졸업작품1

Final Presentation

4분반 4조 김세중 송윤수 이지해

1. Motivation

[청세] '발표 울렁증' 호소하는 대학생이 늘었다? 왜

구현주 국민대 사회학과 학생 | 입력 2022.03.17 12:06 | 수정 2022.03.17 15:32 | 댓글 0

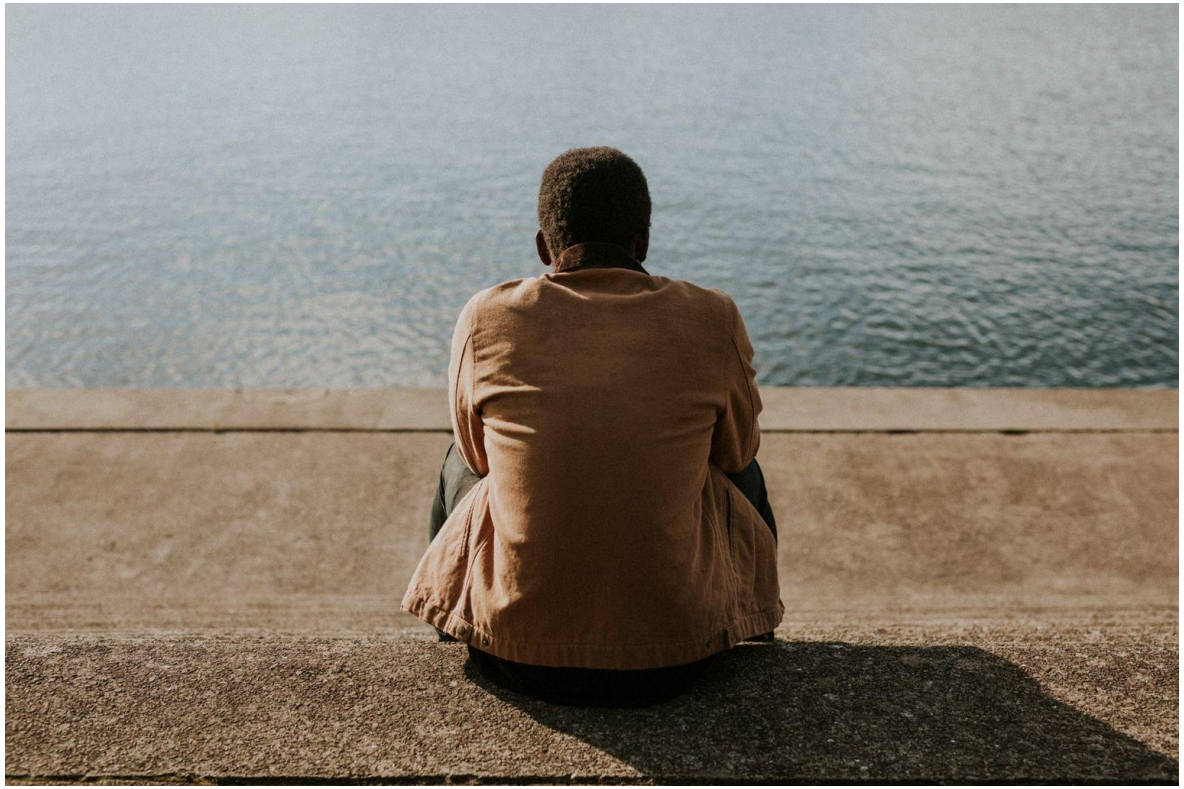
> [청년이 본 세상]
> 코로나 장기화로 비대면 강의 지속이 원인
> "시선 집중, 완벽주의 강박, 소통 어려움"

https://www.womaneconomy.co.kr/news/articleView.html?idxno=210562

Child robot of JST, Japan Science and Technology agency
https://www.youtube.com/watch?v=SE2VCwYDjx0

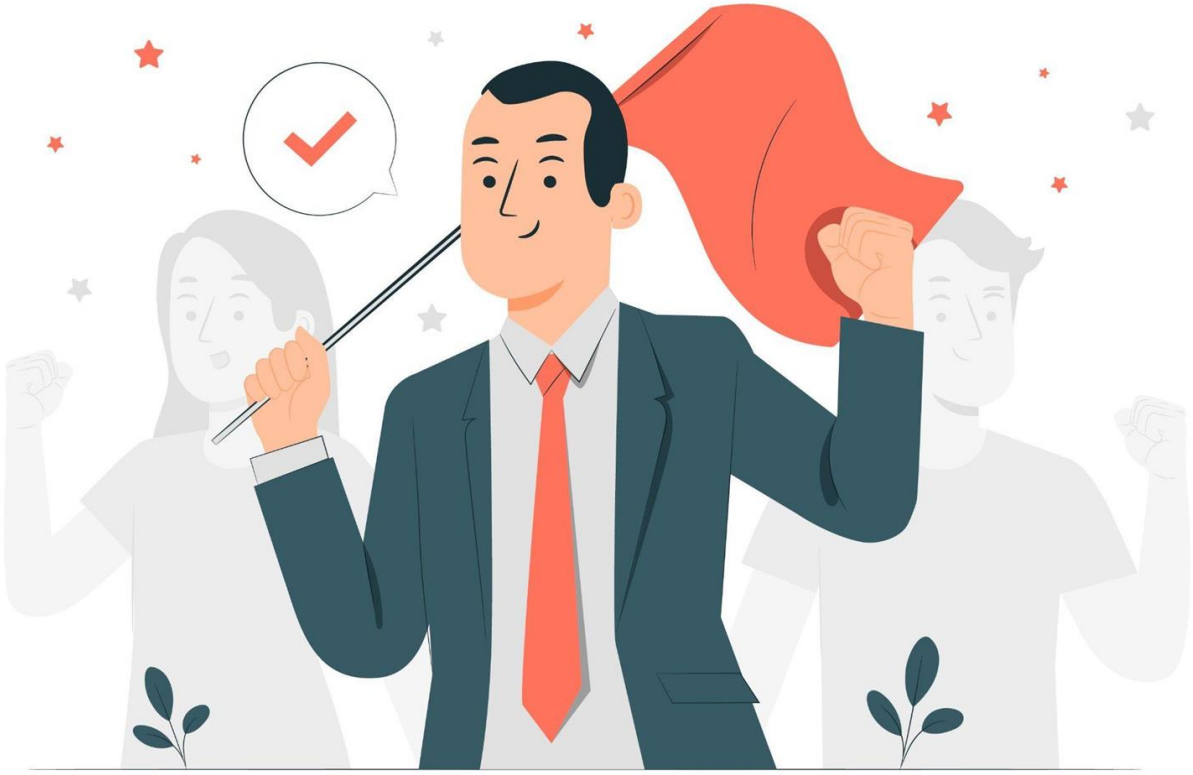
- Everyone knows the importance of presentation.
- But most people are afraid of public speaking.

