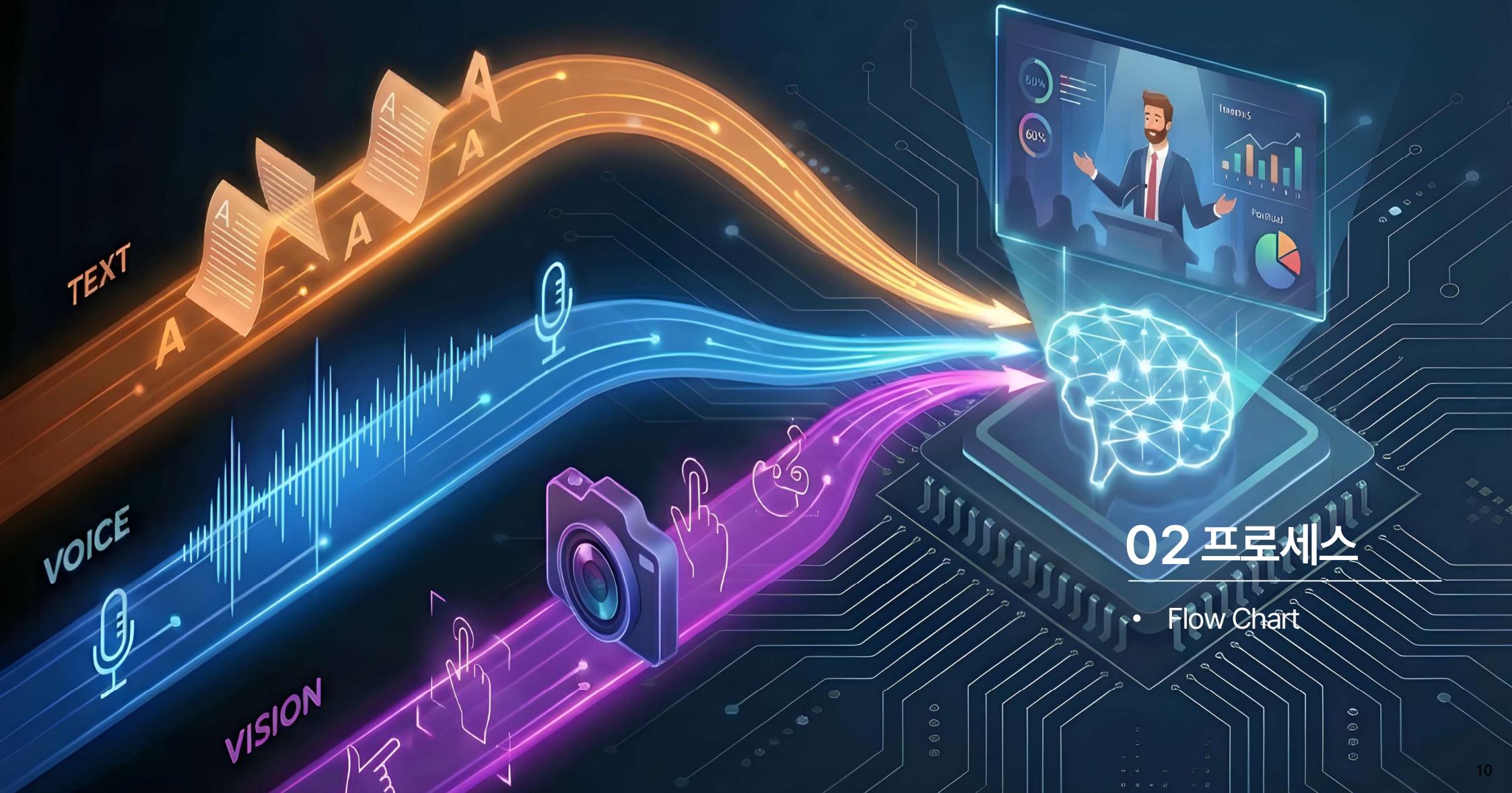
마찬가지로 강조 포인트,
불량한 자세, 음성 안정성 등에 대한
확장 가능성 확인

