ECHO!

한국어 L1 화자를 위한 영어 발음 음소 단위 피드백 서비스

A6 O(1) 중간보고 PT

2026-05-13

김형준, 김지용, 백지영, 이찬영, 신영환, 최현지

목차

RECAP

문제 ① 인식기의 오판

학습자가 정확하게 발음했는데

인식기가 틀렸다고 판정하는 경우가 있다

학습자 발음

f æ ʃ ə n ✓

인식기 출력

p æ ʃ ə n ✕

FRR 5.83%

정답 100개 중 약 6개가 억울하게 오판

문제 ② 피드백을 이해할 수 없다

학습자가 rice 의 r 을 l 로 발음 → lice

기존 앱의 피드백

혀끝을 입천장에
닿지 않게
말아 올리세요

조음 위치 기반 영어 설명

학습자의 반응

한국어 ㄹ 로는
둘 다 같은 소리인데
뭘 다르게 하라는 건지 모르겠다

모국어 직관과 연결되지 않는 피드백은 실행이 어렵다

ECHO 의 접근

문제 ① 인식기 오판 →

FiLM Canonical Injection

정답 음소를 인식기에 주입해
오판률 저감

문제 ② 피드백 이해 불가 →

아학편 기반 L1 피드백

한국어 음운 체계로
이해 가능한 피드백 생성

두 문제를 인식기와 피드백 양쪽에서 동시에 개선

Wav2Vec2 + FiLM

+

LLM + 아학편 지식

현재 진행 상황

프로토타입 구현 완료

로그인 기능 제외 · 전 기능 동작 확인

완료

ML 모델 학습

Wav2Vec2 + FiLM

완료

모델 서버 + 백엔드

FastAPI ↔ Spring Boot 연동

완료

프론트 UI + 연동

React SPA · 13개 화면

새로 발견된 문제

잡음 환경에서 음소 인식률 저하

실제 사용 환경 (카페, 이동 중) 에서 인식 정확도가 떨어지는 현상

기능 소개

기능소개

I. 로그인 기능 및 프로필

ECHO!

아이디

아이디를 입력하세요

비밀번호

비밀번호를 입력하세요

로그인

또는

회원가입

회원가입

ECHO!에 오신 것을 환영합니다

아이디

아이디를 입력하세요

중복확인

비밀번호

비밀번호를 입력하세요

비밀번호 확인

비밀번호를 다시 입력하세요

닉네임

닉네임을 입력하세요

이메일

이메일을 입력하세요

중복확인

서비스 이용약관에 동의합니다

연속 출석 1일!

EXP 1000

구글 데모

이메일

demo@google.example

닉네임

구글 데모

변경

비밀번호 설정

알림

ON

약관 및 정책

홈 통계 프로필

demo@google.example

닉네임

구글 데모

변경

비밀번호 설정

알림

ON

약관 및 정책

회원탈퇴

로그아웃

홈 통계 프로필

기능소개

II. 학습 트랙

학습 종류 선택

연속 출석 0일!

EXP 0

Hello,
구글 데모님

학습 트랙

트랙을 골라 챕터별로
차근차근 익혀요

맞춤 학습

내 대본으로 한 문장씩 연습해요

기초 학습

연속 출석 0일! EXP 0

←

학습 트랙

기본 발음 트랙
잰말놀이부터 R/L · V/B · F/P · TH 까지, 영어 발음의 기본기를 빠르게 잡는 입문 트랙.

챕터 5개

모음 구별 트랙
한국어에 없는 영어 모음 길이/혀 위치 차이를 짝 단어로 익히는 트랙.

챕터 2개

일상 회화 표현
처음 만난 사람과 자연스럽게 인사하고 자기소개하는 짧은 표현을 또렷하게 발음해요.

챕터 2개

홈 통계 프로필

연속 출석 1일! EXP 1000

←

기본 발음 트랙
잰말놀이부터 R/L · V/B · F/P · TH 까지, 영어 발음의 기본기를 빠르게 잡는 입문 트랙.

▶ 처음부터 시작

챕터 목록

1 **영어 잰말놀이**
난이도: MEDIUM

2 **발음 연습: R vs L**
난이도: MEDIUM

3 **발음 연습: V vs B**
난이도: MEDIUM

4 **발음 연습: F vs P**
난이도: MEDIUM

5 **발음 연습: TH vs DH**
난이도: HARD

홈 통계 프로필

연속 출석 1일! EXP 1000

←

모음 구별 트랙
한국어에 없는 영어 모음 길이/혀 위치 차이를 짝 단어로 익히는 트랙.

▶ 처음부터 시작

챕터 목록

1 **발음 연습: A vs E**
난이도: EASY

2 **발음 연습: I vs EE**
난이도: MEDIUM

홈 통계 프로필

연속 출석 1일! EXP 1000

←

일상 회화 표현
처음 만난 사람과 자연스럽게 인사하고 자기소개하는 짧은 표현을 또렷하게 발음해요.