1. Motivation



- Beginners frequently encounter various difficulties while practicing independently, such as:
  - Is the presentation suitable for the topic?
  - Is the pronunciation clear and accurate?
  - Are my body language natural?
  - And so on.

1. Motivation



We will provide services that are helpful to them!

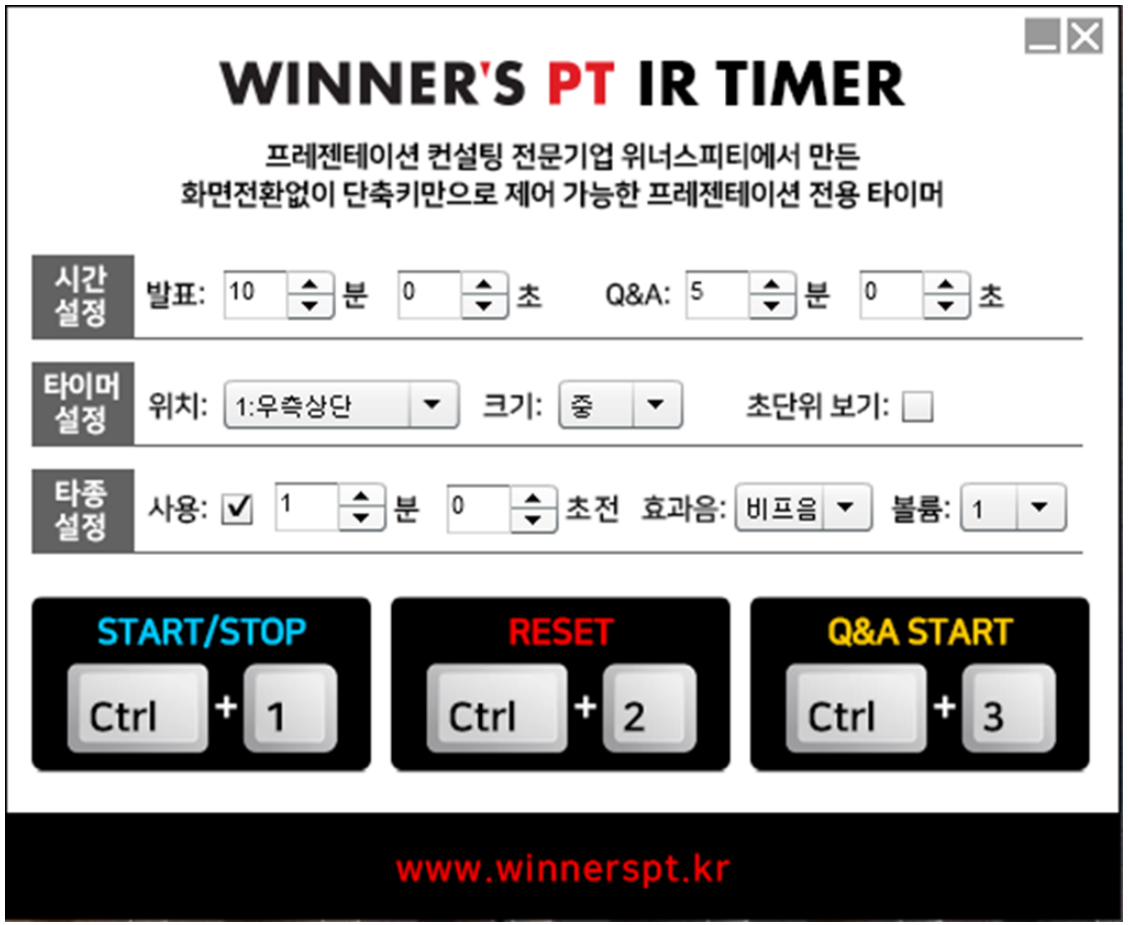
2. Pre-developed technology

CLOVA Note

- Clova notes convert speech into text and analyze keywords.
- In the case of Clova Note, text and keywords are the only purpose of the application.
- However, our project use that function as one of the many functions for presentation analysis.

2. Pre-developed technology

- WINNER'S PT



- Winners PT provides a timer function to check the presentation time.

- The direction of presentation practice is the same as ours, but it simply provides a timer

2. Pre-developed technology

- Samsung #Be fearless of public speaking VR

- This application provides the experience of presenting directly in a virtual environment.
- You can actually have the same experience as presenting in public. Also, You can experience coping with various situations.
- However, it simply helps you have a presentation experience. There is a difference from our application in that it does not analyze actual presentations and present improvements directly.

