**캡스톤 디자인 I**

**종합설계 프로젝트**

|  |  |
| --- | --- |
| 프로젝트 명 | 아나운서 준비생을 위한 맞춤형 AI 스피치 연습 애플리케이션, Loro(로로) |
| 팀 명 | 커비 |
| 문서 제목 | 중간보고서 |

|  |  |
| --- | --- |
| **Version** | 1.0 |
| **Date** | 2024-03-31 |

|  |  |
| --- | --- |
| **팀원** | 안지원 (조장) |
| 김필모 |
| 신민경 |
| 윤하은 |
|  |
| **지도교수** | 이창우 교수 |

|  |
| --- |
| **CONFIDENTIALITY/SECURITY WARNING**  이 문서에 포함되어 있는 정보는 국민대학교 소프트웨어융합대학 소프트웨어학부 및 소프트웨어학부 개설 교과목 캡스톤 디자인I 수강 학생 중 프로젝트 “Loro(로로)”를 수행하는 팀 “커비”의 팀원들의 자산입니다. 국민대학교 소프트웨어학부 및 팀 “커비”의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다. |

**문서 정보 / 수정 내역**

|  |  |
| --- | --- |
| **Filename** | 중간보고서-아나운서 준비생을 위한 맞춤형 AI 스피치 연습 애플리케이션, Loro(로로).doc |
| **원안작성자** | 안지원, 김필모, 신민경, 윤하은 |
| **수정작업자** | 안지원, 김필모, 신민경, 윤하은 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 수정날짜 | 대표수정자 | Revision | 추가/수정 항목 | 내 용 |
| 2024-03-31 | 윤하은 | 1.0 | 최초 작성 |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**목 차**

[1 프로젝트 목표 4](#_Toc411285431)

[2 수행 내용 및 중간결과 5](#_Toc411285432)

[2.1 계획서 상의 연구내용 5](#_Toc411285433)

[2.2 수행내용 5](#_Toc411285434)

[3 수정된 연구내용 및 추진 방향 6](#_Toc411285435)

[3.1 수정사항 6](#_Toc411285436)

[4 향후 추진계획 7](#_Toc411285437)

[4.1 향후 계획의 세부 내용 7](#_Toc411285438)

[5 고충 및 건의사항 8](#_Toc411285439)

# 프로젝트 목표

### 문제 정의

1. 비용적 측면
   * 고가의 등록비 및 수업료: 대부분의 아나운서 학원은 높은 등록비와 수업료를 요구합니다. 이로 인해 경제적으로 부담이 있는 학생들은 학원 수강을 포기하거나, 부모님의 경제력에 따라 선택을 할 수밖에 없는 경우가 있습니다.
   * 추가 비용 부담: 정규수업 외에도 특별 강의, 개인지도 등을 위한 추가 비용이 발생하는데 이로 인해 전체적인 비용 부담이 더 커지는 경우가 있습니다.
   * 아나운서 학원 비용 관련 뉴스 기사 목록
     + <https://www.mediatoday.co.kr/news/articleView.html?idxno=212098>
     + <https://news.mt.co.kr/mtview.php?no=2014051910164095551>

A screenshot of a white table

Description automatically generated

1. 지리적 측면
   * 학원의 위치: 대부분의 아나운서 학원은 대도시나 번화가에 위치해 있기 때문에, 지방이나 외곽 지역 거주자들에게는 접근성이 떨어집니다. 이로 인해 지방 거주자들은 학원까지 이동하는데 추가적인 교통비와 시간을 들여야 합니다.

직접 조사한 결과를 바탕으로 작성한 그래프(네이버지도 검색 기준)

A graph with pink squares

Description automatically generated

### **유사 서비스 조사**

다음은 저희 앱과 유사한 기존의 서비스들을 벤치마킹한 결과입니다.

A screenshot of a phone

Description automatically generated

저희 앱은 다른 앱들과 비교하여 아래의 세 가지 측면에서 차별화된 장점을 가지고 있습니다.