▶ 처음부터 시작

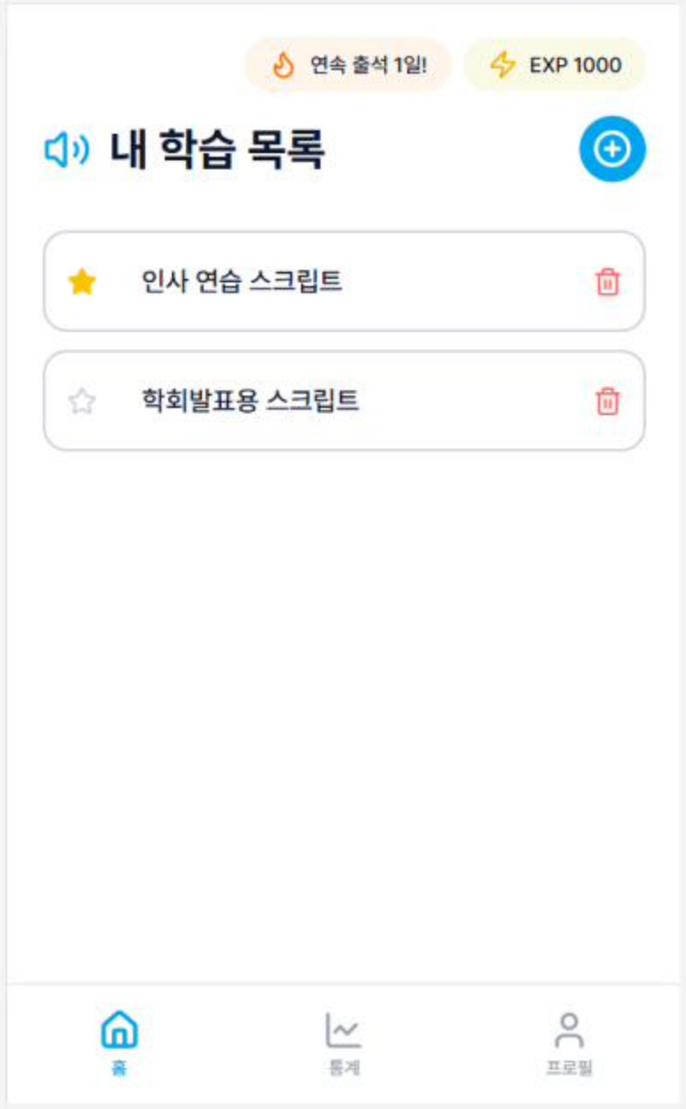
챕터 목록

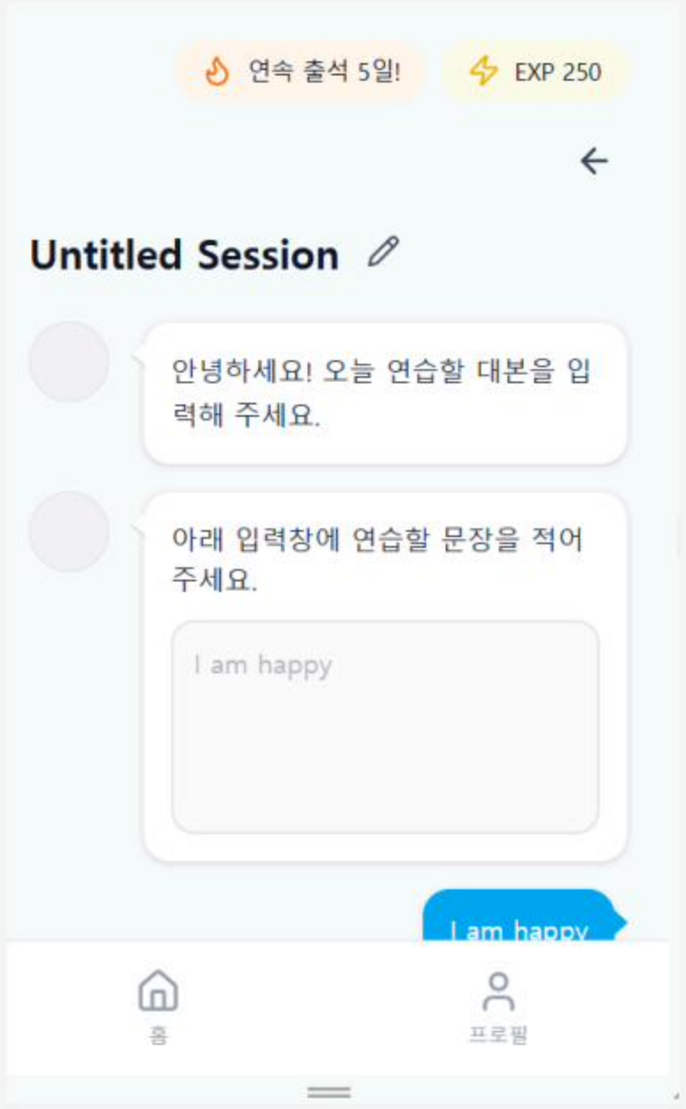
1 **인사말 익히기**
난이도: EASY

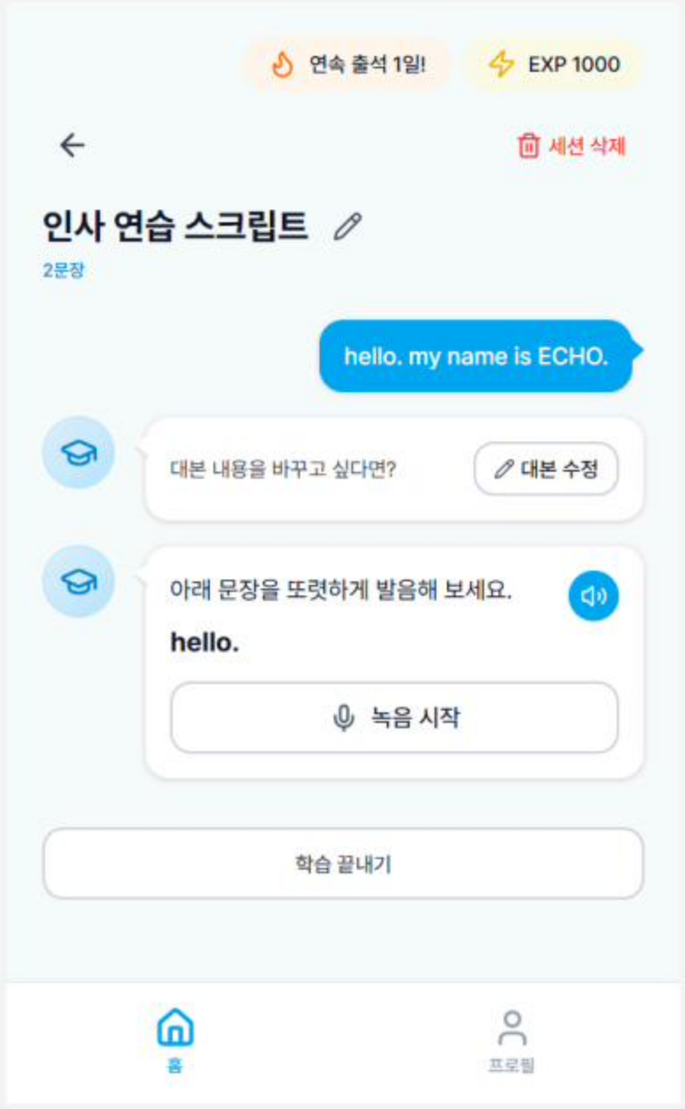
2 **자기소개 표현**
난이도: EASY

홈 통계 프로필

맞춤 학습


연속 출석 1일!
EXP 1000
내 학습 목록
인사 연습 스크립트
학회발표용 스크립트
홈
통계
프로필


연속 출석 5일!
EXP 250
Untitled Session
안녕하세요! 오늘 연습할 대본을 입력해 주세요.
아래 입력창에 연습할 문장을 적어 주세요.
I am happy
I am happy
홈
프로필


연속 출석 1일!
EXP 1000
세션 삭제
인사 연습 스크립트
2문장
hello. my name is ECHO.
대본 내용을 바꾸고 싶다면?
대본 수정
아래 문장을 또렷하게 발음해 보세요.
hello.
녹음 시작
학습 끝내기
홈
프로필

기능소개

III. 발음 연습

피드백 과정

기본 발음 트랙 · 챕터 1/6

영어 잰말놀이

오늘의 잰말놀이에요. 아래 문장을 빠르게 따라 읽어보세요.

I slit the sheet, the sheet I slit, and on the slitted sheet I sit.

녹음 시작

학습 끝내기

전체적으로 발음이 많이 생략되었어요. 'I'는 '아이'로, 'sheet'는 '쉬-이트'처럼 발음해 보세요. 마지막 'sit'는 'y, uw, r'이 아닌 'ㅅ(si)' 발음에 집중해야 해요.

I slit the sheet, the sheet I slit, and on the slitted sheet I sit.

정답

ay	s	l	ih	t	dh	ah	sh
iy	t	dh	ah	sh	iy	t	ay
s	l	ih	t	ah	n	d	aa	n
dh	ah	s	l	ih	t	ah	d
sh	iy	t	ay	s	ih	t	
당신	y	uw	r				

다시 발음하기

학습 마무리

학습 끝내기

이번 학습 정확도는 0.0점이에요.

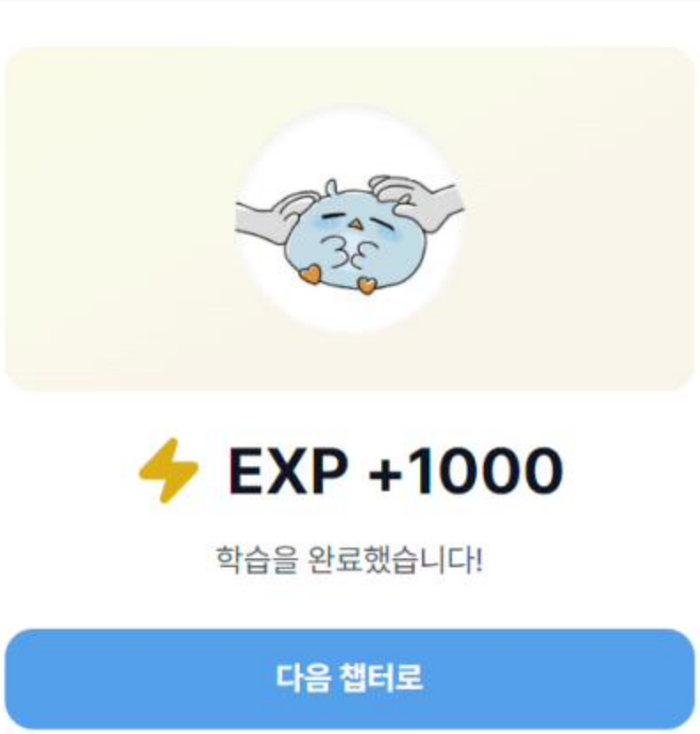
안녕하세요! 잰말놀이 학습 결과, 'T' 발음을 가장 많이 놓치셨네요. 'T' 발음을 좀 더 또렷하게 내기 위해, 혀를 윗니 뒤쪽에 튕기듯이 붙였다 떼는 느낌으로 연습해 보세요. 예를 들어 daughter:(으)도우터, dust:으더스(트) 처럼요.

약점 음소 T 를 한 번 더 연습해 볼까요?

한 번 더 연습하기

학습 완료하고 경험치 획득

피드백 과정



⚡ EXP +1000

학습을 완료했습니다!

다음 챕터로

기능소개

IV. 통계 기능

통계

연속 출석 5일!

EXP 250

2026년 4월

일	월	화	수	목	금	토
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

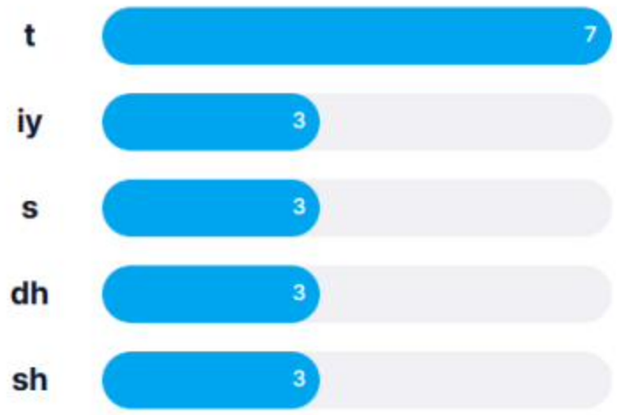
연속 출석 단계

1일 3일 5일 7일+

지난 일주일간 많이 틀린 발음

홈 통계 프로필

지난 일주일간 많이 틀린 발음



업적

첫 학습 완료

3일 연속 출석

7일 연속 출석

2주 연속 출석

30일 연속 출석

잰알늄이 5회 완료

R vs L 마스터

V vs B 마스터

F vs P 마스터

TH 마스터

A vs E 마스터

I vs EE 마스터

인사말 마스터

자기소개 마스터

맞춤 학습 도전자

맞춤 학습 마스터

완벽한 한 판

기능소개

V. 사용 사례

사용영상

설계 및 구현

설계 및 구현

OVERVIEW

ECHO 시스템 구조

Frontend

사용자 학습 화면

채팅형 학습 흐름

마이크 녹음과 업로드

음소 정렬 시각화

출석, EXP, 랭킹, 배지

JWT 인증 보호 라우트

REST

Backend

도메인과 오케스트레이션

학습 자료, 사용자 도메인

점수 산출 정책

EXP, streak, 배지 정책

인증과 권한

외부 AI 호출의 단일 진입점

학습 데이터 시드 적재

비즈니스 예외 코드 (ErrorCode)

JPA

Database

사용자, 학습 자료, 기록

트랙, 챕터, 녹음, 피드백, 세션

개발환경: H2 in-memory (create-drop)

HTTP

External AI

블랙박스

발음 평가 모델

오디오 → 음소와 정렬 결과

/analyze, /g2p

한국어 코칭 LLM

평가 결과 → 학습자 안내문

step, retry, unit, practice-word

사용하는 기존 기법 및 기술

Frontend

React 18
UI 라이브러리

TypeScript
정적 타입 자바스크립트

Vite
개발 서버와 번들러

Tailwind CSS
utility 기반 스타일링

Backend

Spring Boot
Java 기반 웹 프레임워크

Spring Security + JWT
인증과 보호 라우트

Spring Data JPA
엔티티 매핑과 쿼리 자동 생성

H2
in-memory 개발용 데이터베이스

음성 모델

PyTorch
딥러닝 프레임워크

HuggingFace Transformers
사전학습 모델 로딩과 파인튜닝

wav2vec2-base
자기지도 학습된 음성 표현 모델

torchaudio
오디오 로딩과 리샘플링

음소 표기와 정렬

ARPAbet, IPA
영어 음소 표기 체계

CMU Pronouncing Dict
영단어 → 음소 사전

g2p_en
사전 미등재 단어용 G2P 신경망

CTC
가변 길이 시퀀스 학습 손실

FiLM
조건부 feature 변조 (γ, β)

외부 AI

Gemini (Google)
한국어 코칭 문장 생성 LLM

음성 평가 결과 (인식 음소, op 배열, 점수)
를 입력 받아 학습자용 한국어 안내문을
생성한다.