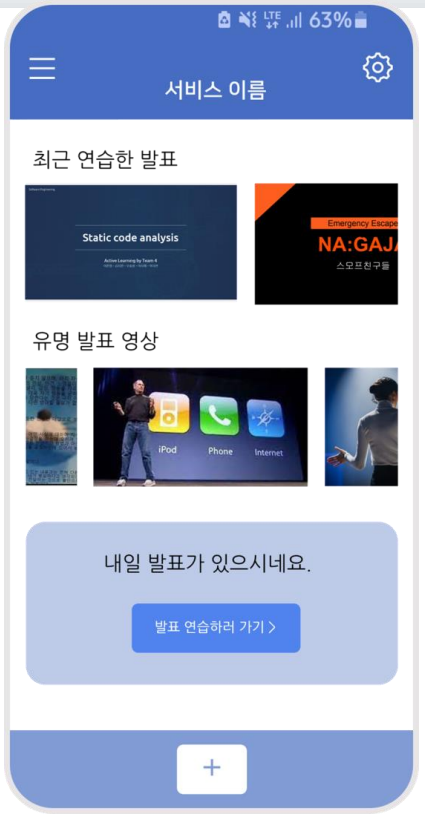
3. Contents

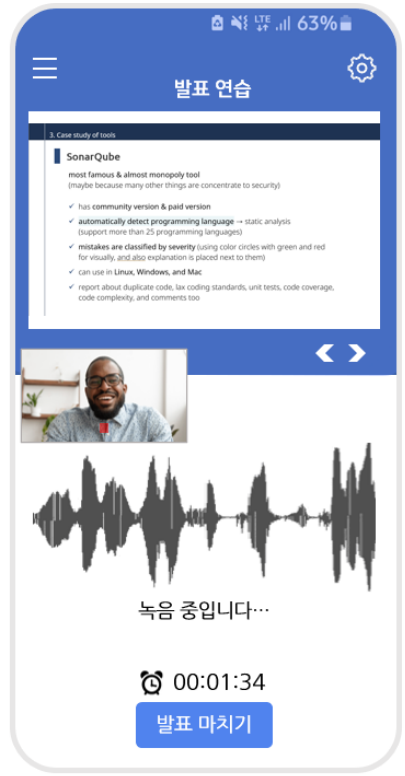
The application for practicing presentations

1) Would be helpful for everyone from people who are afraid of giving presentations to those who simply want a quick practice before a presentation.

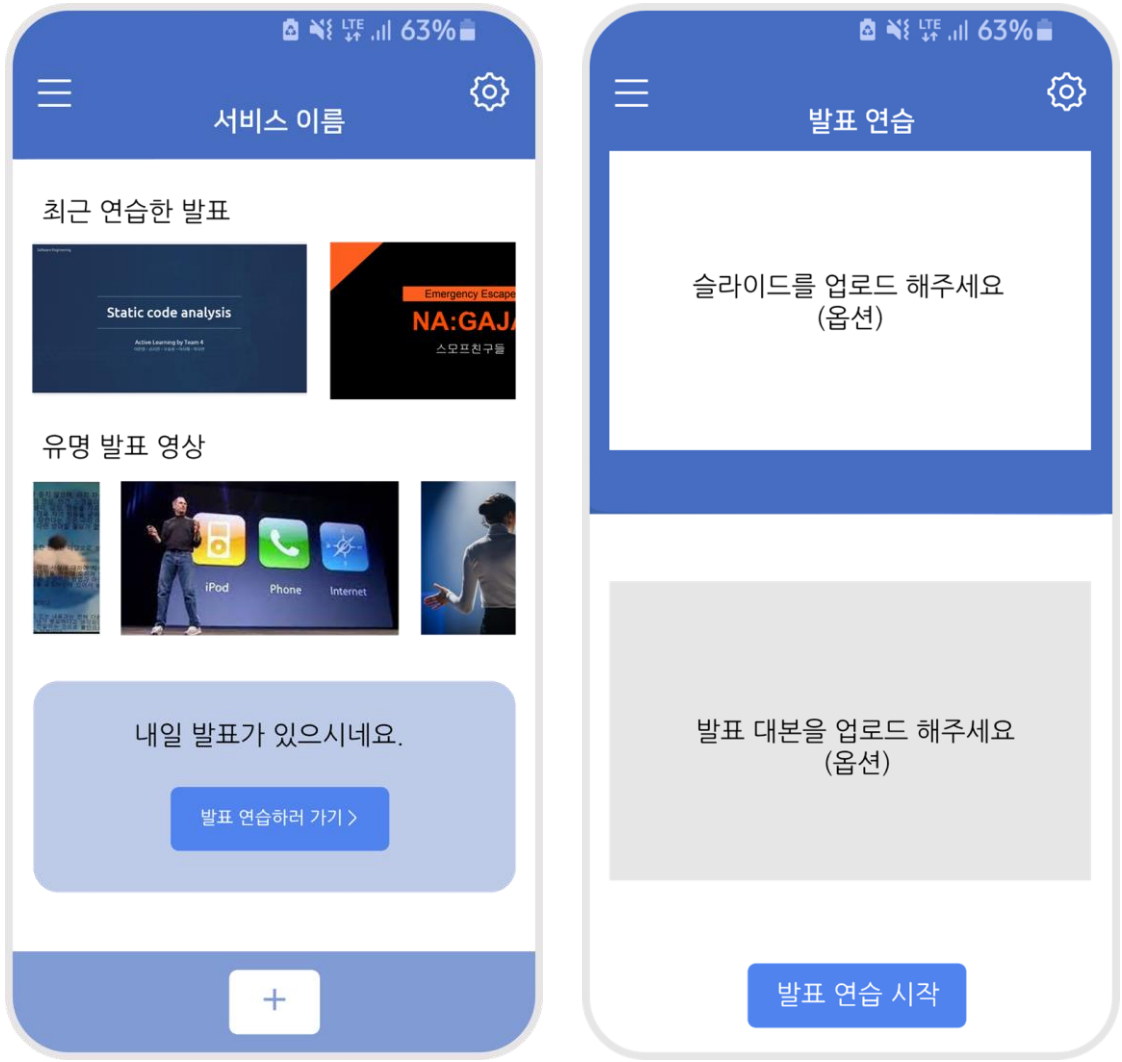
2) Would be easy to use for everyone.

3) Would provide objective and meaningful analysis.