1. **'아나운서' 직업적 특성 반영**  
   : 아나운서를 위한 전용 기능과 특별한 훈련 콘텐츠를 제공하여 사용자가 실제 아나운서로서 필요한 기술과 능력을 향상시킬 수 있습니다. 이를 통해 사용자는 실제 아나운서가 사용하는 발음, 억양, 목소리 조절 등과 관련된 전문적인 기술을 습득할 수 있습니다.
2. **맞춤형 음성 가이드 제공**  
   : 사용자의 음성 특성을 분석하여 개별적인 맞춤형 음성 가이드를 제공합니다. 이를 통해 사용자는 자신의 음성에 맞춰 발음, 억양 등을 개선할 수 있습니다. 또한, 사용자의 발음이나 억양에 대한 실시간 피드백을 제공하여 학습 효과를 극대화할 수 있습니다.
3. **실시간으로 사용자가 원하는 대본 생성**  
   : 사용자가 실시간으로 원하는 대본을 생성할 수 있는 기능을 제공합니다. 사용자는 앱 내에서 원하는 카테고리와 텍스트를 입력하면 해당 내용을 기반으로 뉴스 대본을 생성하여 연습할 수 있습니다.

### 프로젝트 목적

따라서 저희는 Loro를 통해 비용적, 지리적 부담 없이 **언제, 어디서나** 아나운서나 앵커의 발성을 연습하고 피드백 받을 수 있도록 하고자 합니다.

이때 사용자마다 사용자의 음성을 기반으로 한 아나운서 억양 개인화 TTS 모델을 만들어 이를 바탕으로 실제 학원에서 배우는 것과 유사하게 발성을 연습하고 피드백 받을 수 있습니다.

# 수행 내용 및 중간결과

## 계획서 상의 연구내용

## 기능 명세서

* ***회원가입/로그인 화면***

Screens screenshot of a phone

Description automatically generated  
: 사용자가 앱을 처음 실행하면 보게 되는 화면입니다. 회원가입 또는 로그인을 유도하는 역할을 합니다.

* 로그인
  + 유저는 구글 또는 애플 계정으로 회원가입하고 로그인할 수 있습니다.
* 닉네임/캐릭터 설정
  + 닉네임을 정하면 홈화면에서 꾸준히 공부 중인 자신의 이름을 확인할 수 있습니다.
  + 원하는 캐릭터를 선택하세요. 홈 화면에서 유저는 성장하는 캐릭터를 확인할 수 있습니다.
* 음성정보 수집
  + 사용자에 맞는 가이드 음성을 생성하기 위해 음성정보를 수집합니다.
  + 3종류의 글에 대한 음성 수집이 이루어집니다.
    - 다양한 음정이나 감정 특징을 받을 수 있는 문장
* 이용약관
* 개인정보 처리방침

* ***홈***

A screen shot of a phone

Description automatically generated  
: 로그인을 마치면 홈화면으로 이동합니다. 홈 화면에서는 다음과 같이 세가지 정보를 제공합니다.

* + 캐릭터
    - 설정창에서 선택한 캐릭터가 보여집니다. 꾸준히 공부하면 캐릭터가 성장하는 모습을 확인할 수 있습니다.
  + 히스토리
    - 마지막으로 연습한 대본이 보여지며, 하단의 progress bar로 진행률을 확인할 수 있습니다. 클릭 시, 곧바로 학습하는 화면으로 이동이 가능합니다.
  + 연속 공부 일 수
    - (앱 이름)는 꾸준한 학습을 지향합니다. 유저는 홈화면 최하단에서 연속으로 공부한 일 수를 확인할 수 있습니다.
* ***기록***

A screenshot of a phone

Description automatically generated  
: 기록 화면은 사용자가 연습한 내용과 점수와 함께 음성 피드백을 자세히 확인할 수 있는 화면입니다.