프롬프트는 단계별, 재시도, 단원 종료, 단어 연습으로 분리

데이터셋

AIHub 한국인 영어 발음
학습 본체, 약 62,000 샘플

L2Arctic
비원어민 영어 보조, 약 3,500 샘플

MUSAN
환경 잡음 코퍼스 (강건성 평가)

녹음 한 건의 처리 흐름

Frontend

1

녹음 시작과 종료

사용자가 마이크로 영어 발음을 입력한다.

Frontend

2

WAV 인코딩과 업로드

브라우저 오디오를 백엔드로 전송한다.

Backend

3

발음할 영문 문장 결정

트랙 단계 또는 사용자 입력 문장 중 하나를 고른다.

Backend

4

정답 음소 산출 의뢰

텍스트로부터 비교 기준 음소를 받는다.

Backend

5

음성 평가 의뢰

오디오와 정답 음소로 정렬과 오류를 받는다.

Backend

6

점수 산출과 한국어 가이드 요청

정책으로 0에서 100 점을 계산하고 코칭 문장을 요청한다.

Frontend

7

응답과 화면 갱신

프론트가 채팅 버블과 음소 정렬을 그린다.

발음 평가 모델

블랙박스

텍스트 → 음소

오디오 → 정렬 결과

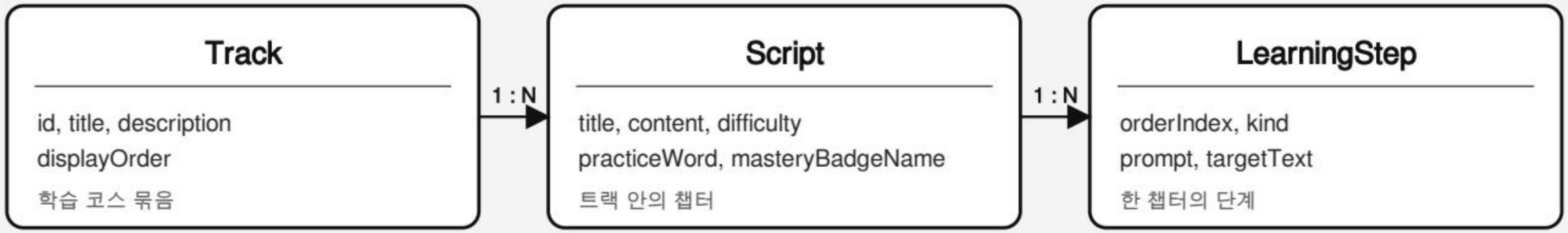
한국어 코칭 LLM

블랙박스

평가 결과 → 학습자 안내문

데이터 모델

학습 자료 (시드 데이터)



사용자와 맞춤 학습

평가 결과 (런타임 누적)

메인 구현 1

피드백 모듈

LLM 피드백 프롬프팅 데이터 선별 및 정제

#/d/ vs /t/
daughter:ᄯᅩ우터,dust:ᄯᅥ스(트),district:씨스트릭(트),deer:씨이어,degree:씨그리이,dish:씨쉬,dawn:ᄯᅡ운,dusk:ᄯᅥ(스크),drought:ᄯᅳ롸웃,dark:ᄯᅡ아(크),deaf:ᄯᅦ(프),drag:ᄯᅳ랙,document:ᄯᅩ큐먼(트),draw:(ᄯᅳ)로우,divine:씨빠인,dye:ᄯᅡ이,dry:ᄯᅳ라이,disorder:씨쏘더 ->으

#/b/
book:ᄲᅳ크,brother:(ᄲᅮ)로ᄯᅥ,brain:(ᄲᅮ)레인,breast:(ᄲᅮ)레스(트),back:ᄲᅢ(크),brook:(ᄲᅮ)룩,bridge:(ᄲᅮ)릿쥐(),bear:ᄲᅦ어,bug:ᄲᅥ(그),brick:ᄲᅳ리(크),bed:ᄲᅦ(드),board:ᄲᅩ(드),bridle:(ᄲᅮ)라이들,bottle:ᄲᅩ틀,ball:ᄲᅩ올,bowl:ᄲᅩ울,basket:ᄲᅡ스켓,bow:ᄲᅩ우,bait:ᄲᅦ이(트),broth:ᄲᅮ로ᄽᅳ,birthday:ᄲᅥᄽᅳ데이,bury:ᄲᅦ뤼,bound:ᄲᅡ운(드), -> 으 삽입

LLM 피드백 프롬프팅 데이터 선별 및 정제

A 방법 제로샷	B 방법 IPA (영어 음소 기호)	C 방법 아학편 (한국어 음소 기호)

총 3종류의 영어 발음 피드백 산출 (Sonnet 4.6)
3개 1세트*6 = 18개의 피드백을 16명 대학생이 평가

LLM 피드백 프롬프팅 데이터 선별 및 정제

차트 복사

(12)
단어: reports
문장: It was my reports from the north which chiefly induced people to buy

피드백:
R 발음 교정
'reports'의 R은 혀를 잇몸에 절대 닿으면 안 돼요! 구한말 교재 아학편에서 ruler를 '(으)룰러', robber를 '(으)로버'로 표기한 것처럼, '으'를 살짝 넣어 '(으)뤼포츠'처럼 시작 해보세요. 혀가 어디에도 닿지 않게 공중에 뜬 채 발음하는 것이 핵심이에요.

'으' 모음 삽입 예시에서 아학편 피드백이 압도적
어두 /r/: C(4.54) > A(3.91) > B(3.62)

LLM 피드백 프롬프팅 데이터 선별 및 정제

차트 복사

(3)
단어: anything
문장: I came for information more out of curiosity than anything else

피드백:
구한말 영어교재 아학편에서는 'th'를 ㅇ계 합용병서로 표기해 마찰음임을 명시했어요
(think→ᄊᆞ잉크). 'anything'의 th(/θ/)는 T처럼 혀를 튕기지 말고, 혀끝을 윗앞니 끝에
가볍게 댄 채 그 사이로 공기를 마찰시키며 내보내세요.

/θ/ 오류에서는 꼴찌
/θ/ 오류: A(4.06) > B(3.90) > C(3.50)
사용자 피로감을 유발하는 합용병서 표기를 현대 한글로 대체 필요

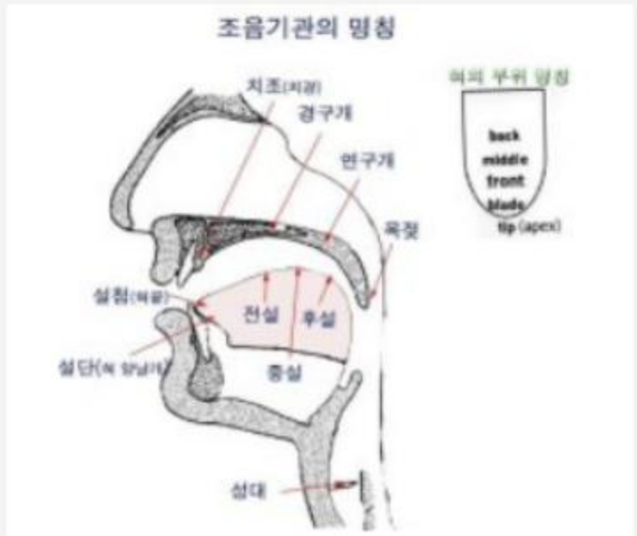
LLM 피드백 프롬프팅 데이터 선별 및 정제

설문조사 결과 분석

- IPA 피드백에서 활용하는 영어 음소 기호별 인지도
/θ/(85.7%), /æ/(64.3%)
/ð/(21.4%), /dʒ/(28.6%), /ɪ/(28.6%)

그러므로 현대 한국어 음소로 대체가 불가능한 아학편 데이터를 모두 영어 음소 기호로 대체하기에 무리가 있다고 판단함.

LLM 피드백 프롬프팅 데이터 선별 및 정제



조음 위치 이미지 예시

현대 한국어 대체 불가능한 음소

: 아학편 인용 없이 조음법에 대한 한국어 설명 + 조음 위치 이미지 자료 피드백 제공

현대 한국어로 대체 가능한 음소

: 어두 /r/, 장음(/iː/)유성음(/d/, /b/, /z/)은 반드시 아학편 예시 2개를 직접 인용해서 피드백 제공

AI피드백제공을위한프롬프트엔지니어링

맥락	역할 부여	너는 한국인 영어 학습자를 돕는 발음 코치다. 나는 한국인을 위한 영어 발음 교정 서비스의 개발자이다.
	타겟 청중	학습자는 한국어가 모국어인 영어 발음 학습자이다.
	매체 또는 채널	출력 결과는 영어 발음 교정 서비스를 제공하는 교육용 앱 내 피드백으로 사용된다.

Jacobsen, L.J.; Weber, K.E. The Promises and Pitfalls of Large Language Models as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback. AI 2025, 6, 35. https://doi.org/10.3390/ai6020035

AI피드백제공을위한프롬프트엔지니어링

미션	미션/질문	미션은 한국인이 듣고 곧바로 따라할 수 있을 만큼 명료하게 틀린 영어 음소에 대한 발음법을 한국어로 풀어 설명하는 것이다.

AI피드백제공을위한프롬프트엔지니어링

명확성과 구체성	형식 및 제약	IPA / 국제음성기호 (ɹ, θ, ʃ 같은 기호) 는 절대 쓰지 않는다. ARPAbet (r, th, sh) 같은 음소 코드도 본문에 쓰지 않는다. 필요하면 "R 발음", "TH 발음" 처럼 쉬운 영어 글자로. 가장 자주 틀린 발음 한 가지를 짚고 다음 학습에 도움될 한 줄 코칭을 더해 두 문장 이내로 작성한다. 이모지 없이, 친근한 존대 어투.