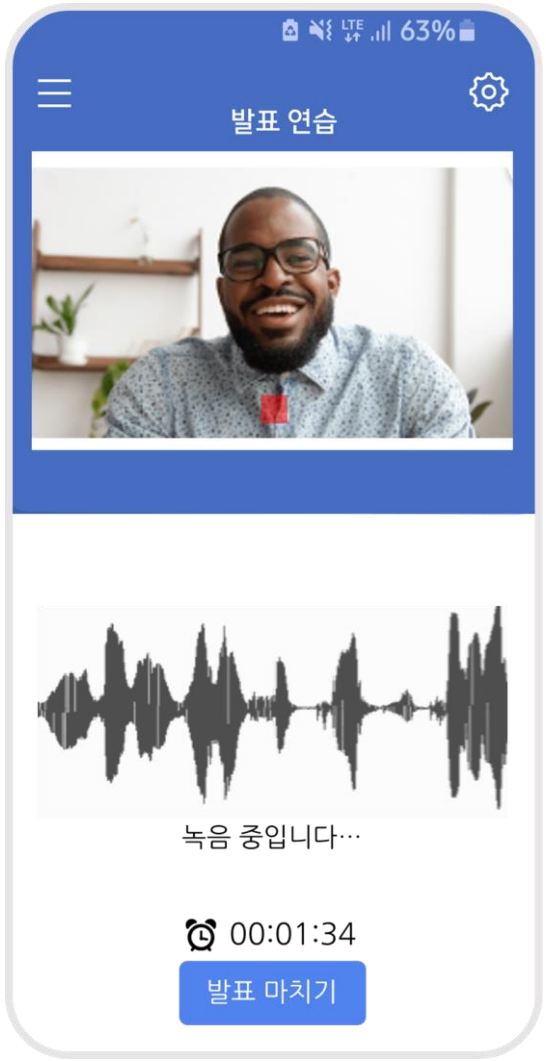
3. Contents

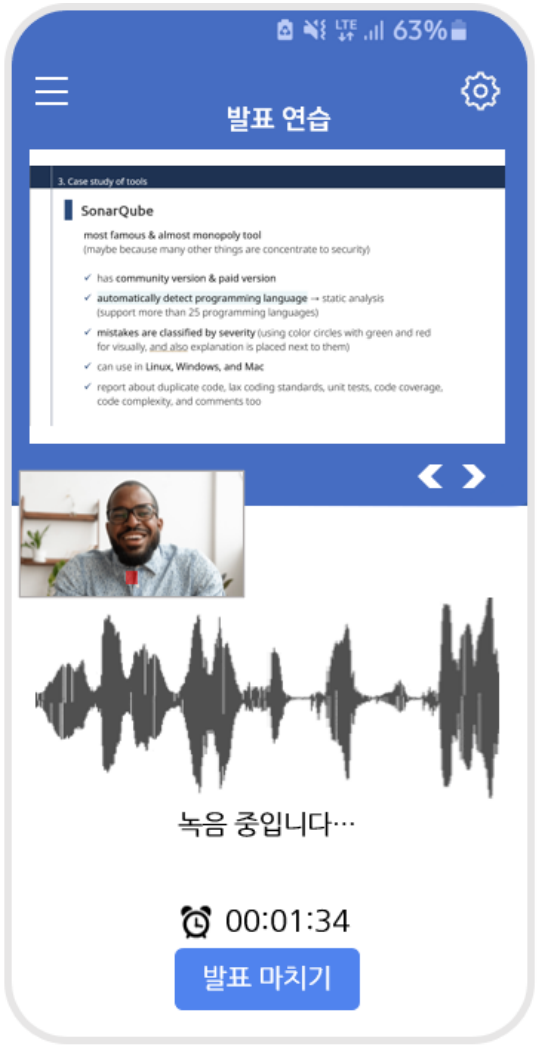


Easy start

For more detailed analysis and practice, users can upload PPT slides and scripts.

They can save recent practice presentations,
so they can practice immediately without additional input.

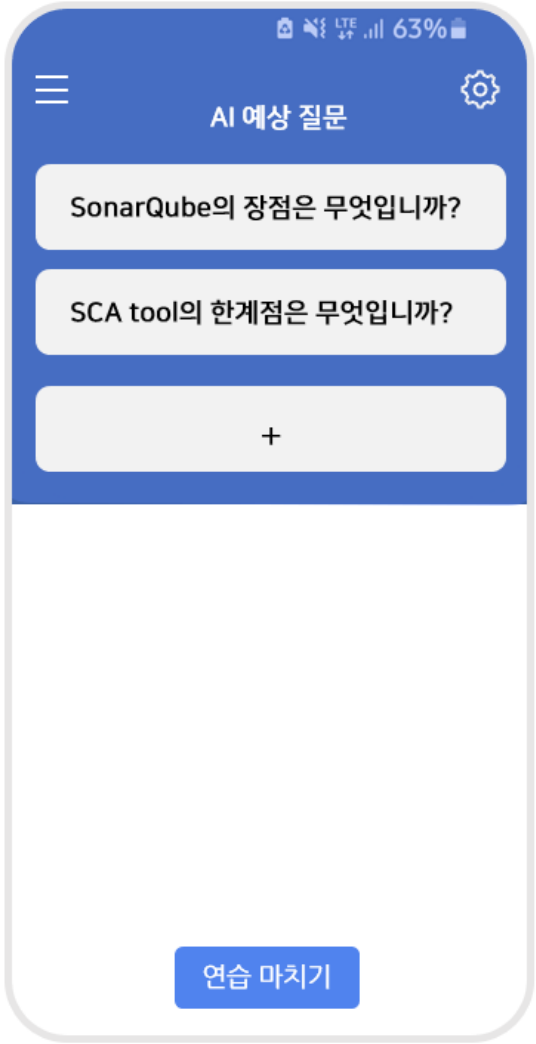




Practice time

It takes video and audio input from the user during practice.