* + 카테고리 설정
    - 대본페이지와 마찬가지로 카테고리에 따라 대본을 필터링할 수 있습니다.
  + 대본 목록
    - 기록 페이지의 대본 목록에서는 각 대본의 점수를 확인할 수 있습니다.
    - 대본을 클릭해 “문장단위”와 “실전 프롬프트” 연습 결과를 확인할 수 있습니다.
      * 저장된 음성과 피드백은 최신 기록만 유지됩니다.
      * “문장단위”
        + 이전에 연습했던 문장들을 확인할 수 있고 각 문장 별로 결과에 대한 피드백을 확인할 수 있고, “다시 연습하기” 버튼을 통해 다시 연습 후 피드백을 받을 수 있습니다.
        + “다음” 버튼을 눌러 대본 내 기록된 다음 문장으로 이동할 수 있으며, 마지막 문장의 경우 “완료” 버튼이 표시되어 대본 페이지로 돌아갈 수 있습니다.
      * “프롬프트”
        + 이전에 연습했던 결과에 대한 피드백을 확인할 수 있고, “다시 연습하기” 버튼을 통해 해당 페이지에서 다시 연습 후 피드백을 받을 수 있습니다.
* ***설정***

A screenshot of a phone

Description automatically generated  
설정 화면은 앱 설정을 관리할 수 있는 화면입니다.

* + 사용자 음성정보 재설정
    - 초기에 설정한 사용자 음성정보 변경을 원한다면 설정창에서 다시 설정할 수 있습니다.
  + 이용약관
    - 서비스를 이용하는 데 있어 사용자와 서비스 제공자 간의 규정된 권리와 의무를 명시합니다.
  + 개인정보처리방침
    - 개인정보를 수집, 보유, 이용, 제공하는 과정에서 이용자의 개인정보를 보호하고 안전하게 처리하기 위한 정책이나 규정을 명시합니다.
    - 수집하는 개인정보의 항목
      * 회원가입
        + 구글 계정
    - 서비스 이용을 위해 필요한 정보
      * 목소리 녹음파일
  + 로그아웃
    - 로그아웃 후 회원가입/로그인 화면으로 이동합니다.
  + 탈퇴하기
    - 탈퇴하기를 통해 현재 사용자 계정을 앱에서 삭제할 수 있습니다.

### **스크립트**

### Screens screenshot of a phone Description automatically generated

* ​스크립트 페이지에서는 연습에 사용할 대본을 정하고, “문장단위” 또는 “실전 프롬프트” 연습을 진행할 수 있습니다.   
  대본의 경우, 제공된 예시 대본을 선택하거나 사용자가 직접 대본을 업로드해 사용할 수 있습니다. 원하는 대본이 없다면, 카테고리와 키워드를 기반으로 생성형 AI를 이용해 새롭게 대본을 생성해주기도 합니다.
* 대본을 선택한 후, “문장단위”, “실전 프롬프트” 중 원하는 연습 방법을 선택하여 진행할 수 있습니다. “문장단위” 연습에서는 한 문장씩 순차적으로 진행하면서 각 문장을 집중적으로 세심하게 연습할 수 있습니다. “실전 프롬프트” 연습의 경우, 연습 화면이 실제 방송 프롬프트처럼 제공되어 대본 전문을 실전처럼 연습할 수 있습니다.
* 피드백은 음성 가이드와 사용자 음성 간의 유사도를 기반으로 판단하여 제공됩니다. ​해당 연습 기록을 저장하여 기록 페이지에서 다시 확인할 수 있습니다.

***대본***

대본은 예시 대본과 사용자 대본으로 나누어집니다.