그만두기 기능 확장하기

고개 움직임 / 눈 깜빡임 / 추임새 횟수 카운팅,
예시 면접 답변 생성 · 제시

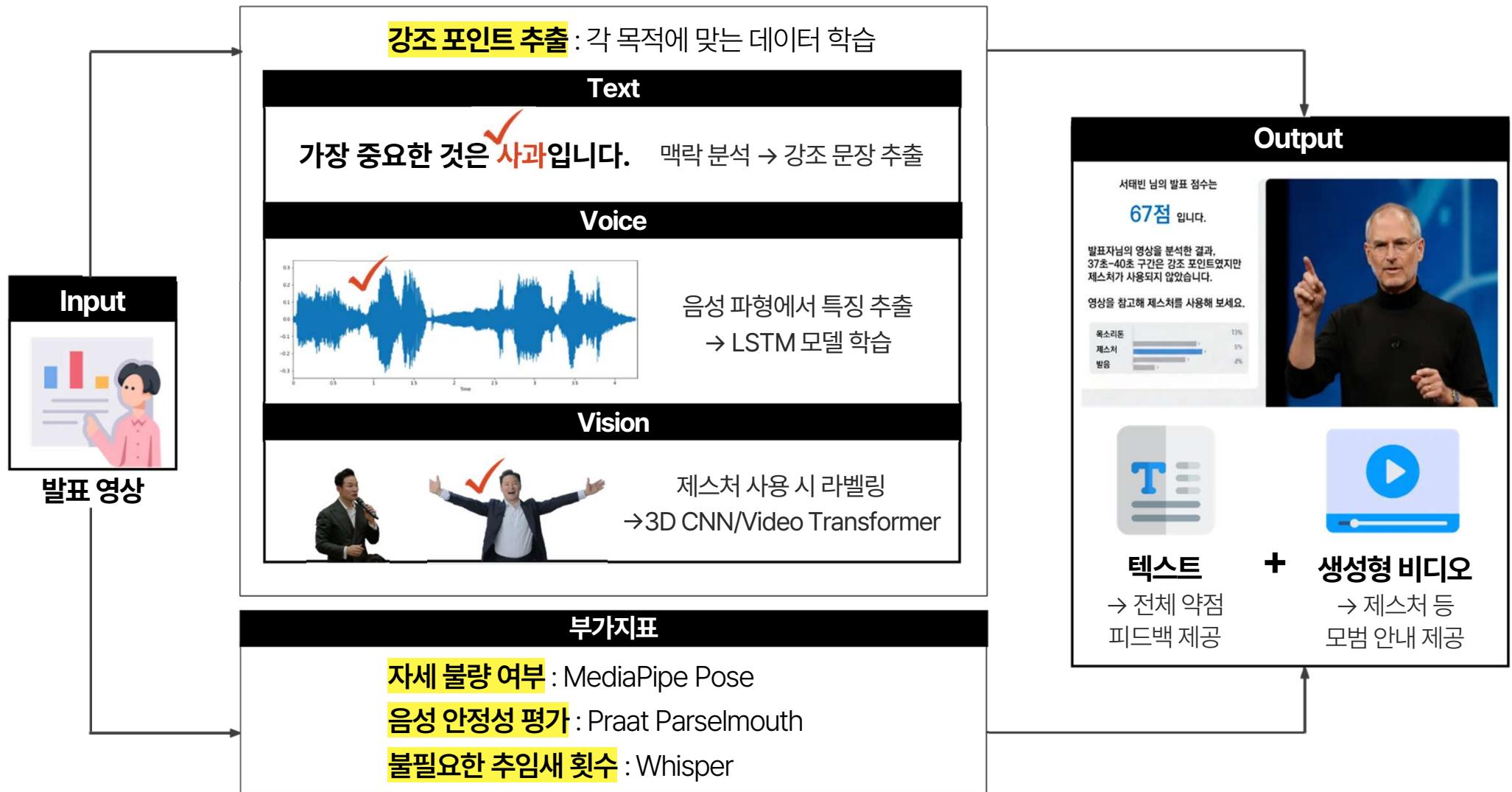
02 프로세스

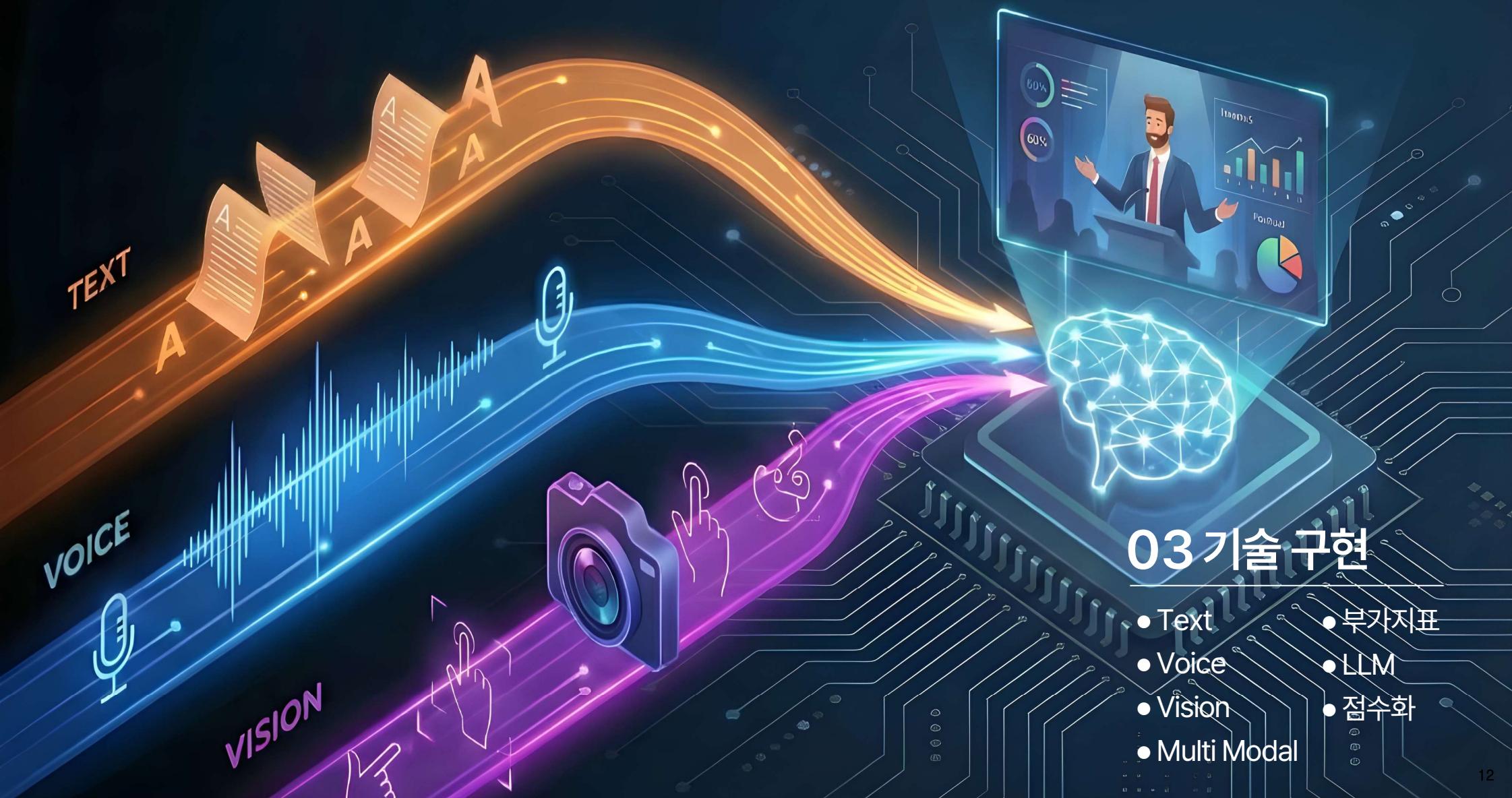
- Flow Chart





2-1 | Flow Chart





03 기술 구현

- Text
- Voice
- Vision
- Multi Modal
- 부가지표
- LLM
- 접수화



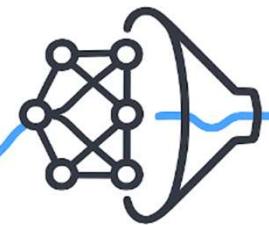
3-1 기술 소개 | Text 구조



데이터 정제



오디오 추출



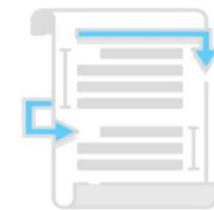
잡음 제거
(Deep Filter Net)



백색소음 주입
(Whisper 환각 방지)



전체 발화 텍스트화
(Whisper Large V3-Turbo)



문장 단위 파싱
(KSS)

Before



"original" : "지금까지 생생지구촌
이었습니다."



"original" : "그것만 해도 코스피의
전체 시총이 코스닥의
약 5.7배 수준이었습니다."



→ "original" : "11월 5일 수요일 스타트
브리핑 시작합니다"



→ "original" : "먼저 조선일보입니다"

After

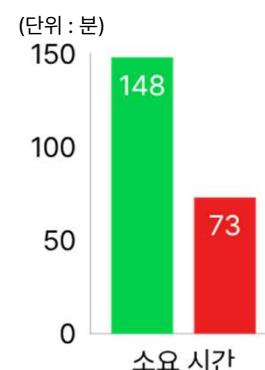
Whisper의 **환각(Hallucination)**
→ 실제 발화와 무관한 문장 생성

잡음 제거 + 백색소음 주입
→ 발표 원문대로 올바르게 추출

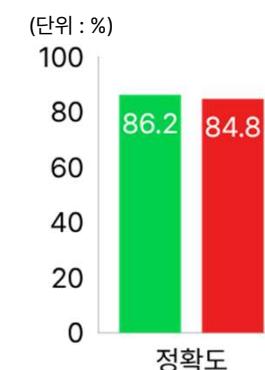
발화 추출 모델 성능 비교

■ Large-V3
■ Large-V3-Turbo

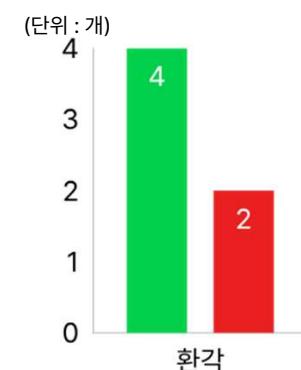
(단위 : 분)



(단위 : %)



(단위 : 개)

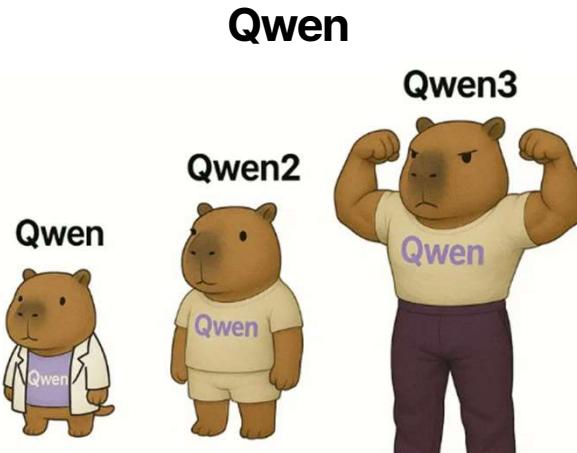




3-1 기술 소개 | Text 모델 비교



* 문맥상 중요한 문장 추출을 위한 모델 후보

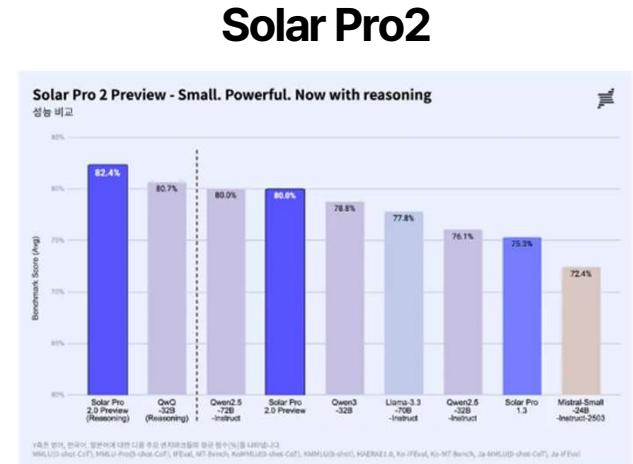


- **Qwen2.5 7B :**
한국어 LLM 성능 평가 플랫폼
'호랑이 리더보드3'의 1위 모델 (15B 미만)
- **Qwen3 32B :**
지시 이행 능력 강화 실험 위해 32B로 체급 상향
- **동아시아 언어 학습**을 통해 불완전한 문장을
보다 정교하게 복원



- **LLaMA3.3 70B :**
다국어 테스트 지표 MGSM에서 타 모델 대비
다국어 추론 능력이 높은 경우 보고됨
- **LLaMA3.2 8B-Ko :**
Hugging Face 라이브러리를 통해 한국어로
미세 조정된 모델로, 기존 모델을 능가하는
SOTA급* 지시 이행 능력 보유