AI피드백제공을위한프롬프트엔지니어링

명확성과 구체성	도메인 특수	입 모양·혀 위치·바람·성대 울림 같은 구체적 단서를 포함해 안내한다. 가능하다면 다음과 같은 용어를 사용하여 설명한다. 제로샷/IPA/아학편
	논리	프롬프트를 제시할 때 정보의 순서 구성

AI피드백제공을위한프롬프트엔지니어링

아학편 도메인 특수 프롬프트 적용 전

안녕하세요! L 발음 연습하시느라 고생 많으셨어요. L 발음을 R 발음이나 N 발음으로 내는 실수가 잦으셨네요. L 발음은 혀끝을 앞니 바로 뒤쪽 잇몸에 붙였다 떼면서 소리를 내는 건데요, 혀가 입천장에 닿지 않도록 주의하면서 다시 한번 연습해 보시면 좋을 것 같아요! 다음번엔 L 발음에 더욱 자신감이 붙으실 거예요.

아학편 도메인 특수 프롬프트 적용 후

L 발음에서 어려움을 느끼시는 것 같아요. "L" 소리는 혀끝을 앞니 뒤쪽 잇몸에 닿게 하고, "을" 소리를 내듯이 발음해야 해요. 예를 들어 "learn:(을)러언", "leg:(을)레그"처럼 연습해보시면 좋을 것 같아요. 혀의 위치에 좀 더 신경 써서 연습해 보세요!

Google Gemini 2.0 Flash 2026-05-12

AI피드백제공을위한프롬프트엔지니어링

아학편 도메인 특수 프롬프트 적용 전

안녕하세요! R 발음 연습하시느라 고생 많으셨어요. L 발음으로 자꾸 바뀌는 실수가 있었네요.

R 발음은 혀를 입천장에 닿지 않게 살짝 말고, '어'와 '으' 중간 발음을 내듯이 굴려주시는 게 중요해요. 마치 강아지가 으르렁거리는 소리처럼요! 다음번엔 혀를 조금 더 안쪽으로 말아 올려서 연습해 보시면 좋을 것 같아요.

아학편 도메인 특수 프롬프트 적용 후

R 발음에서 L 발음으로 대체되는 오류가 잦으셨네요. R 발음을 연습할 때 혀를 입천장 쪽으로 살짝 말아 올리는 느낌을 더 신경 써 보세요.

robber:(으)로버, rice:(으)라이쓰 처럼 R 발음 연습에 집중하면 훨씬 자연스러워질 거예요!

Google Gemini 2.0 Flash 2026-05-12

AI피드백제공을위한프롬프트엔지니어링

LLM 피드백 안정성 평가 방법
한 오류 유형에 대해 동일한 프롬프트로 피드백 5회 산출 후 BERTScore 계산해 피드백 간 유사도 측정
BERTScore 0.75 이상이면 일관된 출력

고민해야 할 문제
step 피드백 / unit 피드백에 아학편 도메인 특수 데이터가 각각 들어가야 하는가?

메인 구현 2

음소인식기 모듈

두 음소 시퀀스 만들기

사용자 발화 처리

사용자 음성

마이크 입력 (16 kHz waveform)

음성 인식 모델

프레임별 음소 분포 산출

중복 프레임을 압축 (CTC 디코딩)

blank 과 sil 토큰 제거

출력 토큰은 ARPAbet 음소 시퀀스

인식 음소

`[ l, ay ]`

실제 발음한 음소

정답 텍스트 처리

정답 텍스트

`Light`

학습 단계의 목표 문장

G2P 변환

CMU 발음 사전을 우선 조회

미등재 단어는 신경망 폴백

스트레스 숫자 제거, 소문자 정규화

단어 사이 공백은 무시

출력 토큰 집합은 인식 모델과 동일

정답 음소

`[ l, ay, t ]`

비교 기준 음소

정렬 비교와 점수 산출

예시: Light 에서 t 누락이 발생한 경우

정답	l	ay	**t**
인식	l	ay	없음
op	=	=	**D**

op 종류

= 매칭 (canonical = perceived)
S 치환 (서로 다른 음소)
I 삽입 (정답에 없는 음소가 인식됨)
D 삭제 (정답 음소가 빠짐)

오류 수 = S + I + D
PER = 오류 수 / 정답 음소 수
PER = 1 / 3 ≈ 0.333
정확도 = (1 − PER) × 100 ≈ 66.7

최종 산출물

인식 음소 시퀀스
perceived ARPAbet
화면 표시와 비교에 사용

자리별 op 배열
C, S, I, D 라벨 시퀀스
음소 정렬 시각화에 사용

PER 과 정확도 점수
0 에서 100 사이 정수
EXP, 통과 판정 입력

음소 단위 오류 목록
PhonemeError 엔티티
한국어 코칭 LLM 입력

원본 데이터셋

AIHub 한국인 영어 발음

한국인 화자가 영어 문장을 낭독한 코퍼스

화자 메타: 성별, 나이, 모국어, 영어 능력 등급

음소 표기: IPA 와 ARPAbet 동시 제공

canonical 과 perceived 정렬에 오류 라벨까지 부여

전체 약 62000 샘플 규모, 학습 데이터의 본체

샘플 항목


```
{
"spk_id": "101868", "gender": "F", "age": 23,
"ability": "HIGH", "duration": 4.87,
"sentence": "It helps the community save many ...",
"canonical_ipa": "ɪ t h ɛ *** l p s ð ə ...",
"perceived_ipa": "ɪ t h ɛ æ l p s ð ə ...",
"canonical_arpabet": "ih t hh eh *** l p s dh ah ...",
"perceived_arpabet": "ih t hh eh ae l p s dh ah ...",
"canonical_aligned": "sil ih t hh eh sil l p s dh ...",
"error_labels_aligned": "C C C C I C C C C C ..." }
```


L2Arctic

비원어민 영어 학습자의 영어 낭독 코퍼스

다양한 모국어 (스페인, 베트남, 아랍 등)

canonical 과 perceived 모두 ARPAbet 정렬 제공

화자별 디렉토리 구조, 한 화자 약 1000 문장

전체 약 3500 샘플 규모, 보조 데이터

샘플 항목


```
{
"spk_id": "ABA",
"wav": "data/l2arctic/ABA/wav/arctic_a0003.wav",
"duration": 4.92,
"wrd": "For the twentieth time that evening ...",
"canonical_aligned":
  "sil f ao r sil dh ah t w eh n t iy ih th ...",
"perceived_aligned":
  "sil f ao r sil dh ah t w eh n t iy ih th k sil ..." }
```

전처리 단계

1 음소 표기 통일

IPA 와 ARPAbet 가 섞인 라벨을 ARPAbet 소문자로 모은다.
스트레스 숫자를 떼고 모델 인벤토리에 있는 음소만 남긴다.

단일 음소 집합으로 두 데이터셋이 같은 사전을 공유한다.

2 앞뒤 sil 추가, 중복 합침

발화 앞뒤에 sil 토큰을 더한다.
연속해 나오는 같은 음소나 sil 은 하나로 합친다.

CTC 디코딩이 연속 동일 토큰을 자동 압축하기 때문.

3 Levenshtein 정렬과 op 라벨링

정답과 인식 음소를 자리별로 짝짓는다.
자리마다 C, S, I, D 라벨 (매칭, 치환, 삽입, 삭제) 을 매긴다.

삽입과 삭제 자리는 *** 로 표시하되 학습 타겟에서는 제거.

4 음소를 정수 ID 로 매핑

phoneme_to_id.json 으로 음소 토큰을 정수 인덱스로 변환.
0번 인덱스는 CTC blank, 1번 인덱스는 sil 토큰.

총 42 개 클래스 (blank + sil + 40 개 ARPAbet 음소).

5 배치 패딩과 길이 전달

가변 길이 오디오와 음소 시퀀스를 배치 안 최대 길이로
0 패딩한다.

CTC 손실이 패딩을 무시하도록 실제 길이를 함께 전달한다.

6 샘플 필터링

오디오 파일 누락 또는 학습 타겟이 빈 항목을 제외한다.
최대 길이 (16 kHz 기준 10 초) 초과 샘플도 제외한다.

학습 인덱스 단계에서 한 번 거르고 그 뒤로는 패딩만 처리.

화자 단위 70 / 15 / 15 분할

전체 통계

총 65,486 샘플, 906 명 화자

AIHub 한국인 영어 발음 61,990 샘플 (95 %)

L2Arctic 3,496 샘플 (5 %)

분할 비율 (샘플 기준)

train.json

70 %

45,804 샘플
634 명

val.json

15 %

9,847 샘플
136 명

test.json

15 %

9,835 샘플
136 명

FiLM 모델 구조

음향 경로

오디오 waveform

`[B, S]` (16 kHz)

Wav2Vec2 + BaseEncoder

사전학습 wav2vec2-base
LayerNorm, Dropout

출력 h [B, T, H=768]

정답 음소 경로

canonical IDs

`[B, L]` (정답 음소 인덱스)

Embedding + Pool + Linear

Embedding (E = 64 / 128 / 512 실험)
Masked Mean Pool, Linear (E to 2H)

출력 (γ, β) [B, H=768]

FiLM 변조

$$h' = (1 + \gamma) \odot h + \beta$$

γ, β 초기값 0 으로 시작 (Identity 시작)

h' [B, T, 768]