* **예시 대본** 
  + 개요
    - ​사용자를 위해 다양한 카테고리의 예시 대본을 제공합니다.
  + ​화면 설명​
    - 하단 네비게이션 바의 “스크립트”를 클릭하면, 예시 대본 화면으로 연결됩니다.
    - 정치, 경제, 사회 등 ​다양한 카테고리의 예시 대본을 확인할 수 있습니다.
    - ​연습하고자 하는 대본을 클릭하면, 대본의 전문을 확인할 수 있습니다.
    - ​이전에 연습한 이력이 있는 대본이라면, “기록보기” 버튼을 통해 이전 연습 기록을 확인할 수 있으며, “다시 연습” 버튼을 통해 연습 화면으로 이동할 수 있습니다.
    - 이전에 연습한 이력이 없는 대본이라면, “연습하기” 버튼을 눌러 연습 화면으로 이동할 수 있습니다.
* **사용자 대본**
  + 개요
    - 사용자가 직접 대본을 생성하고, 생성한 대본으로 연습할 수 있습니다.
  + 화면 설명
    - 대본 생성하기

Screens screenshot of a phone

Description automatically generated

* + - “나만의 대본 만들기” 버튼을 클릭하면, 대본 생성 화면으로 연결됩니다.
    - 제목과 내용을 입력하고, 카테고리를 선택할 수 있습니다.
    - 자신이 업로드한 대본을 그대로 사용할 수 있으며,
    - AI 생성하기 버튼을 통해서 입력한 내용을 기반으로 대본을 생성하는 것 또한 가능합니다.
    - 대본 생성을 완료하면, 생성된 대본이 문장 단위로 분할된 것을 확인할 수 있으며, 각 문장으로 눌러 원하는 형태로 수정하거나 새로운 문장을 추가할 수 있습니다.
    - “저장 후 나가기” 버튼을 누르면 사용자 생성 대본 화면에서 해당 스크립트를 확인할 수 있습니다.
      * “연습하기” 버튼을 통해 곧바로 “실전 프롬프트” 또는 “문장단위” 연습 화면으로 이동할 수 있습니다.

***연습***

* 대본 선택을 완료하면 “실전 프롬프트” 또는 ”문장단위”로 연습을 진행할 수 있습니다.
  + **음성 가이드** 
    - 사용자는 연습을 시작하기 전 음성 가이드를 듣습니다. 가이드 음성은 개인화 TTS 모델을 이용해 사용자의 목소리와 아나운서의 특성을 반영하여 생성됩니다.
  + **문장단위**
    - 개요
      * 선택한 대본을 문장별로 순차적으로 연습할 수 있습니다.
    - 화면 설명
      * 화면 상단에 연습할 문장이 표시되고, 음성 모델이 생성한 예시 음성을 들을 수 있습니다.
      * 하단의 연습하기 버튼을 통해 해당 문장을 연습할 수 있으며, 음성 가이드와 사용자의 연습 음성을 함께 제공해 사용자가 다시 들어보며 직접 비교 분석할 수 있습니다.
      * 또한 음성 가이드와 유사도가 일정 범위 이상 벗어나는 부분은 빨간색으로 표시해 피드백을 제공합니다.
        + 피드백 어떤 식으로 줄 지 더 알아봐야 함
      * 저장하기 버튼을 누르면 해당 문장은 기록 화면에서 다시 확인할 수 있습니다.
  + **프롬프트**
    - 개요
      * 실제 방송 프롬프트 화면으로 구성해 대본 전문을 실전처럼 시뮬레이션 해볼 수 있습니다.
    - 화면 설명
      * “시작하기” 버튼을 눌러 음성 가이드를 듣거나 “바로 연습하기” 버튼을 통해 연습을 바로 시작할 수 있습니다.
      * “시작하기” 버튼을 누르면 3초 카운트다운 후 프롬프트 화면에 맞추어 음성 가이드를 들을 수 있습니다. 음성 가이드를 다 들은 후, “음성 가이드 다시 듣기” 버튼을 눌러 해당 음성을 다시 듣거나 “연습하기” 버튼을 눌러 연습 화면으로 넘어갑니다.
      * “바로 연습하기” 버튼을 누르면 3초 카운트다운 후 실제 방송 환경처럼 사용자가 프롬프트를 보며 대본 전문을 연습할 수 있습니다.
      * 연습을 완료하면, 피드백 페이지로 이동합니다. 피드백 페이지에서는 음성 가이드와 사용자의 연습 음성을 함께 제공해 사용자가 다시 들어보며 직접 비교 분석할 수 있습니다.
      * 또한 음성 가이드와 유사도가 일정 범위 이상 벗어나는 부분은 빨간색으로 표시해 피드백을 제공합니다.