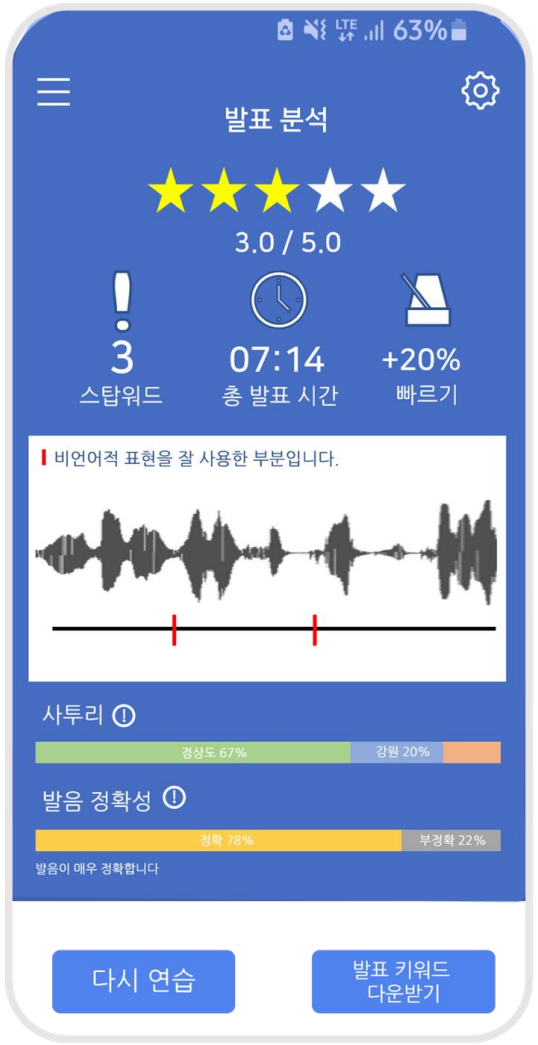
If the user has uploaded PPT slides, it displays the slides alongside.

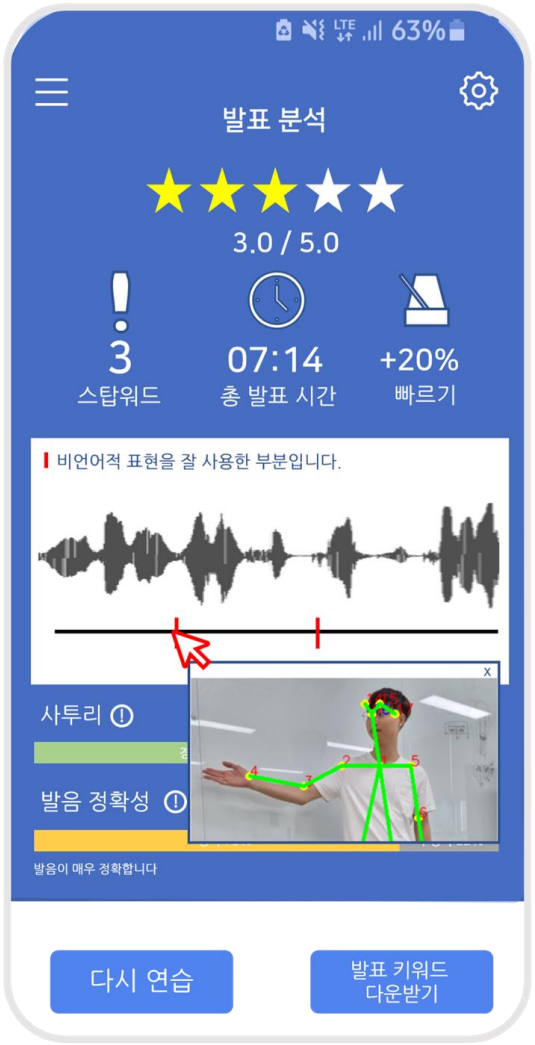


Dealing with questions

Practice responding to questions provided by generative AI.

3. Contents





Presentation Analysis

It shows how often you use filler words (um, uh, etc.)

It shows how fast you speak compared to a normal speaker.

It shows whether you use gestures well.

It indicates whether you use dialects or whether your pronunciation is accurate.

Presentation Analysis

It shows the key words extracted from the user's practice presentation.
Users can check for themselves whether these are the points they wanted to emphasize.
If the points that user wanted to emphasize were not highlighted, users can correct this in their next practice.

If a PPT and script were inputted, it displays the key words extracted from each slide based on the PPT and script.
It also indicates whether these keywords were emphasized in the user's presentation.

Summary

- Speech to Text
  - ASR API
- Slide to Text
  - OCR
- Expected questions generation
  - GPT API + Fine tuning
- Keyword Extraction
- Evaluation
  - Pronunciation, 사투리, filling word, gesture, …

4. Implementation

Speech to Text

- It is necessary **for use in natural language processing** such as keyword extraction.
- There are open sources like Zeroth and Kospeach.
- However, we will use **ETRI '음성인식' API** for the following reasons.

⚠ Development Cost:

Our project is not ASR technology developmenting project but presentation analysis. If we implement the underlying technology first, the development COST will increase considerably.

⚠ Accuracy:

If we produce the ASR model we made, it is difficult to guarantee its stability at the moment. However, the ETRI API has already been verified.

⚠ Speed:

When using the model, it requires more computing power than API and can be slowed down

4. Implementation

Speech to Text

- A preprocessing process is required for API use.
  - 1. Sampling frequency should be 16 kHz.
  - 2. It should be encoded with Base64.
  - 3. Split every 20 seconds.
- Request REST API from ETRI OPEN API server using HTTP method after preprocessing.

```
[Segment 0]
택배 기사가 짐이 가득 실린 카트를 끌고 경사진 길을 오릅니다 아파트 입구에서 가장 가까운 동까지 삼백 미터 거리를 차량 없이 오가야 합니다 아파트 단지에서 십 년 전쯤부터
[Segment 1]
택배 차량의 지상 출입을 막고 지하 주차장만 허용했기 때문입니다 하지만 차량 높이가 주차장 입구를 넘어서 들어갈 수가 없다 보니 몇 배나 힘이 드는 일을 매번 반복할 수밖에 없습니다
[Segment 2]
차 타고 들어갔을 때에는 거의 한 3,40분 정도면 끝나는 시간인데 이제 구름알을 끌고 이제 왔다 갔다 한 시간 반에서 한 시간 40분에서 50분 정도 소요가 돼요 경기 수원에 있는 대단지 아파트는
[Segment 3]
주제 택배 대란입니다. 이달부터 택배 차량의 지상 출입을 막자 반발한 기사들이 밖에다 택배를 쌓아둔 겁니다.
[Segment 4]
아이들 안전과 차없는 쾌적한 환경을 위한 조치라지만 주민들 사이에서도 입장이 엇갈립니다.
[Segment 5]
을 지난 2018년 이후 곳곳에서 비슷한
[Segment 6]
```

ETRI제공 '음성인식'API 시현 결과

4. Implementation

Slide to Text

- For NLP such as keyword extraction...