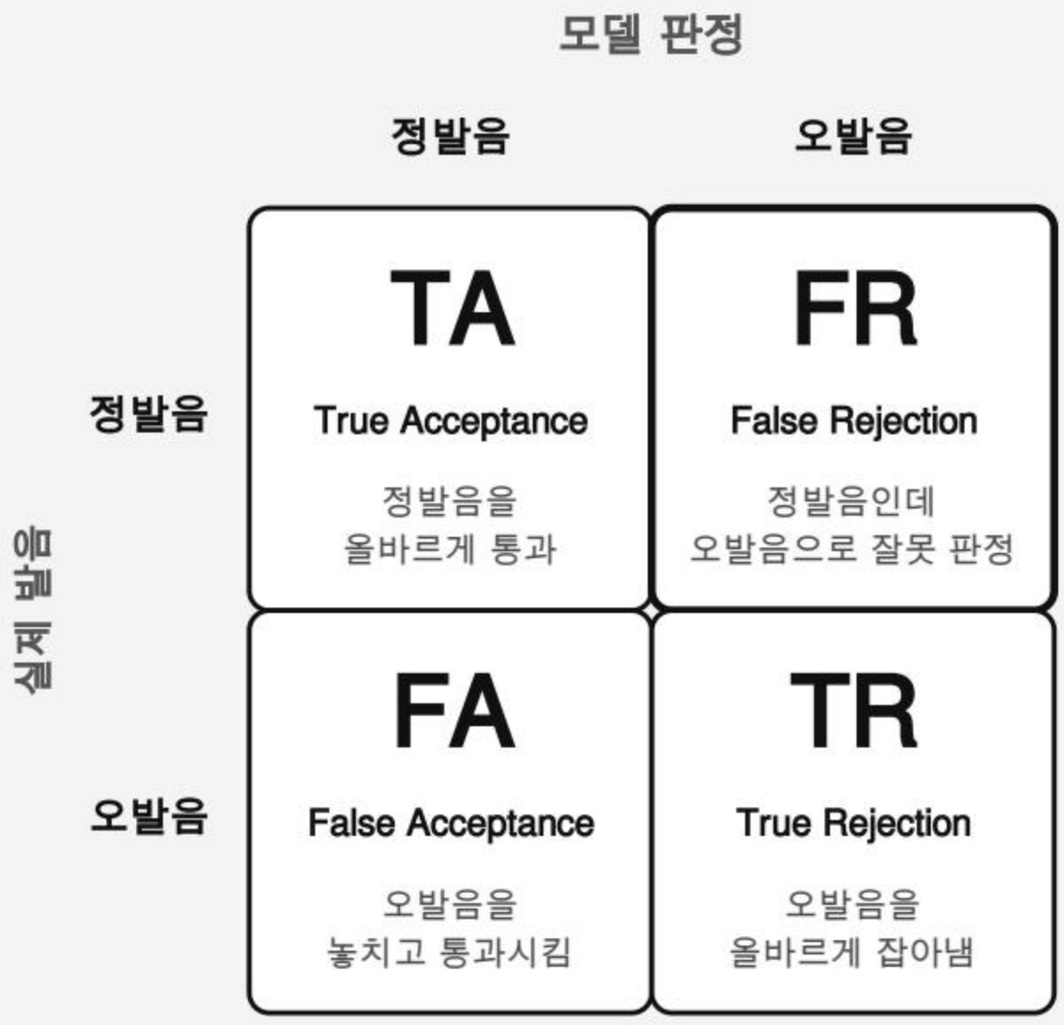
PhonemeHead

Linear, GELU, Linear, CTC 손실로 학습

logits [B, T, V=42]

발음오류 검출 (MDD) 평가 지표

실제 발음과 모델 판정의 2 × 2 분류



지표 정의

FRR = FR / (TA + FR)

정발음이 거부된 비율

FAR = FA / (FA + TR)

오발음이 통과된 비율

Precision = TR / (TR + FR)

"오발음" 판정의 정확도

Recall = TR / (TR + FA)

오발음을 놓치지 않은 비율

MDD F1 = 2 × P × R / (P + R)

Precision 과 Recall 의 조화평균

Table 1: 전체 9,835개 발화에 대한 음소 오류율(PER, %)과 발음오류 검출 성능(MDD F1, %)으로, PER은 **낮을수록 좋고**, MDD F1은 **높을수록 좋음**. FiLM-128이 PER, MDD F1 모두에서 가장 우수한 성능을 보임.

Model	PER (%)	S (%)	D (%)	I (%)	MDD F1 (%)
Baseline	15.93	9.50	3.69	2.75	76.50
FiLM-64	15.56	9.39	**3.69**	2.48	77.00
FiLM-128	**15.39**	**9.19**	3.97	**2.23**	**77.51**
FiLM-512	15.55	9.31	3.82	2.43	77.06

Table 2: 발음오류 검출(MDD) 세부 지표(%)로, Precision/Recall은 **높을수록 좋고**, FRR/FAR은 **낮을수록 좋음**. FRR(False Rejection Rate)은 정발음을 "오발음"으로 잘못 판정한 비율, FAR(False Acceptance Rate)은 오발음을 "정발음"으로 놓친 비율을 의미함.

Model	Precision (%)	Recall (%)	FRR (%)	FAR (%)
Baseline	78.61	74.50	5.83	25.50
FiLM-64	79.33	74.80	5.62	25.20
FiLM-128	79.41	**75.71**	5.67	**24.29**
FiLM-512	**79.50**	74.76	**5.56**	25.24

Table 3: 데이터셋별 전체 지표 종합

Model	Source	PER (%)	MDD F1 (%)	FRR (%)	FAR (%)
Baseline	L2-ARCTIC	16.99	60.34	9.95	28.81
Baseline	AIHub	15.90	76.68	5.76	25.47
FiLM-64	L2-ARCTIC	17.00	59.66	10.74	27.80
FiLM-64	AIHub	15.51	77.20	5.53	25.17
FiLM-128	L2-ARCTIC	16.59	60.51	10.07	28.25
FiLM-128	AIHub	15.35	77.70	5.59	24.26
FiLM-512	L2-ARCTIC	16.46	61.19	9.85	27.68
FiLM-512	AIHub	15.52	77.24	5.48	25.22

Table 4: 데이터셋별 발음오류 검출 F1 점수(MDD F1, %)로, **높을수록 좋음**. L2-ARCTIC은 다국적 화자의 더 다양한 발음 변이로 인해 모든 모델에서 AIHub 대비 약 16%p 낮은 MDD F1을 보인다.

Model	L2-ARCTIC MDD F1 (%)	AIHub MDD F1 (%)
Baseline	60.34	76.68
FiLM-64	59.66	77.20
FiLM-128	60.51	**77.70**
FiLM-512	**61.19**	77.24

Table 5: 데이터셋별 False Rejection Rate(FRR, %)로, **낮을수록 좋음**. FRR은 *정발음을 모델이 "오발음"으로 잘못 판정한 비율*로, 사용자 경험에 직접적인 영향을 미친다. L2-ARCTIC에서 FRR이 약 10%로 AIHub(약 5.6%) 대비 거의 두 배에 달해, 다국적 화자의 발음 변이를 모델이 잘못된 발음으로 거부하는 경향이 두드러진다.

Model	L2-ARCTIC FRR (%)	AIHub FRR (%)
Baseline	9.95	5.76
FiLM-64	10.74	5.53
FiLM-128	10.07	5.59
FiLM-512	**9.85**	**5.48**

Table 6: 데이터셋별 False Acceptance Rate(FAR, %)로, **낮을수록 좋음**. FAR은 *오발음을 모델이 "정발음"으로 놓친 비율*로, 평가의 엄격성과 관련됨. 두 데이터셋 모두 FAR이 25% 이상으로 오발음 검출이 어려운 영역임을 시사한다.

Model	L2-ARCTIC FAR (%)	AIHub FAR (%)
Baseline	28.81	25.47
FiLM-64	27.80	25.17
FiLM-128	28.25	**24.26**
FiLM-512	**27.68**	25.22

Table 8: AIHub 학습자 능력 등급별 PER(%)로, **낮을수록 좋음**. 능력 등급(HIGH/MID/LOW)이 낮을수록 발음 변이가 커져 PER이 증가하며, LOW 학습자는 HIGH 대비 약 2배의 PER을 보이는 점에서 학습자 수준에 대한 모델의 격차가 존재함.

Model	HIGH (n=1,305)	MID (n=4,148)	LOW (n=3,946)
Baseline	9.50	15.28	18.79
FiLM-64	9.39	14.96	18.22
FiLM-128	**9.20**	**14.72**	**18.16**
FiLM-512	9.38	14.94	18.28

잡음 강건성 평가 방법

잡음 데이터 (MUSAN)

MUSAN 환경 잡음 코퍼스

실내 잡음, 기계 소음, 군중 소음, 자연 소음 등 환경 잡음 모음

774 개 파일, 약 7 시간

train 619 / val 77 / test 78 로 분할

평가용 잡음 (test split)

test split 에 속한 78 개 잡음 파일을 사용한다.

발화별 잡음 파일을 시드로 고정해 합성한다.

평가 절차

학습자 발화
원본 음성

→

잡음 합성
`y = speech + α * noise`
α : 목표 SNR (dB) 기준 스케일

→

잡음 섞인 발화
동일 발화의 노이즈 버전

→

모델 평가
PER, MDD F1, FRR, FAR

SNR 단계

Clean	20 dB	15 dB	10 dB	5 dB
잡음 없음	조용한 실내 수준	일상 사무실 수준	카페, 거리 수준	시끄러운 환경

Table 9: 잡음 환경에서의 음소 오류율(PER, %)로 **낮을수록 좋음**. 그러나 SNR이 낮아질수록 모든 모델의 PER이 급격히 증가하는 것을 볼 수 있음.

Model	Clean	20dB	15dB	10dB	5dB
Baseline	15.93	18.87	21.80	27.40	37.02
FiLM-64	15.56	18.66	21.51	26.88	36.24
FiLM-128	**15.38**	**18.42**	**21.35**	**26.77**	36.23
FiLM-512	15.55	18.77	21.72	27.08	**36.02**

Table 10: 잡음 환경에서의 발음오류 검출 F1 점수(MDD F1, %)로, **높을수록 좋음**.