## 아키텍처

A diagram of a model

Description automatically generated

Loro는 사용자 맞춤형 음성 피드백을 위해 아나운서 억양의 개인화 TTS 모델을 제공합니다. 이를 위해 전반적인 모델의 학습과 추론, 배포의 과정은 AWS 내의 Sagemaker 상에서 이루어지도록 했습니다.

해당 모델을 학습하기 위한 데이터와 학습 완료 후 만들어진 모델은 Amazone S3에 저장합니다. 이후 Inference 단계에서 다시 S3에서 해당 모델을 가져옵니다. 프론트로부터 사용자 음성 데이터와 텍스트를 받아와 모델에 입력으로 전달하여 해당 텍스트에 대한 음성 데이터를 추출합니다. AI와 프론트 사이에서 데이터를 주고받는 과정은 FastAPI를 이용해 백엔드를 구축하여 수행합니다.

사용자 정보는 파이어베이스의 Authentication을 사용하여 관리하고, 사용자가 생성한 대본이나 예시 대본, 사용자 음성 데이터, 대본별 피드백 기록(점수, 음성 가이드, 연습한 내 음성 데이터) 은 Firestore 및 Storage를 이용해 저장합니다.

스크립트 생성 기능을 위해 사용자로부터 제목과 카테고리를 입력받고, Langchain과 OpenAI API를 사용해 실시간으로 대본을 생성해주고 있습니다.

## 데이터

**아나운서 음성 데이터**  
: 음성합성을 위한 데이터는 AI-hub의 “뉴스 대본 및 앵커 음성 데이터” 데이터셋을 가공하여 사용할 예정입니다.

* <https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=71557>

**뉴스 기사 데이터**  
: 대본은 공공데이터포털의 ‘한국언론진흥재단\_뉴스빅데이터\_메타데이터\_언론’ 데이터셋을 가공하여 제작할 예정입니다.

* <https://www.data.go.kr/data/15093736/fileData.do>

## 모델

### **대본 생성**

**예시 대본 제공**

사용자가 다양한 카테고리의 대본으로 언제든 연습할 수 있도록 예시 대본을 제공합니다. LangChain의 DocumentLoader와 BeautifulSoup을 이용해 뉴스 텍스트를 읽어들이고, TextSplitter와 OpenAI를 활용해 아나운서의 뉴스 대본에 적합하도록 가공하여 대본을 만들 예정입니다. 추후 언제든 예시 대본을 추가할 수 있도록 시스템을 구축해 사용자를 위한 대본을 지속적으로 업데이트하려고 합니다.

**생성형 AI를 이용한 실시간 대본 생성**

예시 대본 외에 다른 대본이 필요한 경우를 대비해 생성형 AI를 이용해 실시간으로 대본을 만들어주는 기능도 추가했습니다. OpenAI API를 활용해 사용자가 입력한 제목과 카테고리를 기반으로 뉴스 대본을 생성해줍니다. 이때 모델은 gpt-3.5-turbo를 이용했으며, tempareture를 1로 설정해 사용자가 다양한 대본을 받아볼 수 있도록 했습니다. 또한 LangChain의 ChatPromptTemplate을 활용해 질 좋은 대본을 생성할 수 있도록 프롬프트 엔지니어링을 진행했습니다. ‘아나운서’라는 직업적 특성과 ‘시청자들에게 정보를 전달하는 뉴스’라는 상황적 특성을 반영해 사용자의 니즈에 맞는 대본을 생성하고자 했습니다.