*AI·머신러닝 분야 '최첨단 성능'을 달성한 모델·알고리즘



- 한국 AI 스타트업 업스테이지에서 올해 공개된
한국어 특화 대형 언어모델
- 출시 직후 한국어 MMLU / Hae-Rae / Ko-Reasoning 등 여러 벤치마크에서
GPT-4 계열 위협하는 성능 보임
- 문단·단락·문맥 흐름을 세밀하게 이해하도록
학습돼 **"담화 분석"** 작업에 매우 강함



3-1 기술 소개 | Text 모델 선정

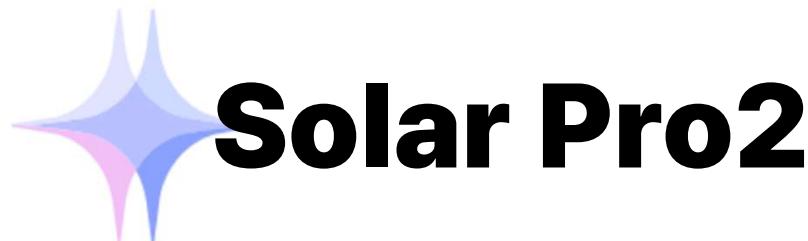


모델 평가용 정답지

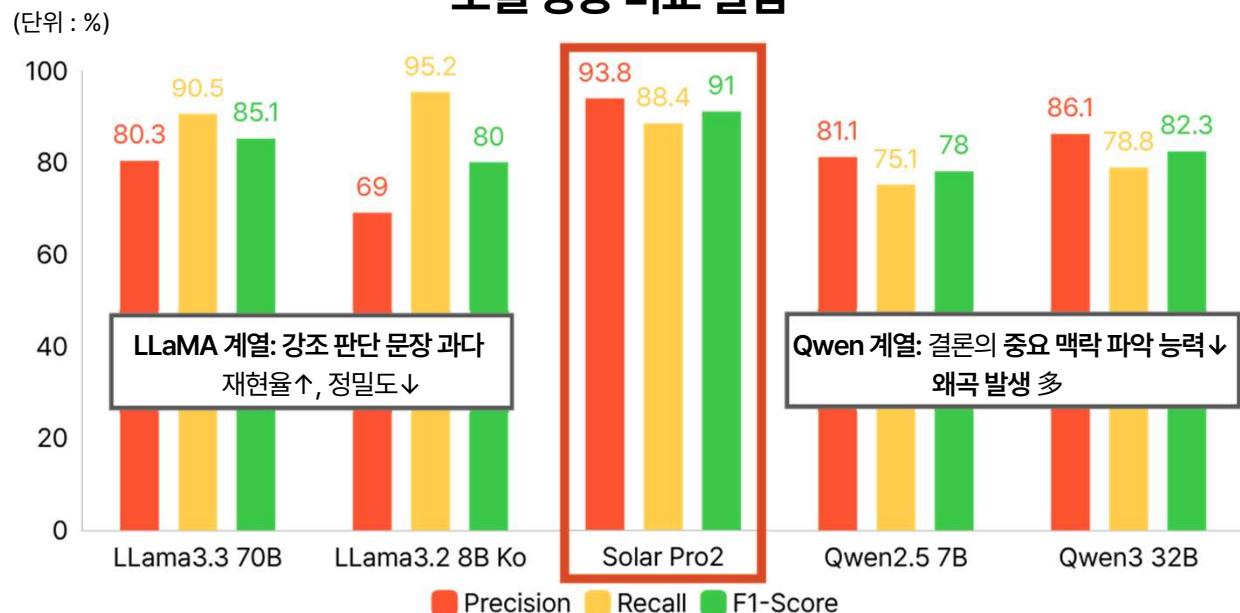
sentences	score
11월 5일 수요일 스타트 브리핑 시작합니다	0.05
먼저 조선일보입니다	0.05
코스피와 코스닥의 시가총액 격차가 최근 10년 사이 최대 수준으로 벌어졌습니다	0.9
코스피가 올해 72% 오를 때 코스닥은 37% 올라서 상대적으로 부진했기 때문입니다	0.85
올해 조만 해도 코스피의 전체 시총이 코스닥의 약 5.7배 수준이었습니다	0.8
그런데 지금은 그 차이가 7.2배로 더 커졌습니다	0.85

너는 한국어 발표·스피치 코칭 전문가이다.
입력으로 주어지는 [뉴스 전문]을 완벽히 이해한 상태에서,
[분석 대상 문장들]이 이 글 전체 맥락에서 얼마나 중요한지
0.0 ~ 1.0 사이의 점수(score)로 평가하라.

*score 0.7 이상 : 정답(1), 미만 : 오답(0)



모델 성능 비교 실험

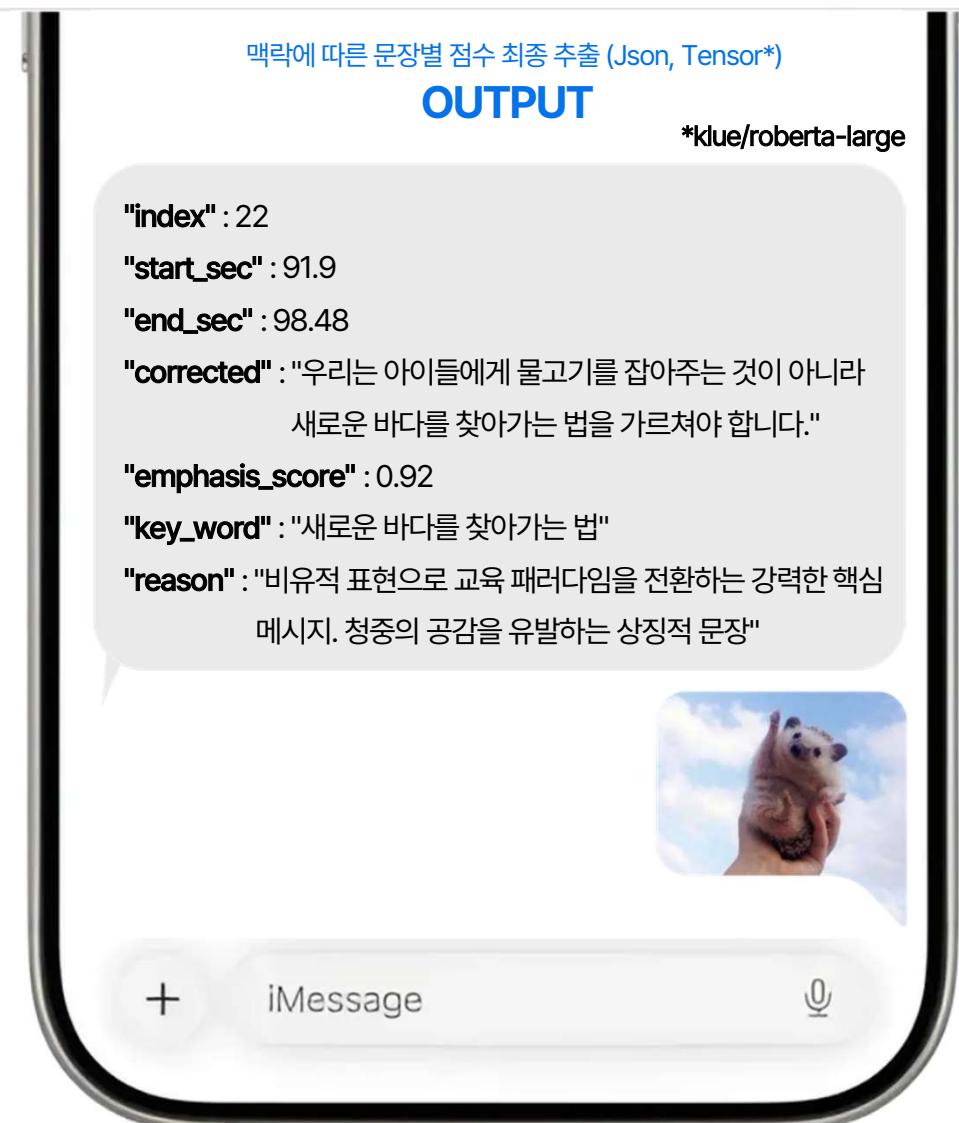
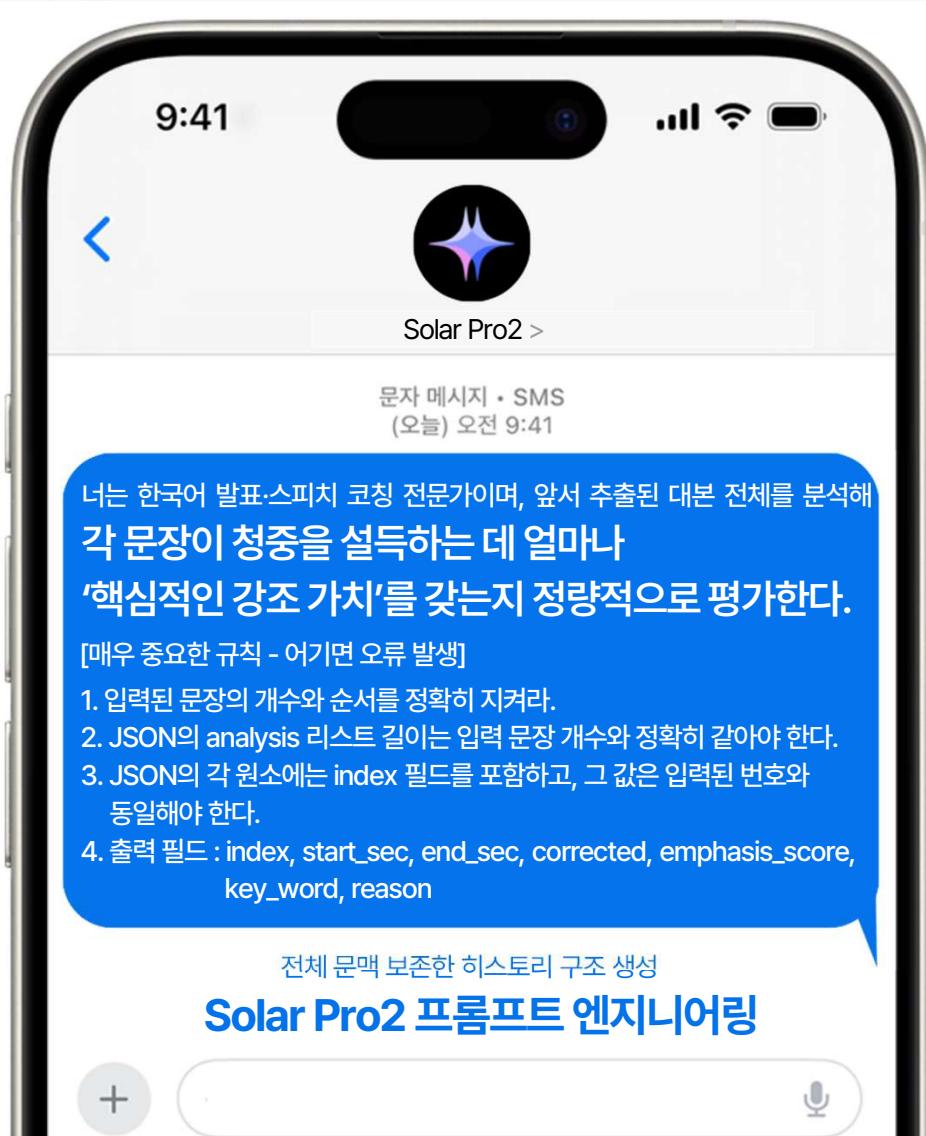