Model	Clean	20dB	15dB	10dB	5dB
Baseline	76.49	73.16	70.11	64.84	57.16
FiLM-64	76.99	73.32	70.28	65.21	**57.79**
FiLM-128	**77.50**	**73.95**	**70.79**	**65.39**	57.63
FiLM-512	77.05	73.09	69.89	64.65	57.57

Table 12: 환각 음소 삽입률(Insertion Rate, %)으로, **낮을수록 좋음**. 잡음으로 인해 발화하지 않은 음소를 모델이 잘못 삽입하는 비율을 의미함. **FiLM-128이 잡음 환경에서도 환각 발생을 가장 잘 억제한다.**

Model	Clean	20dB	15dB	10dB	5dB
Baseline	2.75	2.83	2.87	2.92	2.88
FiLM-64	2.47	2.43	2.43	2.50	2.61
FiLM-128	**2.22**	**2.24**	**2.17**	**2.08**	**2.03**
FiLM-512	2.42	2.29	2.26	2.27	2.44

메인 구현

변경 이력

설계서 변경 이력

달성 목표 ④ 잡음 환경 환각 음소 억제 (신규 추가)

변경 전	변경 후
잡음 강건성 관련 목표 없음	**5dB 환각 음소 삽입률 2.88% → 2% 이하** 현 FiLM-128 2.03% 달성

사유 잡음 환경에서 발화하지 않은 음소를 모델이 잘못 검출하는 현상을 별도 지표로 추적.

아학편 기반 피드백 보완 (한국어 부재 음소 처리)

변경 전	변경 후
/θ/ 등 모든 음소를 아학편 합용병서로 표기	**한국어 부재 음소는 보완 표기 / 설명 적용** 합용병서 의존 축소

사유 벤치마크 평가에서 /θ/ 항목 점수 최저 (4.06 / 3.90 / 3.50). 합용병서 표기가 학습자 피로감을 유발.

테스트 시나리오 합격 기준 구체화 (TC-04, TC-09 / TC-10)

변경 전	변경 후
TC-04 : 입력 배경 잡음, 출력 분석 성공 / 실패 TC-09 / 10 : ?지표 OO 이상 → 정답처리	**TC-04 : 10dB SNR 합성, 환각 삽입률 ≤ 2%** **TC-09 / 10 : 학습 단계 정답률 ≥ 80% → 통과**

사유 모호한 판정 기준 ("성공 / 실패", "?지표") 을 측정 가능한 임계값으로 교체.

TC-18 시각 자료 피드백 (신규 추가)

변경 전	변경 후
시각 보조 자료 첨부 시나리오 없음	**필요한 음소는 텍스트 피드백 + 관련 이미지 표시** 혀 위치, 입 모양 등 시각 보조

사유 음성학 설명만으로 직관적 이해가 어려운 음소(/θ/ 등) 에 시각 보조를 더해 학습 효율을 보완.

달성 목표

달성 목표 ① 인식기 FRR 저감

베이스라인

5.83%

현재 측정값

→

목표

4%대

FiLM Canonical Injection

자체 구축 데이터셋에서 평가

달성 목표 ② 피드백 평가 벤치마크 구축

발음 오류 케이스	비교 시스템	한국인 평가자
50 개	**4 가지**	**N 명**

비교 시스템

IPA 설명	음소 매핑	제로샷 LLM	아학편 기반
음성학 용어로 기호 차이 안내	스픽식 단순 지시 (혀 위치 등)	프롬프트 없이 일반 코칭 응답	한국어 음운 체계로 설명 (본 프로젝트)

달성 목표 ③ 테스트 시나리오 전부 통과

테스트 케이스	검증 영역	목표 통과율
18 개	5 영역	100 %

검증 영역

발음 인식

정상 발음
오발음
묵음

피드백 교정

모음 구분
연음 처리
반복 오류 추적

사용자 적응

추천 학습
맞춤 학습
신규 온보딩

엣지 케이스

소음 환경
마이크 권한
네트워크 끊김

학습 데이터

이력 저장
통계 리포트

달성 목표 ④ 잡음 환경 환각 음소 억제

베이스라인

2.88%

5dB 잡음, 현재 측정값

→

목표

2% 이하

FiLM-128 적용, 현 2.03% 달성

잡음 강건성 측정 조건

MUSAN test split 잡음을 5dB SNR 로 합성한 발화에서 평가
환각 음소 = 발화하지 않은 음소를 모델이 잘못 검출한 것

100 개 음소 중 약 2 개 미만이 잘못 추가되는 수준

테스트 시나리오

테스트 시나리오 (1 / 3)

ID	항목	입력	기대 출력	통과
TC-01	정상 발음 인식	"apple" 을 정확히 발음	정확도 90% 이상, 통과 표시	●
TC-02	오발음 감지	"apple" 을 /æ/ 대신 /ɑ/ 로 발음	오류 감지 후 교정 피드백 제공	●
TC-03	묵음 처리	"knight" 의 k 발음	묵음 처리 올바른지 확인	○
TC-04	**소음 환경 환각 억제** (목표 ④ 와 연결)	10dB SNR 잡음 합성 발화	**환각 음소 삽입률 ≤ 2%**	○
TC-05	무음 입력	아무 말 없이 녹음 종료	"다시 시도해보세요" 안내 메시지	○
TC-06	모음 구분	/iː/ vs /ɪ/ (sheep vs ship)	i 발음법 안내 및 예시 제공	○

● 통과 ○ 미통과 / 검증 예정

진하게 표시된 행은 변경 이력 반영 항목

테스트 시나리오 (2 / 3)

ID	항목	입력	기대 출력	통과
TC-07	연음 처리	"want to" → "wanna" 허용 여부 정책 적용	정책대로 처리, 발음별 정확도 추이 그래프	○
TC-08	반복 오류 추적	같은 발음을 n 회 연속 틀림	맞춤 LLM 피드백 제공 (조선시대 영어교재 아학편 기반)	●
TC-09	추천 학습 사용자	**추천 학습 단계 정답률 ≥ 80%**	추천 학습 단계 통과 처리	●
TC-10	맞춤 학습 사용자	**맞춤 학습 단계 정답률 ≥ 80%**	맞춤 학습 단계 통과 처리	●
TC-11	신규 사용자 온보딩	최초 실행	학습 목록 표시	●
TC-12	한국어 억양 간섭	한국어식 영어 발음 입력	한국인 공통 오류 패턴 피드백	○

● 통과 ○ 미통과 / 검증 예정

진하게 표시된 행은 변경 이력 반영 항목

테스트 시나리오 (3 / 3)

ID	항목	입력	기대 출력	통과
TC-13	매우 빠른 발음	정상 속도의 2 배로 발음	속도 경고 또는 정상 처리	○
TC-14	마이크 권한 거부	마이크 접근 차단 상태	권한 요청 안내 화면 표시	●
TC-15	네트워크 끊김	오프라인 상태에서 분석 요청	오프라인 모드 안내 또는 캐시 처리	○
TC-16	학습 이력 저장	세션 종료 후 재접속	이전 진도, 오류 이력 유지	●
TC-17	통계 리포트	7 일 사용 후 리포트 열람	출석 달력, 틀린 발음 순위표 제공	●
TC-18	**시각 자료 피드백** (신규 추가)	시각 보조가 필요한 음소 (/θ/, /r/ vs /l/ 등)	**텍스트 피드백 + 관련 이미지** (혀 위치, 입 모양)	○

● 통과 ○ 미통과 / 검증 예정

진하게 표시된 행은 변경 이력 반영 항목

메인 구현

일정

간트차트

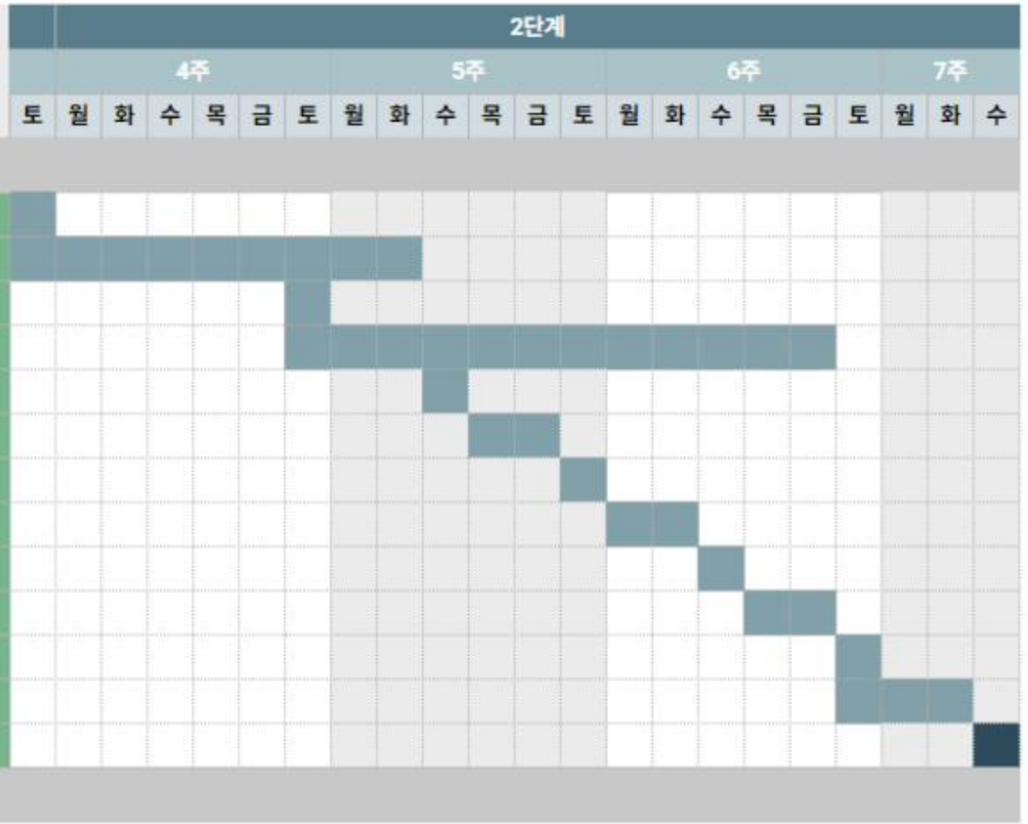
종합설계 간트차트

프로젝트 이름	ECHO! 프로젝트	**팀명**	O(1)
프로젝트 관리자	신영환, 김형준, 김지용, 백지영, 이찬영, 최현지	**날짜**	26년 3월 4일 ~ 26년 6월 10일