## 음성 모델

**음성 합성 TTS 모델에서 갖춰야 할 점, 두 가지**

1. 발화가 어색하지 않고 자연스러워야 한다.
2. 문장마다 알맞은 감정이 담겨있어야 한다.

**Matcha-TTS**

* [GitHub - shivammehta25/Matcha-TTS: [ICASSP 2024] 🍵 Matcha-TTS: A fast TTS architecture with conditional flow matching](https://github.com/shivammehta25/Matcha-TTS)
* Matcha-TTS는 ODE(Ordinary Differential Equation) 기반의 TTS 모델로, 이를 사용하여 자연스럽고 고품질의 음성을 생성합니다. 확률적 흐름 매칭(OT-CFM)을 통해 학습되어 음성 생성 과정에서 동일한 품질에 대해 훨씬 빠른 합성이 가능합니다. 또한, 텍스트 인코더와 디코더의 구조를 최적화하여 메모리 소비를 줄이고 평가 속도를 빠르게 하여 synthesis 속도를 향상시킵니다.
* **장점**
  + 속도와 효율성  
    : Matcha-TTS는 빠르고 효율적인 모델로, 긴 발화에 대한 빠른 합성이 가능합니다. 저희 앱은 사용자가 만든 긴 대본에 대한 음성합성을 해야하는 데 이러한 대량의 텍스트를 처리해야 하는 상황에서 다른 모델에 비해 강점이 있다고 판단했습니다.
  + 커뮤니티 지원  
    : Matcha-TTS는 오픈 소스로 제공되며, 관련 코드와 자료가 공개되어 있습니다. 메인테이너의 issue feedback이 빠르고 자세한 편이며 발표한 지 얼마 되지 않았는데 사용자가 많아(fork, star 수) 같은 어려움을 겪는 사람들한테 도움을 받기 좋았습니다.
  + 확장성  
    : Matcha-TTS는 다중 화자 지원과 같은 확장 기능을 추가할 수 있으며, 또한 확률적 지속 시간 모델링과 같은 기술을 통해 음성의 더 복잡한 특징을 모델링할 수 있습니다. 프로젝트 진행을 하다보면 요구사항이 달라지는 경우가 빈번한데 그에 맞게 모델을 수정하고 발전시킬 수 있는 유연성이 있다고 판단했습니다.
  + 고품질 음성 생성
* **Encoder**
  + 기존의 모델들은 Matcha-TTS와 마찬가지로 encoder-decoder 아키텍처를 사용하여 Transformer 블록을 encoder에 사용합니다. FastSpeech와 같은 많은 모델은 **위치 종속성을 위해 sinusoidal position embeddings을 사용**하는데, 이는 **긴 시퀀스에 일반화하기 어렵다**는 것이 밝혀졌습니다.
  + GlowTTS, VITS 및 Grad-TTS는 대신 **relative positional embeddings**을 사용합니다. 그러나 이런 방법들은 **짧은 컨텍스트 창 외부의 입력을 "bag of words"로 처리해 종종 자연스럽지 않은 운율**이 생깁니다.
  + LinearSpeech는 rotational position embeddings (RoPE)을 사용했습니다. 따라서 Matcha-TTS는 인코더에서 RoPE를 사용하여 RAM 사용량을 줄입니다.
* **Decoder**
  + 현대 TTS 아키텍처는 decoder 네트워크 설계 측면도 다릅니다. 정규화 플로우 기반의 방법인 Glow-TTS와 OverFlow는 **확장된 1D-convolution**을 사용합니다. DPM 기반의 방법은 mel-spectrograms를 합성하기 위해 **1D-convolution**을을 사용합니다.
  + 반면, Grad-TTS는 **2D-convolution을 사용하는 U-Net**을 사용합니다. 이는 mel-spectrogram을 이미지로 취급하고 시간 및 주파수 모두에 대해 변환 불변성을 가정합니다. 그러나 음성 mel-spectrogram은 주파수 축을 따라 완전히 변환 불변이 아니며, 2D 디코더는 텐서에 추가 차원을 도입하여 일반적으로 더 많은 메모리를 필요로 합니다.
  + 한편, FastSpeech와 같은 확률적이지 않은 모델은 **1D 트랜스포머를 사용한 decoder**를 사용하여 장거리 의존성을 학습하고 빠르고 병렬적인 합성을 달성했습니다. Matcha-TTS도 decoder에서 트랜스포머를 사용하지만, 2D U-Net에서 영감을 받은 **1D U-Net 설계를 사용**합니다.
* **특징**
  + 추론 속도가 빠르고 메모리를 덜 사용합니다.
  + 다른 모델과 비교했을 때, 파라미터 수가 적은 편에 속합니다.
  + 합성 속도가 빠릅니다.
  + MAT-2 합성 속도가 긴 발화에서 FS2와 비슷하며, MAT-4도 크게 뒤지지 않음을 확인할 수 있습니다.
  + 발화가 길어질수록 MAT와 GRAD 사이의 격차가 더 커집니다.