정밀도 · 재현율 모두 최상위권 → 가장 신뢰할 수 있고 균형 잡힌 결과

추가로 진행한 '문서 요약 평가 말뭉치'* 활용 실험 역시 5가지 모델 중 1위

*국립국어원 전문가들이 기사 전문 중 핵심 문장을 판단

3-1 기술 소개 | Text 구조



3-2 기술 소개 | Voice Feature



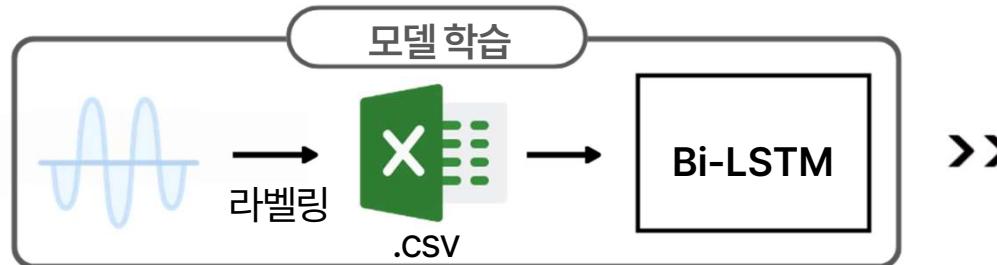
파이프라인



Input 영상



음성 추출·정규화

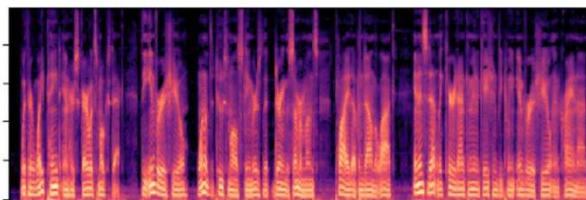


.json



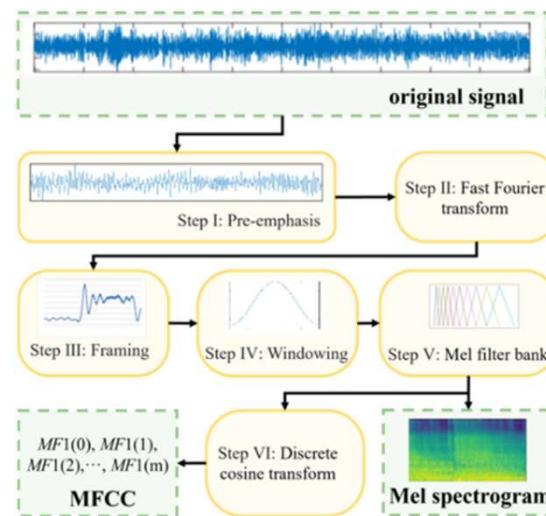
tensor

Mel-Spectrogram



- 시간의 흐름에 따른 주파수 변화를 시각화한 2차원 이미지 형태의 특징
- 원본 소리에 STFT·Mel-Scale Filter → 음성의 많은 특징을 보존

[두 기술의 작동 방식]