- There are many open sources.
  - **pytesseract**
    - pytesseract is a simple ocr library.
    - ⚠ However, the recognition rate of Hangeul is not good.
  - **easyOCR**
    - It supports multiple language.
    - The recognition rate is better than that of pytesseract.

We will use **easyOCR**!

4. Implementation

pytesseract

```
4. 아스트 패이지입니다

0)₩6556「80아를 테스트하기 위한 글입니다.

소웨 최고
가나다라마바사

각맑빨
```

easyOCR

```
['4. 테스트 폐이지입니다:', 'pytesseract를 테스트하기 위한 글입니다:',
'소웨 최고', '가나다라마바사', '꿀닭빰']
```

4. Implementation

Expected question generation

- Method 1

  Train the model using the KorQuAD dataset

  - which consists of question-answer pairs based on passages.

  ⚠ But, this dataset only contains questions that can be found within the given passages.

- Method 2

  Use GPT API and fine-tune it.

  Unlike Method 1, this approach generates questions that may not be explicitly mentioned in the presentation content.

4. Implementation

Keyword extraction

- It helps users memorize and practice the entire flow with only keywords.

- There are several ways to extract keywords.
  - TF-IDF
    the easiest way. It extracts keywords based on frequency.
  - YAKE!
    - Based on statistical techniques, we provide Python api.
  - KeyBERT
    Based on Bert, it provides Python api.
  - GPT
    There are no keyword extraction model using GPT

➡ **Try all four methods, compare performance, and choose the best way to perform**

4. Implementation

Keyword emphasis analysis

- Analysis of highlights in presentation exercises based on keywords

- No related open source or thesis found.

- We're going to briefly analyze it in the following order

  1) Pre-processing such as tokenization, elimination of non-terminal terms, word rematification, etc

  2) Comparison of similarity between previously extracted keywords and each word in the script (use cosine similarity)

     - If there is a word with a similarity above the threshold,
     it is judged that the keyword is emphasized.

4. Implementation

Evaluation - pronunciation

- Required to analyze user presentation
  - Accurate pronunciation is one of the important factors in delivering presentations.
- How to Implement Features
  - 1) Using Model
  - 2) Using API

4. Implementation

1) Using Model

- STEP
  - 1. Voice MFCC extraction
  - 2. HMM model training

도면 1

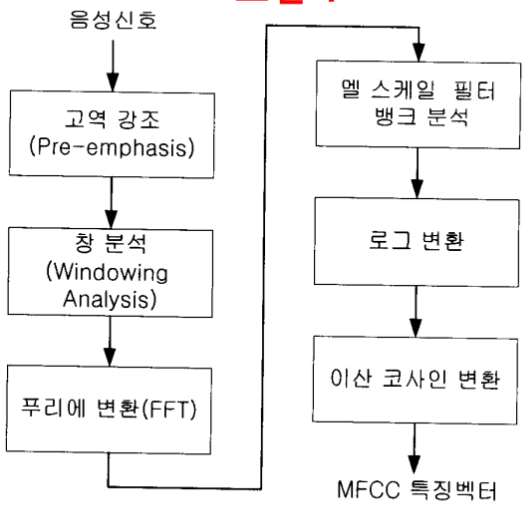


그림 1. MFCC 특징 벡터 변환 과정

도면 2

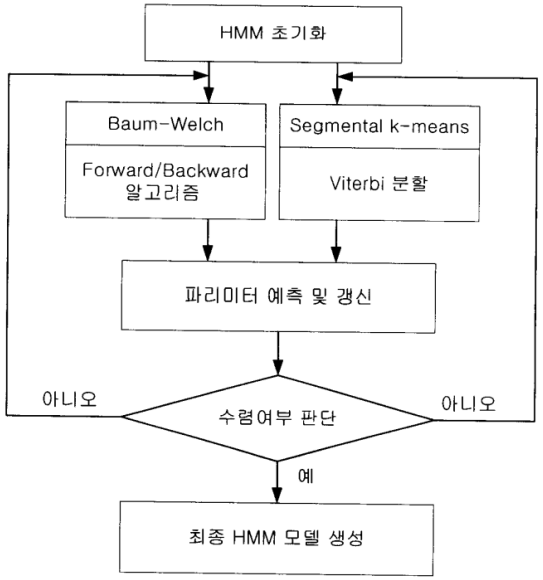


그림 2. 음향 모델 훈련 과정

등록특허 10-0362292
음성인식 기술을 이용한 영어 발음 학습 방법 및 시스템

Advantages

-1. Usability: It is possible to create a FIT model on our application

Disadvantage

-1. Speed: It requires more computing power than api and can be done by slowing down.

-2. Accuracy: There is no verification that the model made is superior to the api.