**기존 모델 학습법 문제**

* Multi-Speaker Model
  + 한계
    - **화자 확장에 제한**  
      : 새로운 화자를 추가할 경우 Embedding Matrix에 vector가 추가되어야하는데, 이 임베딩 벡터는 Encoder, Decoder와 함께 학습되는 파라미터이기 때문에 모델 구조가 변경됩니다. 결론적으로 모델 전체를 다시 학습되어야 합니다.

A computer screen shot of a computer

Description automatically generated

* Zero shot approach
  + 한계
    - Speaker Encoder가 speaker를 구분하기 위한 모델입니다. 따라서 decision boundary를 잘 긋기 위해서 모델은 speaker들을 다 떨어뜨려 둡니다.
      * **학습된 화자 공간이 sparse 합니다.**
    - 같은 사람이지만 의미가 다른 문장을 읽었을 경우 다른 발화 특성을 보입니다.
      * 짧은 오디오에서 관찰할 수 있는 발화 특성은 contents에 align되어있는데, contents를 여러가지를 보지 못한 상태에서 **화자의 전반적인 발화 특성을 모델링하기 어렵습니다.**

A computer screen shot of a computer

Description automatically generated

**우리가 구현하고자 하는 개인화 TTS 모델 구조 → VITS2**

<https://vits-2.github.io/demo/#ss>

**VITS2** **모델 학습 방법**

1. 음성 오디오의 분포를 알 수 없기 때문에 예측하고 다룰 수 있는 분포로 Transform
   * VAE를 사용하여 Gaussian 분포에 fit이 되도록 학습합니다.
2. Target Speaker의 다양한 오디오로부터 분포에 매핑된 데이터 포인트를 찍는데, 클러스터링을 통해 여기서 대표되는 feature를 뽑습니다.
   * 클러스터링 한 센트로이드들을 모아서 화자의 음성 특징이라 정의했습니다.
   * 연속공간에서 데이터가 분포되어 있었을 것이기 때문에 중간점이 있어야 음성을 복원할 수 있습니다.

**중간점을 어떻게 뽑아올 것인가에 대한 솔루션**

* 두 개의 벡터의 선형 결합을 하면 벡터 사이의 어떤 점이라도 연속적인 공간에서 샘플링하는 것이 가능합니다.
  + Sum to one이 되는 weight를 사용하는 **Weighted Sum**을 사용함으로써 해결했습니다.
  + **Attention**이 softmax attention score를 사용하는 경우에 sum to one이 되는 weight로 weighted sum을 하게 됩니다.

A diagram of a speech features

Description automatically generated

* Phoneme Sequence가 Encoder에 들어갈 경우 각 time step에 대한 hidden representation이 만들어집니다.
* Time step에 대한 hidden representation과 Speech feature matrix와 attention을 합니다.
* 구해진 attention score matrix는 해당 알파벳을 화자에 맞게 가장 잘 표현할 수 있도록 feature들을 조합하는 방법을 나타냅니다.
* Vector들을 attention score에 맞게 weighted sum 후 다시 넣어주면 continuous discretize된 데이터 포인트에서 continuous space에서 샘플링하는 효과를 얻을 수 있게 됩니다.
* Attention matrix는 표현하려고 하는 contents를 대상 화자로 가장 잘 표현할 수 있게 해주는 speech feature를 조합하는 방법을 나타냅니다.