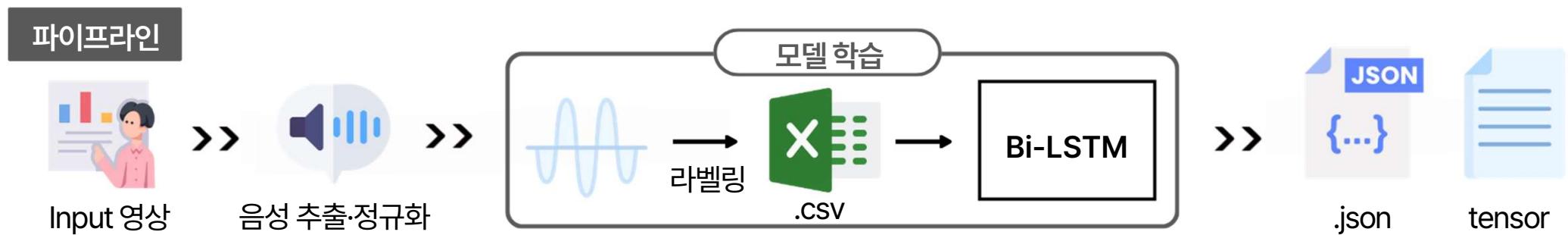


MFCC

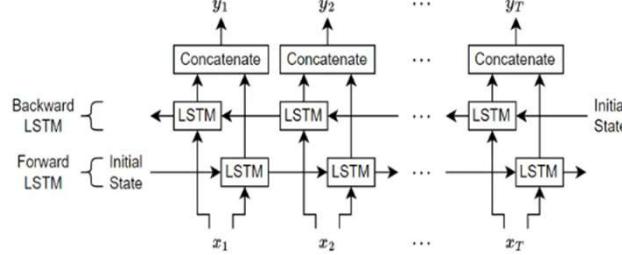


- 사람의 발화 특징과 성대 진동 특징을 분리하여 압축하는 특징 벡터
- Mel-Spectrogram에 로그·DCT → 핵심 데이터만 남음

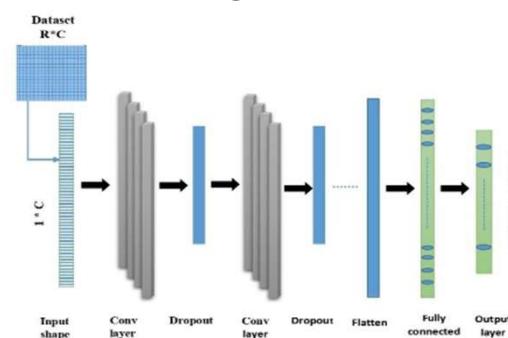
3-2 기술 소개 | Voice Feature



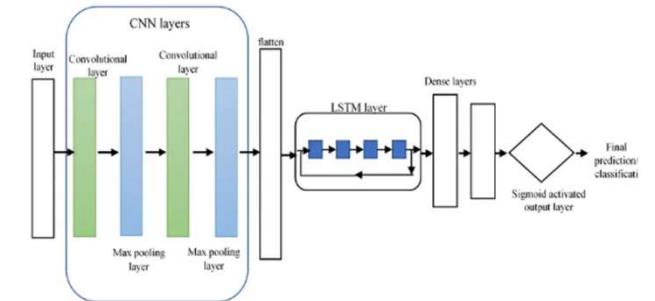
Bi - LSTM



CNN



CNN + LSTM



- 시계열 데이터의 전/후 문맥을 **양방향 학습**
- '침묵 후 발화'와 같이 **전후 관계**가 중요한 강조 유형 탐지에 유리

- **필터(Kernel)**를 통해 오디오의 **지역적 패턴(Local Feature)** 추출
- 급격한 음량·톤 변화 등 순간적 특징 포착에 강점

- 앞단에선 CNN, 뒷단에선 LSTM 진행
- CNN의 Pooling 과정에서 중요 특징만 남김
→ 단독 LSTM보다 **Noise**에 강점



3-2 기술 소개 | Voice 성능 비교



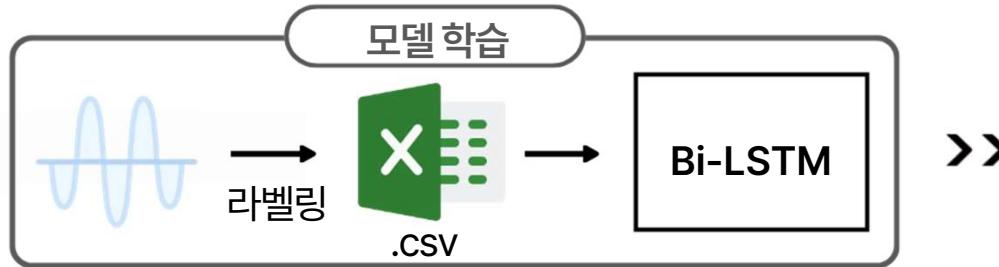
파이프라인



Input 영상



음성 추출·정규화

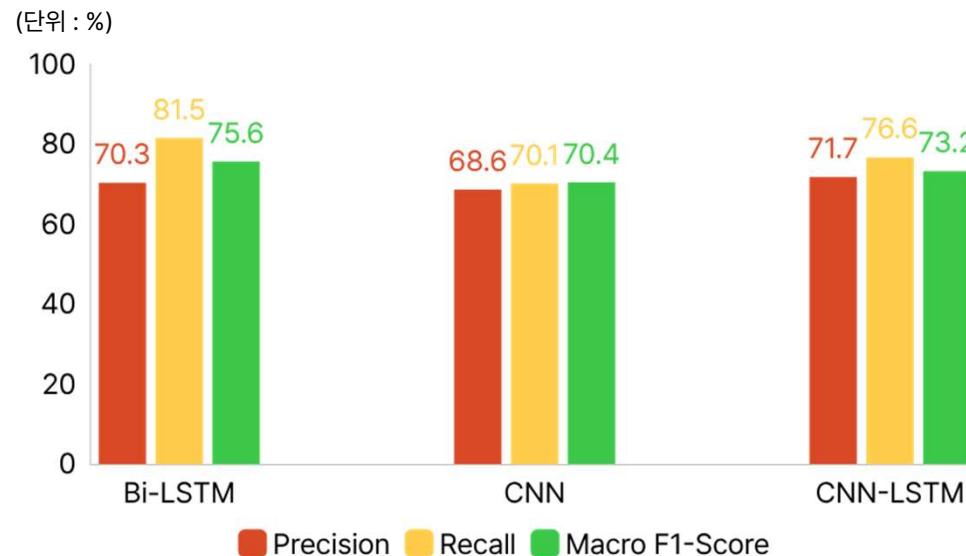


.json

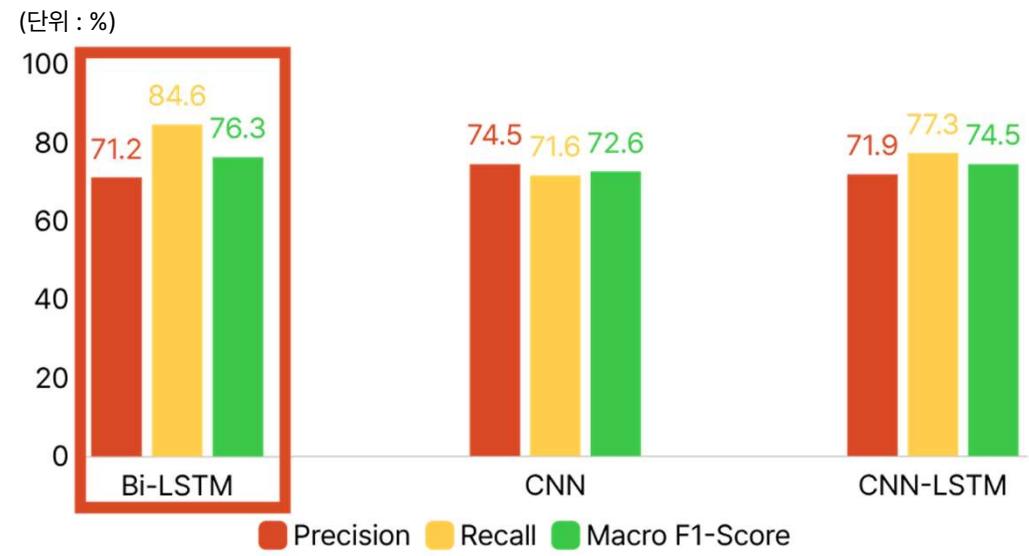


tensor

Feature = MFCC



Feature = Mel





3-3 기술 소개 | Vision 성능 비교



파이프라인

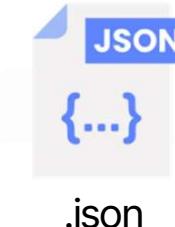


Input 영상

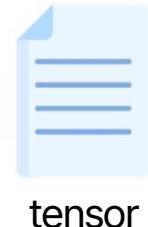
>>



>>



JSON
{...}
.json



tensor

mc3_18

- 공간(자세/손 위치) + 시간(움직임)을 섞어서 보는 구조
- 3D Conv: 움직임 여부에 대한 시간 정보 감지
- 2D Conv: 프레임별 자세/손 형태 정리

r3d_18

- 3D Conv로 공간(손/팔/상체) + 시간(동작의 흐름) 확인
- 강조 제스처 움직임의 변화 학습

r2plus1d_18

- 3D Conv를 공간(2D)과 시간(1D)으로 분해한 구조
- 2D Conv: 손/팔 위치, 상체 자세, 프레임 내 형태 이해
- 1D Conv: 속도 변화 등 시간 패턴 학습

3D CNN

Vision 성능비교

(단위 : %)



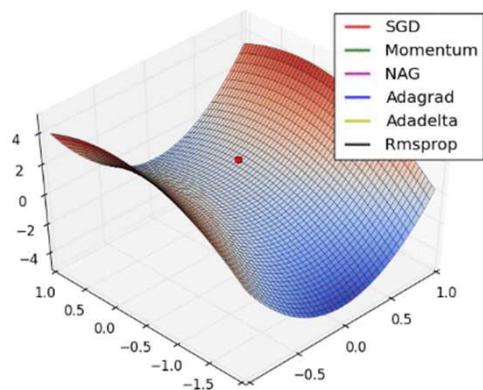


3-3 기술 소개 | Vision 성능 비교



3D-CNN r3d_18 성능 개선을 위한 주요 파라미터 조정

optimizer (최적화 방법)