-3. Dataset: Difficulty in obtaining large, reliable datasets

4. Implementation

2) Using API

- We will use the ETRI '발음평가' API.
- The reasons are as follows.
  - ACCURACY

    ETRI api is well-tested, and it looks like stable presentation analysis will be possible using this.
  - Cost

    The Pronunciation Analysis API is provided by ETRI, such as the ASR API. Accordingly, a file preprocessing process or the like may be omitted, and through this, the overall COST may be lowered.

```
[responseCode] 200
[responBody]
{"result":0,"return_type":"com.google.gson.internal.LinkedTreeMap","return_object":{"recognized":"안녕하세요 오늘도 멋진 하루 되세요.","score":"1.557736"}}
```

ETRI제공 '발음평가'API 시현 결과

4. Implementation

Evaluation – Dialect(사투리)

- It is a function for those who want to speak the standard language correctly.
- Dialect characteristics can be mixed between utterances without even realizing it. It detects this and shares the status of dialect usage with the user.

Implementation Step

- Speech Processing: Noise Removal and Normalization Processing
- Speech characteristics: Speech vectorization using MFCC feature extraction method
- Model: Random Forest Model Learning
- Result: Digitize classification with SoftMax

Fig. 1. System Architecture

김영국 and 김명호. (2021). 음향 특성에 따른 한국어 방언 분류 모델의 성능 비교. 한국컴퓨터정보학회논문지, 26(10), 37-43.

4. Implementation

- Good points
  - Guide: The author wrote an end-to-end process from data preprocessing to evaluation.
- Limit
  - It shows a low accuracy of around 65%.
- Improvements
  - It will be released as a beta service, not an official service.
  - By feature engineering, Try to improve accuracy
  - Voting segments.

Table 3. Table of Classification Model Performance with Scaled MFCC

	Accuracy	Precision	Recall	F1-score
SVM	64.02	62.69	64.02	61.73
RF	64.50	63.97	64.50	62.02
DNN	**65.00**	**64.49**	**65.00**	**63.43**
RNN	63.66	58.05	63.66	60.29
LSTM	64.56	63.39	64.56	61.42
Bi-LSTM	61.84	58.47	61.84	58.47
GRU	62.62	56.99	62.62	59.28
1D-CNN	64.23	58.80	64.23	

김영국 and 김명호. (2021). 음향 특성에 따른 한국어 방언 분류 모델의 성능 비교. 한국컴퓨터정보학회논문지, 26(10), 37-43.

Evaluation – filling word

- To analyze how many times a user has said filling words that are distracting

- step
  - 1. Build filling word Dataset
    -> Build filling word datasets such as "아", "음", and "그".
  - 2. Convert voice to text using ASR technology.
    -> The ETRI '음성인식' API currently used in this project detects filling words.
  - 3. Using the filling word dataset, search for filling words in converted text.

4. Implementation

Evaluation – speed of utterance

- To analyze how fast the user's speech

step

- 1. Calculate the number of utterance syllables per unit time from collected data and write statistical data
- 2. Calculate the amount of time except where the user's voice does not utter
- 3. Calculates the number of phonetic syllables per unit time based on the text obtained in asr
- 4. Compare user data with statistical data.

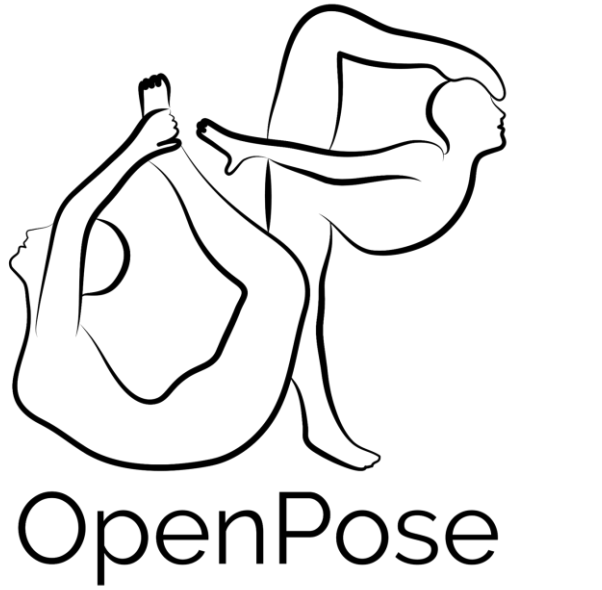
⚠ The data must be authenticated Korean data.

⚠ The reason for selecting Korean data is that the average speaking speed may vary depending on the language.*

*이은성. "화자의 발화속도, 발음, 소음이 통역 청취에 미치는 영향." 국내박사학위논문 韓國外國語大學校 通飜譯大學院, 2022. 서울, 180p

4. Implementation

Evaluation - gesture



- In presentation, gesture is effective way in attracting audience's attention.
  In result, analyze gesture and provide in the form of timeline.
- OpenPose: One of the open sources that predicts person's joint position in the photo or video.
- COCO: Provides datasets used in the field of computer vision.

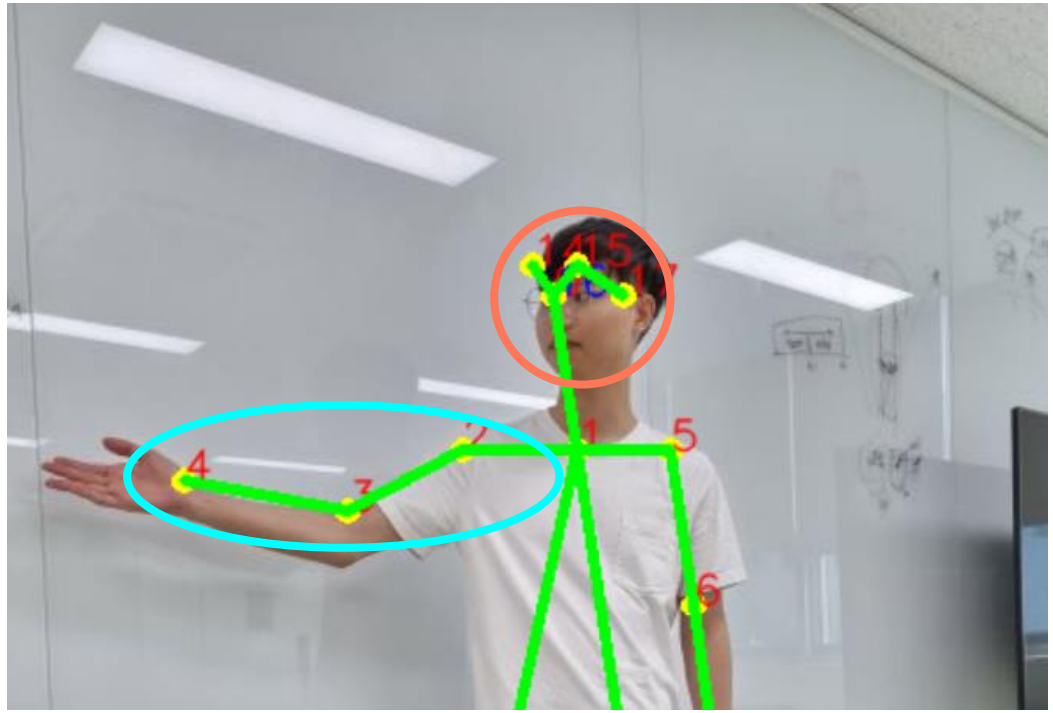
- Limitation: Given model takes long time to analyze videos.
  Expected solution
  1) Reduce the number of frames of presentation video to analyze
  2) Lightweighting the model