**성능 평가 지표**

A screenshot of a graph

Description automatically generatedA screenshot of a document

Description automatically generated

* **Speaker blending** 
  + Speaker feature space에서 다른 speaker의 feature들을 잘 조합하면 또 다른 speaker를 만들어 낼 수 있습니다.
  + Speaker space가 sparse 하기 때문에 out of distribution data point가 뽑힐 수 있는 가능성이 높았는데, 본 논문에서 제시한 모델의 경우 Speaker blending에서도 좋은 성능을 보여줍니다.

## 수행내용

### **프로젝트 진행 방향 및 관리 방법**

A blurry image of a project

Description automatically generated

**현재 진행 상황**

* 프론트 : UI 설계가 완료, 개발 착수
* 백엔드 : 사용자 정보를 저장하기 위해 파이어베이스의 구조설계 완료, 개발 착수
* AI : TTS 모델 구현 방식으로 VITS2를 선정 완료, 개발 착수

다음은 프로젝트 진행 및 관리 과정입니다. 저희는 애자일 방법론을 도입해 이번 프로젝트를 진행하고 있습니다. 먼저 여러 자료를 참고해 애자일과 스크럼에 대해 공부하고, 현재 프로젝트 관리 도구로 지라와 컨플루언스를 사용하여 프로젝트를 체계적으로 관리하고 있습니다.

저희 프로젝트는 1~2주 단위의 스프린트로, 스프린트 플래닝 - 진행 - 스프린트 회고를 거치며 진행 중입니다.

스프린트 플래닝 단계에선 스프린트를 진행하기 전 이번 스프린트 동안 무엇을 어떻게 할 지 팀원들과 의논하고, 백로그를 작성합니다.

스프린트 진행 단계에선 플래닝을 통해 만들어진 백로그를 지라를 통해 관리합니다. 깃과 지라를 연동하여 커밋과 풀리퀘스트를 백로그와 함께 기록하고 있습니다. 또한 데일리 스크럼을 작성하여 개발 진척도를 파악하고, 서로가 맞이한 block item을 확인하여 문제를 해결하고자 합니다.

스프린트 회고 단계에선 스프린트의 마지막 날, KPT 회고를 진행합니다. 회고를 통해 발견한 문제를 해결하기 위해서 와카타임과 같은 아이템들도 중간 중간 도입하여 적용하고 있습니다.

앞으로 이러한 단계를 거치며 점진적으로 프로젝트를 발전시키고자 합니다.

**문서 관리**

또한 프로젝트를 진행하며 학습한 지식이나, 개발 과정 중 생긴 버그에 대한 리포트, 회의록 등은 현재 모두 컨플루언스에 기록하여 추후에도 확인할 수 있도록 문서화하고 있습니다.

**회의**

UI 설계 및 모든 프레임워크를 선정할 때, 팀원끼리 충분한 회의를 거치며 여러 선택지를 비교 분석하여 결정했습니다. 그만큼 의논할 사항이 많고 서로가 하는 작업을 이해해야하기에 번거롭고 시간이 많이 걸리지만 단순히 금전적인 측면뿐만 아니라 팀의 기술력, 프로젝트의 목표와 요구사항을 명확히 이해하면서 모두가 함께 프로젝트를 진행하고 있습니다.

# 수정된 연구내용 및 추진 방향

## 수정사항

# 향후 추진계획

## 향후 계획의 세부 내용

* CI/CD 도입하기
* 사용자(언론학 관련 학생들에게) 설문조사를 통해 사용자 피드백 받기
* 구글 스토어에 앱 출시

# 고충 및 건의사항