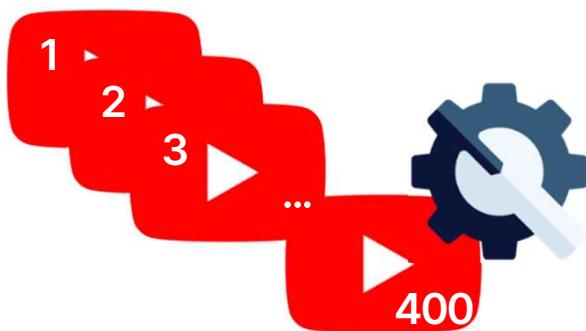


과적합을 완화하고 검증 성능을 유지

AdamW 채택

- TD Conv: 속도 향상 및 시간 패턴 학습

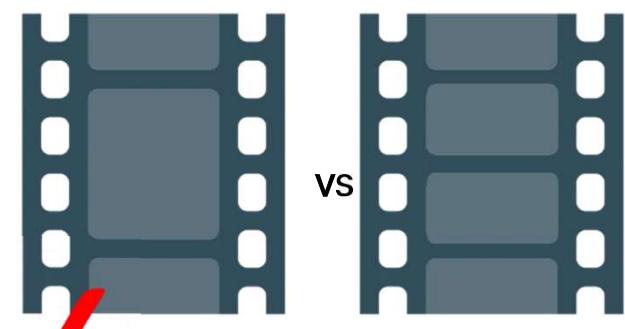
pretrained (사전학습 모델 여부)



kin 400으로 사전학습된 3d cnn 모델을 활용,
제스처로 강조한 구간 탐지에 맞게 파인튜닝

Kinetics-400 사전학습

clip_len (영상 프레임 수)



서버 시스템 특성을 고려하여

16 프레임 선정



3-4 | Multi Modal



Text

Solar-Pro2

```
"index": 22  
"start_sec": 91.9  
"end_sec": 98.48  
"corrected":  
"우리는 아이들에게 물고기를 잡아주는 것이 아니라  
새로운 바다를 찾아가는 법을 가르쳐야 합니다."  
"emphasis_score": 0.92  
"key_word": "새로운 바다를 찾아가는 법"  
"reason":  
"비유적 표현으로 교육 패러다임을 전환하는 강력한  
핵심 메시지. 청중의 공감을 유발하는 상징적 문장"
```

Voice

Bi - LSTM

```
"start_sec": 6.3  
"end_sec": 7.2  
"type": "Loud"  
"class": 3  
"scores": { "Normal": 0.2891,  
            "Pause_Talk": 0.0,  
            "High_Tone": 0.1667,  
            "Loud": 0.5352 }
```

Vision

3D CNN r3d_18

```
"start_sec": 222.36  
"end_sec": 222.86  
"gesture_emphasis_prob": 0.58
```

Tensor : 1024

Tensor : 128

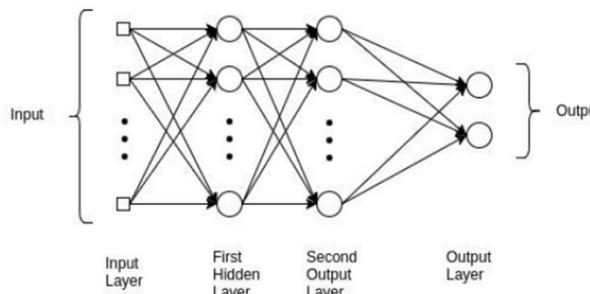
Tensor : 512

Tensor : 256

Multi Modal Model

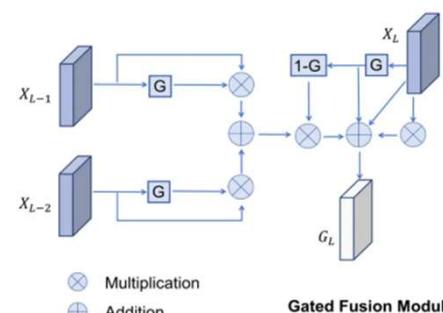
3-4 | Multi Modal 선정

Simple MLP



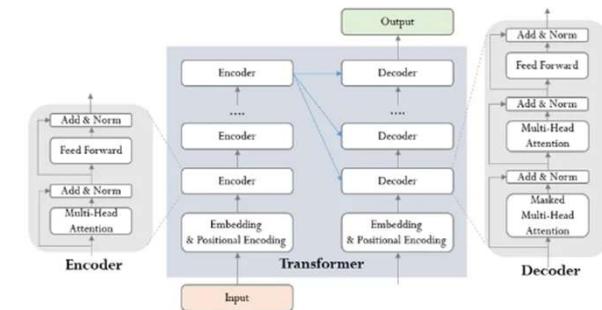
텍스트, 음성, 영상 특징을 Concat
세 모달을 동일한 비중으로 취급

Gated Fusion

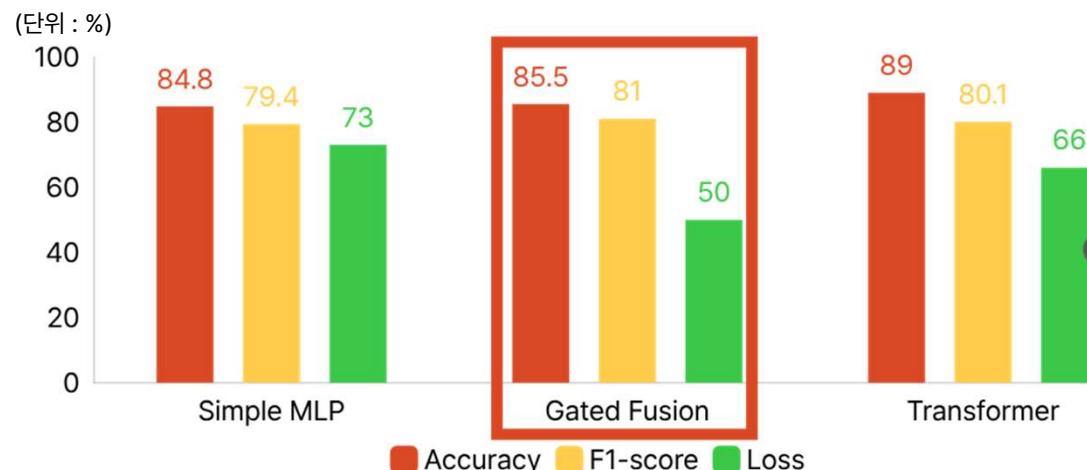


텍스트, 음성, 영상별 가중치 두어 상황별 중요도 조절
실제 환경에서 발생하는 노이즈, 불확실성에 강한 구조

Transformer



Text, Voice, Vision간 상호작용과 시간적 관계 학습
“어떤 말이 어떤 톤·제스처와 함께 나왔는지” 고려



대규모 데이터셋에서 Transformer의
압도적 성능이 선행 연구를 통해 입증되었으나*
본 프로젝트의 소규모 데이터셋에서는
Gated Fusion이 더 좋은 성능을 보임

F1-score가 가장 높고, Loss가 가장 낮은

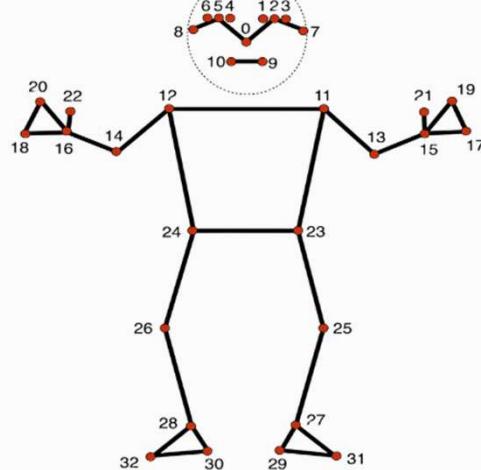
Gated Fusion

*MLP Architectures for Vision-and-Language Modeling: An Empirical Study

3-5 | 부가지표



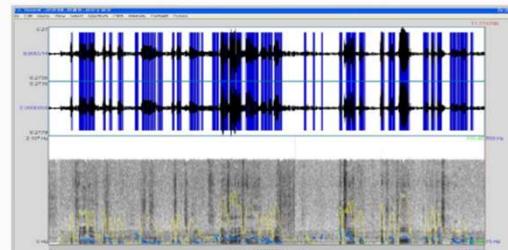
자세 불량 여부



Media Pipeline Pose

- **Media Pipeline Pose**: 신체 33개 랜드마크
- 팔꿈치 각도, 손목 위치, 발 각도 등을 계산
- 짹다리, 팔짱, 주머니 손 넣기 카운트

음정 안정성 평가



사회적으로 영향력 있는 사람들의 평균 Jitter · Shimmer 값*

Jitter	Shimmer
2.90 %	1.280 dB

*Identification of Voice for Listeners who Feel Favor Using Voice Analysis
Ji Hyun Choi, Dong Uk Cho, Yeon Man Jeong, 2015



Jitter

성대 진동의 변화량 → **톤·음정 안정성**



Shimmer

진폭 변화의 규칙성 → **성량·호흡 안정성**

불필요한 추임새 횟수

"habit_words": [

{"word": "그", "count": 5, "category": "추임새",
"comment": "문장 연결 시 '그'를 반복적으로 사용."},

{"word": "근데", "count": 4, "category": "접속 습관",
"comment": "대화체 전환 시 '근데'를 빈번히 사용."},

{"word": "좀", "count": 5, "category": "부사 습관",
"comment": "강조 없이 습관적으로 '좀'을 삽입."},

{"word": "이렇게", "count": 6, "category": "부사 습관",
"comment": "시범 설명 시 '이렇게'를 반복 사용."}],

"overall_feedback": "특히 '45도'와 '이렇게'의 반복 사용이 두드러지며, 추임새 '그'와 접속어 '근데'가 발표흐름을 방해합니다. 핵심 용어 외 불필요한 반복을 줄이는 연습이 필요합니다."