WBS 번호	작업 제목	시작일	마감일	기간	작업 완료 비율	1단계														
						1주						2주						3주		
						월	화	수	목	금	토	월	화	수	목	금	토	월	화	수
1	**프로젝트 구상**																			
1	팀 구성	26년 3월 4일	26년 3월 5일	2	100%			■	■											
2	프로젝트 아이디어 제안	26년 3월 7일	26년 3월 7일	1	100%						■									
3	현황 조사	26년 3월 9일	26년 3월 11일	3	100%							■	■	■						
4	정기회의	26년 3월 14일	26년 3월 14일	1	100%												■			
5	제안서 작성	26년 3월 14일	26년 3월 14일	1	100%												■			
6	PPT 제작	26년 3월 15일	26년 3월 17일	3	100%													■	■	
7	프로젝트 제안 발표	26년 3월 18일	26년 3월 18일	1	100%															■

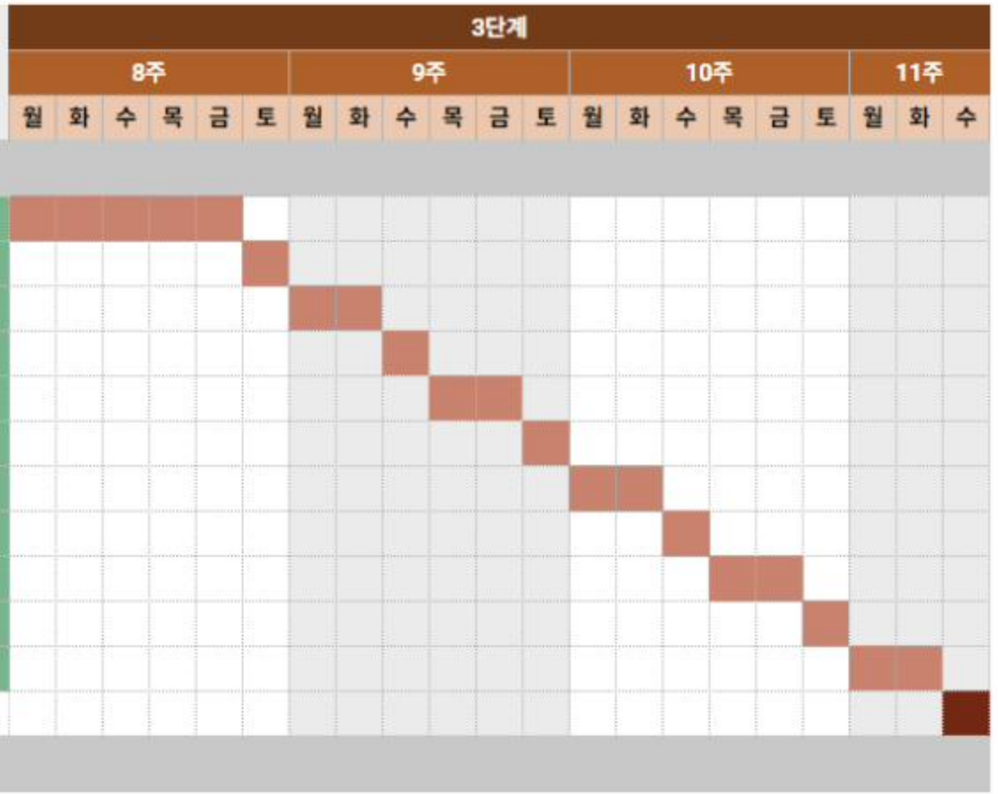
간트 차트

WBS 번호	작업 제목	시작일	마감일	기간	작업 완료 비율
2	**프로젝트 설계**				
1	정기회의	26년 3월 21일	26년 3월 21일	1	100%
2	데이터 전처리, 기능 설계	26년 3월 21일	26년 3월 31일	11	100%
3	정기회의	26년 3월 28일	26년 3월 28일	1	100%
4	모델 학습	26년 3월 28일	26년 4월 10일	14	100%
5	팀활동 - 현황 점검	26년 4월 1일	26년 4월 1일	1	100%
6	아학편 정리, 백엔드 구상	26년 4월 2일	26년 4월 3일	2	100%
7	정기회의	26년 4월 4일	26년 4월 4일	1	100%
8	프론트 수정, 백엔드 설계	26년 4월 5일	26년 4월 7일	3	100%
9	팀활동	26년 4월 8일	26년 4월 8일	1	100%
10	PPT 구상, 백엔드와 모델 구성	26년 4월 9일	26년 4월 10일	2	100%
11	정기회의	26년 4월 11일	26년 4월 11일	1	100%
12	발표 준비	26년 4월 11일	26년 4월 14일	4	100%
13	프로젝트 설계 발표	26년 4월 15일	26년 4월 15일	1	100%



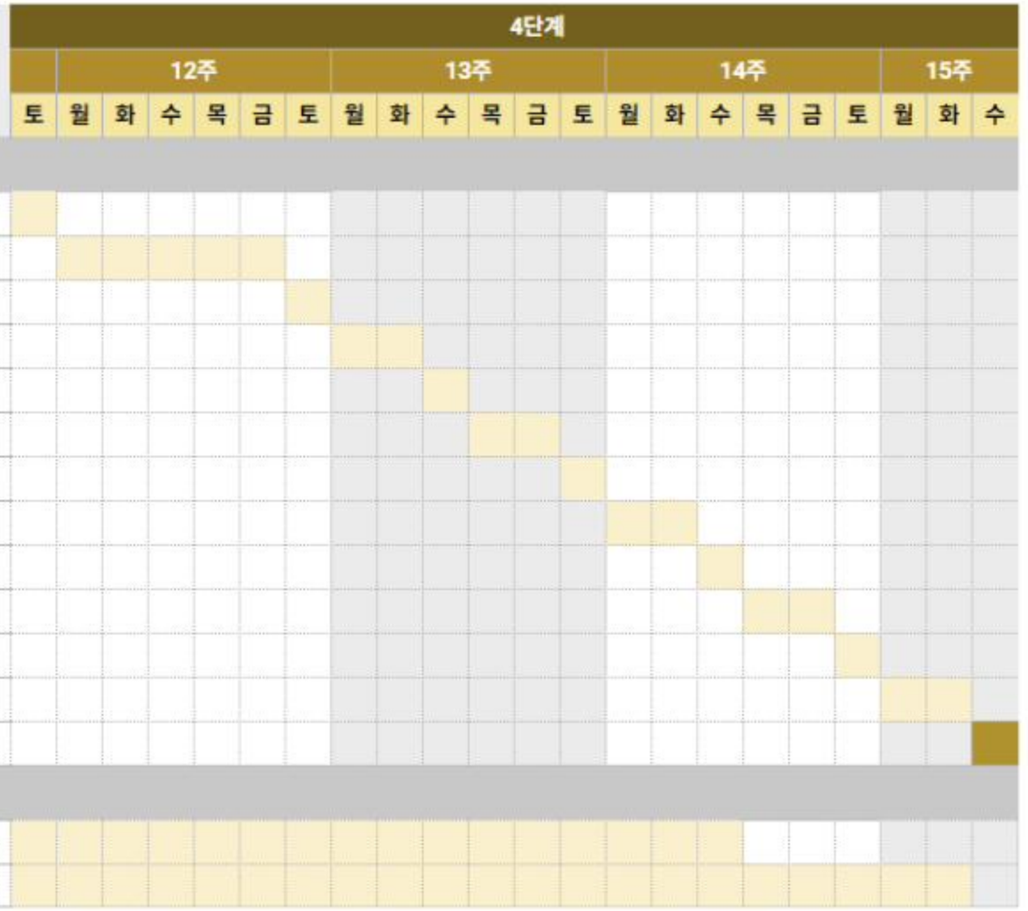
간트 차트

WBS 번호	작업 제목	시작일	마감일	기간	작업 완료 비율
3	**프로젝트 중간 보고**				
1	피드백 정리, 아학편 데이터 정리	26년 4월 20일	26년 4월 24일	5	100%
2	정기회의	26년 4월 25일	26년 4월 25일	1	100%
3	아학편 데이터 마무리	26년 4월 26일	26년 4월 28일	3	100%
4	팀활동	26년 4월 29일	26년 4월 29일	1	100%
5	프론트 QA	26년 4월 30일	26년 5월 1일	2	100%
6	정기회의	26년 5월 2일	26년 5월 2일	1	100%
7	백엔드 - 프론트 연결	26년 5월 3일	26년 5월 5일	3	100%
8	팀활동 (LLM 연결, 아학편 데이터셋)	26년 5월 6일	26년 5월 6일	1	100%
9	프롬프트, 데이터 완성, QA 수정	26년 5월 7일	26년 5월 8일	2	100%
10	정기회의	26년 5월 9일	26년 5월 9일	1	100%
11	발표 준비	26년 5월 10일	26년 5월 12일	3	100%
12	프로젝트 중간 발표	26년 5월 13일	26년 5월 13일	1	0%