4. Implementation

Evaluation - gesture

- Arm movement

  Find movement using arm coordinates



4. Implementation

Client application

FrontEnd: Android using Android Studio

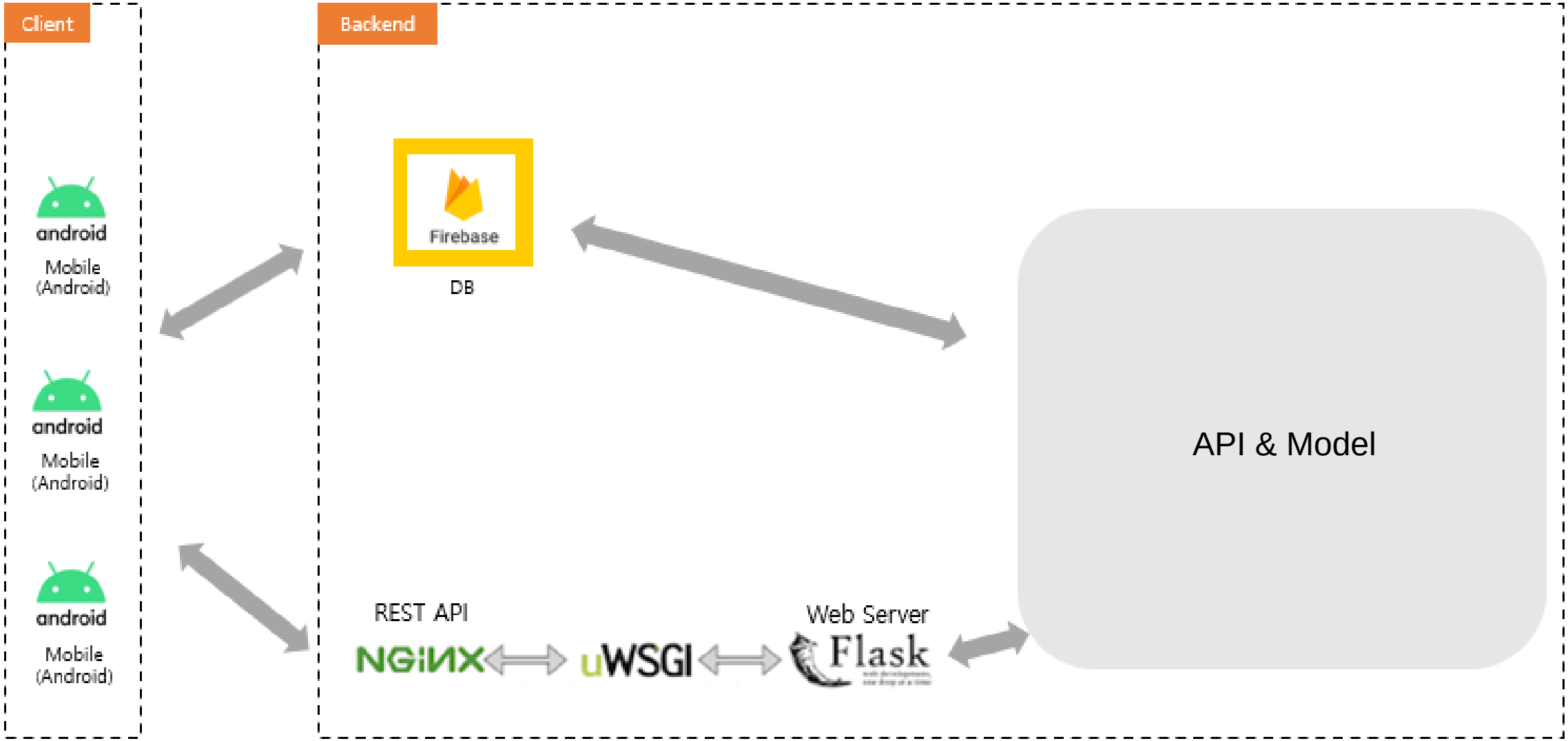
UI/UX: Overall UI will be designed using Pigma

Server & DB

DB: Firebase realtime database

Server: Flask

6. Architecture



7. Development Plan and Role

July	August	September	October	November

Study

Server, REST API,

deep learning,

how to use Figma

Implementation

- Dialect detection
- Keyword extraction
- Figma app design

Implementation

- Text extraction from slides
- Keyword emphasis analysis
- Gesture detection

Testing

- User testing
- Correction

Implementation

- Filler words
- Pronunciation
- ASR API connection

Implementation

- App development
- Server construction

7. Development Plan and Role

- Develop organically together
- Below are the areas where one takes on the role of leader.

이지해	송윤수	김세중
- Generative AI - NLP	- Server - Gesture	- DB - ASR & MFCC

Thank You

For Listening