• **Whisper**로 STT 실시

• Solar-Pro2 프롬프트 엔지니어링 실시

• 불필요한 추임새 카운트



3-6 | LLM 프롬프트



.json

Multi Modal Model

Gated Fusion

```
{ "index": 1  
"start_sec": "25.3"  
"end_sec": "28.6"  
"score": "0.65" }
```

unimodal

Text

```
"index": 1  
"start_sec": "25.3"  
"end_sec": "28.6"  
"emphasize_score":  
"0.95"
```

Voice

```
"start_sec": "25.3"  
"end_sec": "28.6"  
"Loud_score": "0.65"
```

Vision

```
"start_sec": "25.3"  
"end_sec": "28.6"  
"gesture_score":  
"0.55"
```

Added

```
"불량자세횟수": "0"  
"Jitter": "3.1"  
"Shimmer": "1.4"  
"불필요한 추임새 횟수":  
"3"
```

Prompt

1. 너는 대한민국에서 제일 발표를 잘하고 발표 · 스피치 피드백 코치야
 2. 너는 피드백을 잘 해내야만 발표 · 스피치 코칭 사업을 계속 영위할 수 있고,
그렇지 않으면 (즉, 피드백에 오류가 있거나 이상한 부분이 있다면) 사업이 망해서 굽어 죽어.
 3. **첨부된 5개의 JSON 파일들을 빈틈 없이 철저히 분석해 발표자에게 피드백을 해줘야 해**
- •
•



3-6 | LLM 비교·선정



	Gemini 3.0	ChatGPT 5.2
자아 수준(0)	12(회)	14(회)
과제 수준(1)	14(회)	15(회)
과정 수준(2)	16(회)	13(회)
점수	1.10	> 0.98

▲ Hattie 모델 기반 피드백 수준별 가중 평가

- 0(자아 수준) : 단순 칭찬, 인사, 격려
- 1(과제 수준) : 맞고 틀리다, 크고 작다 등 결과와 현상에 대한 단순 교정
- 2(과정 수준) : 구체적 전략, 인과관계, 비유를 사용해 원리 설명
- 점수 산정: $0 * (0\text{ 횟수}) + 1 * (1\text{ 횟수}) + 2 * (2\text{ 횟수}) / (0\sim2\text{ 전체 횟수})$

Vision 피드백

"start_sec": 120.9
 "end_sec": 134.12
 "text": "시공간을 확인하는 MC3, 시공간을 한 번에
 보는 R3D, 처음부터 사진과 시간을 분리해
 단계적으로 학습하는 R2 플러스 1D 성능을
 비교했습니다."
 "advice": "복잡한 모델명을 나열하는 구간이라
 청중이 지루해할 수 있습니다. 모델을
 언급할 때마다 손으로 위치를 다르게
 가리키는 (좌, 중, 우) 제스처를 사용해
 시각적으로 구분해 주세요."

Voice 피드백

"start_sec": 90.9
 "end_sec": 101.1
 "text": "그 결과 보시다시피 멜스펙트로그램과
 BI-LSTM 조합을 최종적으로 선택하게
 되었습니다."
 "advice": "발표의 하이라이트 중 하나입니다.
 제스처 활용은 좋으나 목소리가 다소 평이
 합니다. '최종적으로 선택'이라는 말에서
 힘찬 목소리로 확신을 보여주어야 합니다."

Gemini 선정

- 어려운 내용도 청중의 머릿속에 그림 그려지듯 쉽게 설명할 수 있도록 도움
- 발표의 결정적 순간에 청중의 귀에 꽂히는 **확실한 정보 전달**
- 맥락에 최적화된 구체적 전략을 제시해 발표자 **스스로 연출력 성장**





3-7 | Multi Modal 점수 산정



기준표			
Text	Voice	Vision	점수
0	1	1	-2
1	0	1	+1
1	1	0	+1
1	1	1	+3

* 비강조구간에서 강조한 경우 / 강조구간에서 강조부분
누락 위주로 점수 산정

67 스티브 따라잡스
발표 평가
발표자 : 서태빈

	11.3초 ~ 12.6초	25.6초 ~ 28.4초	62.1초 ~ 65.4초
Text	O	O	X
Voice	X	O	X
Vision	X	O	O
점수	69	72	71

The application interface includes a video player showing Steve Jobs speaking, a summary text, and a bar chart.

Summary Text:

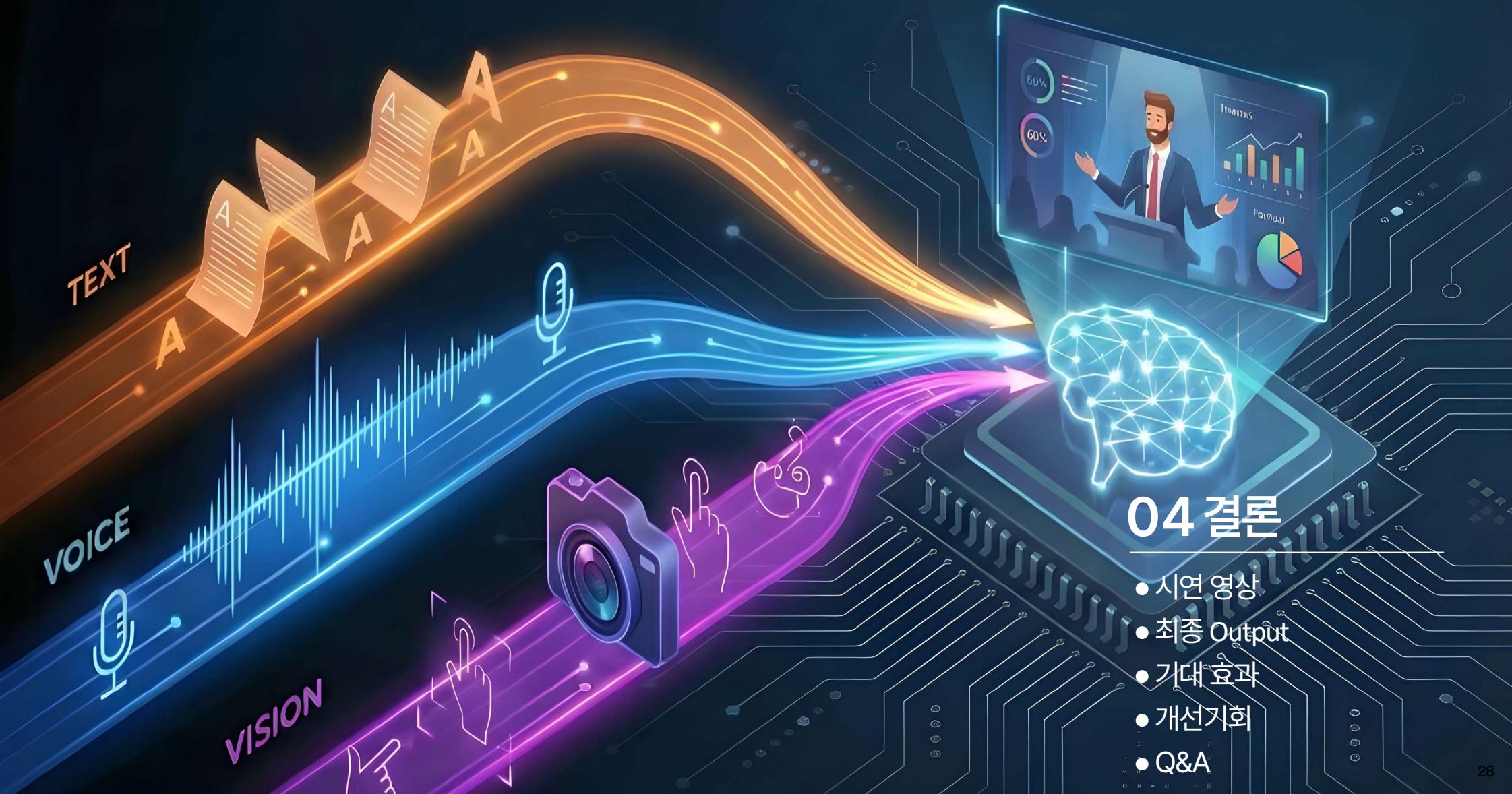
서태빈 님의 발표 점수는
67점입니다.
발표자님의 영상을 분석한 결과,
37초-40초 구간은 강조 포인트였지만
제스처가 사용되지 않았습니다.
영상을 참고해 제스처를 사용해 보세요.