간트 차트

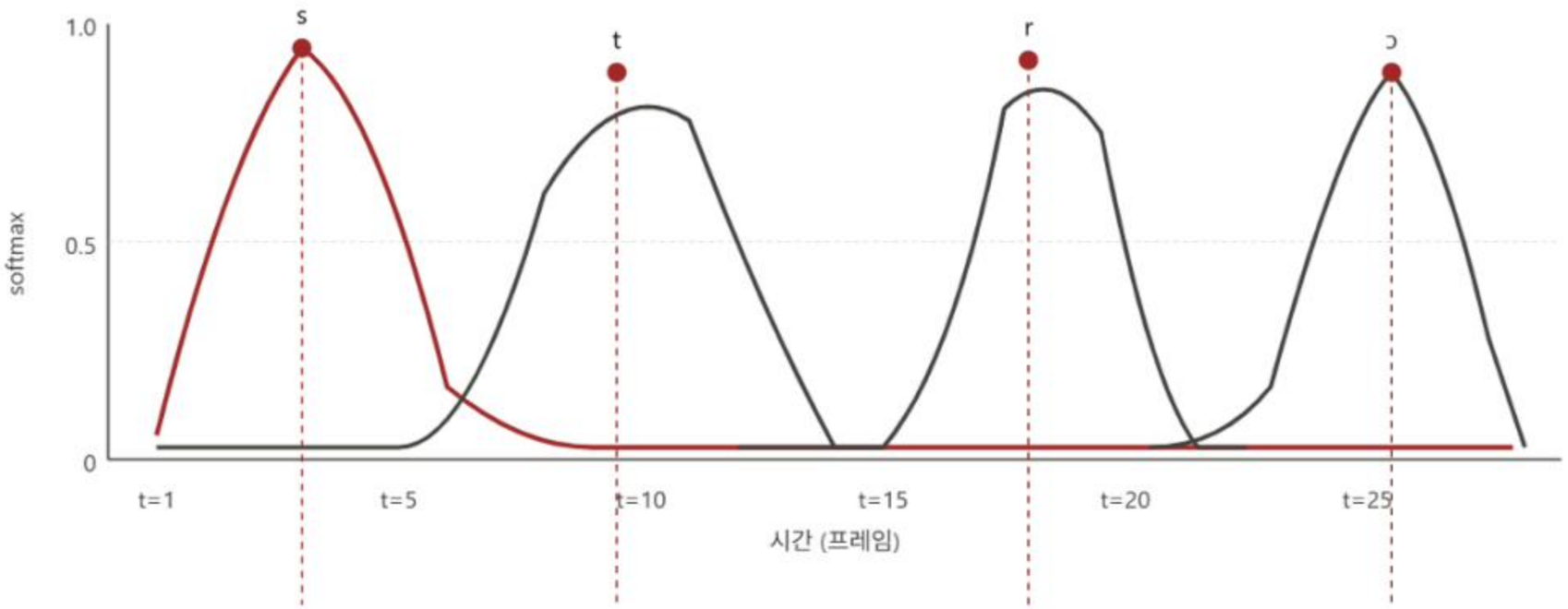
WBS 번호	작업 제목	시작일	마감일	기간	작업 완료 비율
4	**프로젝트 최종 정리**				
1	정기회의	26년 5월 16일	26년 5월 16일	1	0%
2	프론트 디자인 마무리	26년 5월 17일	26년 5월 22일	6	0%
3	정기회의	26년 5월 23일	26년 5월 23일	1	0%
4	신규 유저 모집	26년 5월 24일	26년 5월 26일	3	0%
5	팀활동	26년 5월 27일	26년 5월 27일	1	0%
6	유저 가입 및 사용 테스트	26년 5월 28일	26년 5월 29일	2	0%
7	정기회의	26년 5월 30일	26년 5월 30일	1	0%
8	수정사항 개선	26년 5월 31일	26년 6월 2일	3	0%
9	추가 회의	26년 6월 3일	26년 6월 3일	1	0%
10	배포 준비	26년 6월 4일	26년 6월 5일	2	0%
11	정기회의	26년 6월 6일	26년 6월 6일	1	0%
12	배포 및 발표 준비	26년 6월 7일	26년 6월 9일	3	0%
13	프로젝트 최종 발표	26년 6월 10일	26년 6월 10일	1	0%
장기	tracks.json 개선	26년 5월 16일	26년 6월 3일	~	0%
장기	잡음 이슈 해결	26년 5월 16일	26년 6월 9일	~	0%



Q&A

CTC 디코딩과 peak

프레임별로 각 음소에 대한 softmax 확률이 나온다



각 음소의 peak 위치에서 확률을 취해 시퀀스를 만든다

s	t	r	ɔ
peak 0.94	peak 0.91	peak 0.93	peak 0.92

peak 사이 구간은 blank 또는 연속 중복 · CTC 디코딩에서 제거

랭킹 점수 계산

학습자가 strawberry 를 발음

음소	s	t	r	ɔ	b	ɛ	r	i
softmax	0.93	—	0.88	0.81	0.95	—	0.86	0.91
판정	○	×	○	○	○	×	○	○

정확도

정답 음소 수 / 총 음소 수

6 / 8

= 0.75

확신도

정답 음소 위치 softmax 평균

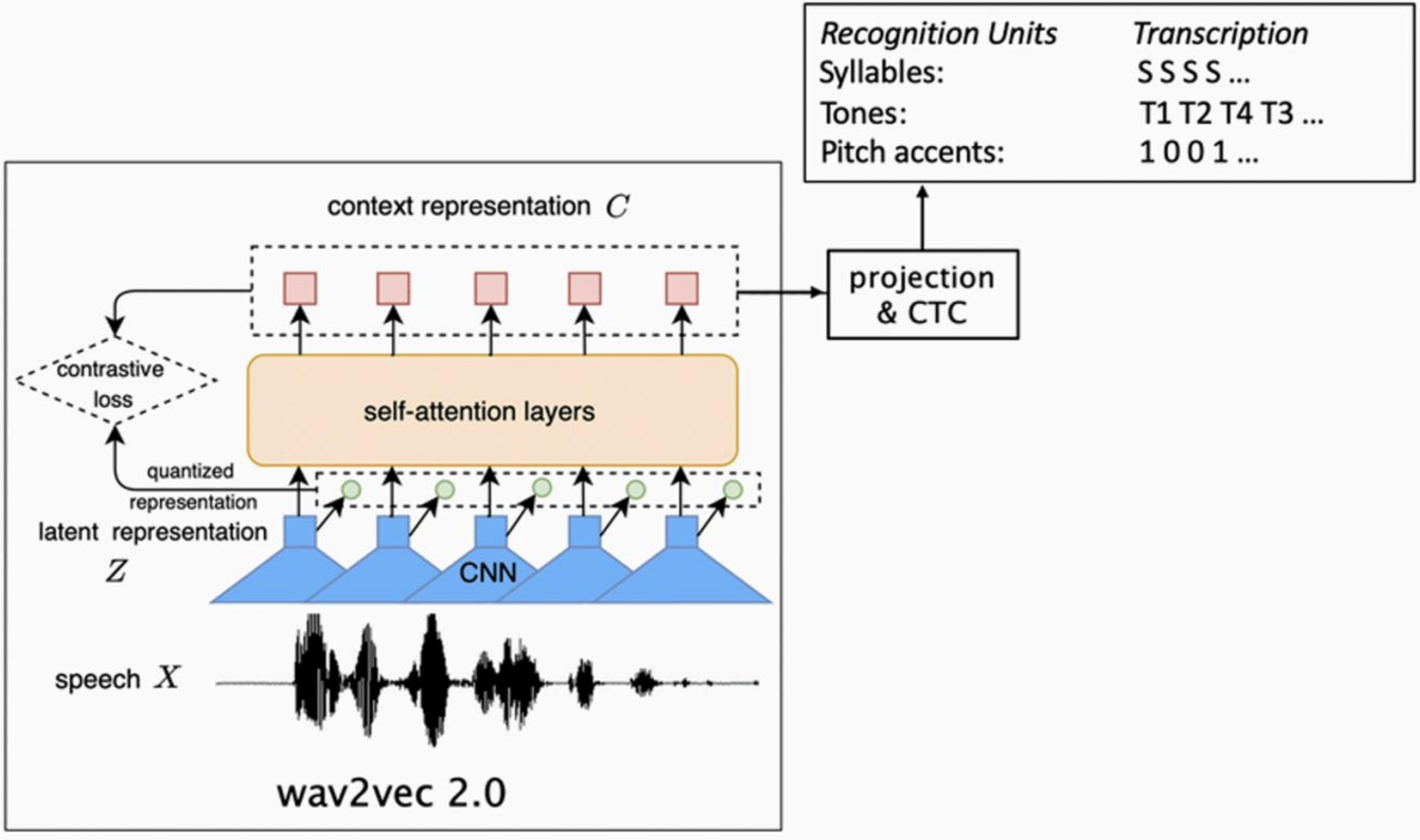
(0.93 + 0.88 + ...) / 6

= 0.89

최종 랭킹 점수

0.75 × 0.89

= 0.67


Recognition Units
Transcription
Syllables:
S S S S ...
Tones:
T1 T2 T4 T3 ...
Pitch accents:
1 0 0 1 ...
context representation C
projection & CTC
contrastive loss
self-attention layers
quantized representation
latent representation Z
CNN
speech X
wav2vec 2.0

용어 정리

분야

CAPT	Computer-Assisted Pronunciation Training · 컴퓨터 보조 발음 학습
MDD	Mispronunciation Detection and Diagnosis · 오발음 검출과 진단
L1 · L2	학습자의 모국어 · 학습 대상 제2언어

평가 지표

PER	Phoneme Error Rate · 음소 인식 오류율
MDD F1	오류 탐지의 정밀도와 재현율을 합친 종합 점수
FRR	False Rejection Rate · 정답 발화를 오류로 잘못 판정한 비율
FAR	False Acceptance Rate · 오발음을 정답으로 놓친 비율

모델

Wav2Vec2	Facebook AI 의 자기지도 학습 기반 음성 표현 모델
CTC	Connectionist Temporal Classification · 정렬 없이 시퀀스를 학습하는 손실 함수
FiLM	Feature-wise Linear Modulation · $h' = \gamma \cdot h + \beta$ 형태의 조건부 변환
canonical · perceived	정답 음소 · 학습자가 실제로 발음한 음소