Bar Chart Data:

목소리톤	13%
제스처	5%
발음	4%

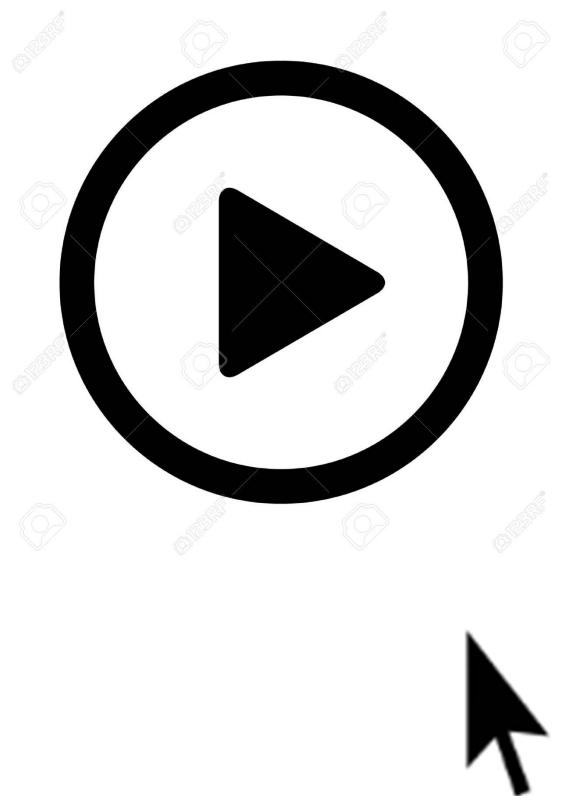
Video Player Controls:

▶ 0:38 / 1:23 ⏴ ⏵ ⏵ ⏵





4-1 시연 영상





4-2 | 최종 Output



피드백, 그 이상의 피드백

가장 완벽한 발표, 그것은 아마도 '스티브 따라잡스'에서.

당신의 발표 점수는 **86점**입니다.

미세한 단점이 조금 있었나 보군요.
오직 당신만을 위한, 잠스의 맞춤형 코칭 영상을 참고해보시는 건 어떤가요?

#1

발언 내용

다음은 보이스 부분입니다.

피드백

자연스럽게 잘 처리하셨습니다. '보이스 부분'과 관련된 내용을 차분한 어조로 전달하여 안정감을 주고 있습니다.

#2

발언 내용

보이스에 관한 두 개의 피처를 소개하겠습니다.

피드백

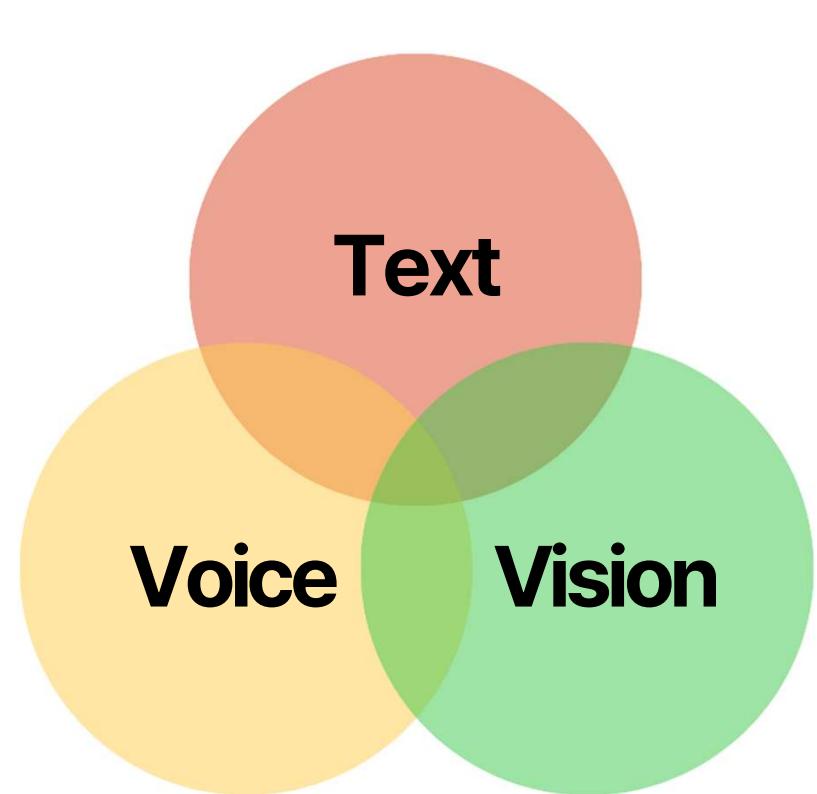
두 개의 를 말할 때 손가락으로 숫자 2를 표시해봐. 시선을 사로잡게 될 거야.

잠스라면 이렇게 했을 겁니다...

이 부분은 '퓨처에 대한 핵심 설명'으로 더 강한 강조가 필요합니다. 현재 높은 톤(High Tone)인 감지되었고 제스처는 부족합니다. 두 개의 를 말할 때 손가락으로 숫자 2를 표시하거나, '퓨처'에서 손바닥을 펴는 제스처를 추가하여 청중의 시선을 사로잡으세요.



4-3 기대 효과



언행일치 (Alignment) 분석을 통한 피드백

- Text · Voice · Vision 모두 고려한 강조 분석
- '문맥상 강조이나, 목소리가 작습니다' 식의 구체적 피드백 가능



시간 · 공간 · 비용으로부터 자유로운 서비스

- 영상만 업로드하면 언제든지 사용 가능
- 비싼 '스피치 학원'에서 받는듯한 피드백을 제공



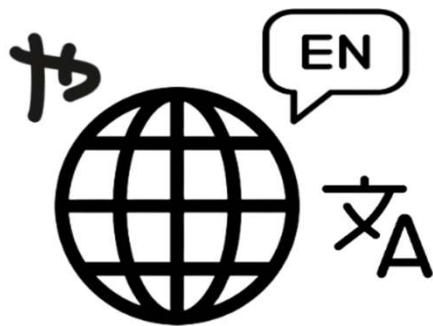
발표 역량의 '메타 인지' 강화

- 주관적 감각이 아닌 데이터에 기반한 객관적 피드백 제공
- 사용자의 제스처, 목소리, 발화 습관을 스스로 인식하고 개선 가능

4-4 개선기회

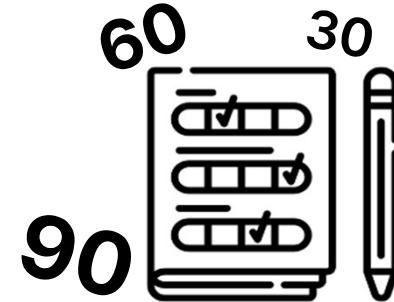


언어 다양성 확장 가능성



- 다양한 언어적 특성 반영한 데이터셋 확장
→ 글로벌 학습자 대상으로 폭넓은 피드백 제공 가능
- 다국어에 대해 다양한 모델을 비교 연구
→ 연구의 깊이와 완성도 향상
- 언어별 표현 특성에 최적화된 피드백 가능
→ 언어마다 다른 강조 어순, 억양 반영 분석

정량적 점수 구체화



- 피드백 요소 간 우선 순위 설정 불명확
→ 신뢰도 있는 비교 및 평가에 제약
- 객관적 점수화의 낮은 우선 순위
→ 평가의 객관성 및 깊이 부족
- 발표에 관한 도메인 지식 적극적 활용
→ 신뢰도 · 구체적 수식으로 고도화된 점수 제공

Transformer



- 데이터 규모의 제약
→ 선행 연구에서 입증된 Transformer 성능을 적용하지 못함
- 데이터셋 추가 확보로 Transformer와 Gated Fusion에 대한 추가 검증 연구 수행 기대
→ 모델 선택 기준과 성능 차이 분석 가능



감사합니다