한모아

교차언어 음성 합성 프로젝트
2025-2학기 캡스톤디자인종합프로젝트1 최종 발표

팀번호: 25캡스톤1_4
팀원: 김형준(팀장), 김명현, 김재우, 배지영, 홍지우
작성자: 김형준, 김명현, 김재우, 배지영, 홍지우

작성일: 2025.12.02

목차

01
프로젝트 목표

02
세부 목표

03
시스템 구조도

04
AI 파이프라인

05
시연 영상

06
Q&A

프로젝트 목표

영어 화자의 음색과 운율을 보존한 자연스러운 한국어 더빙 음성 생성

화자 특성

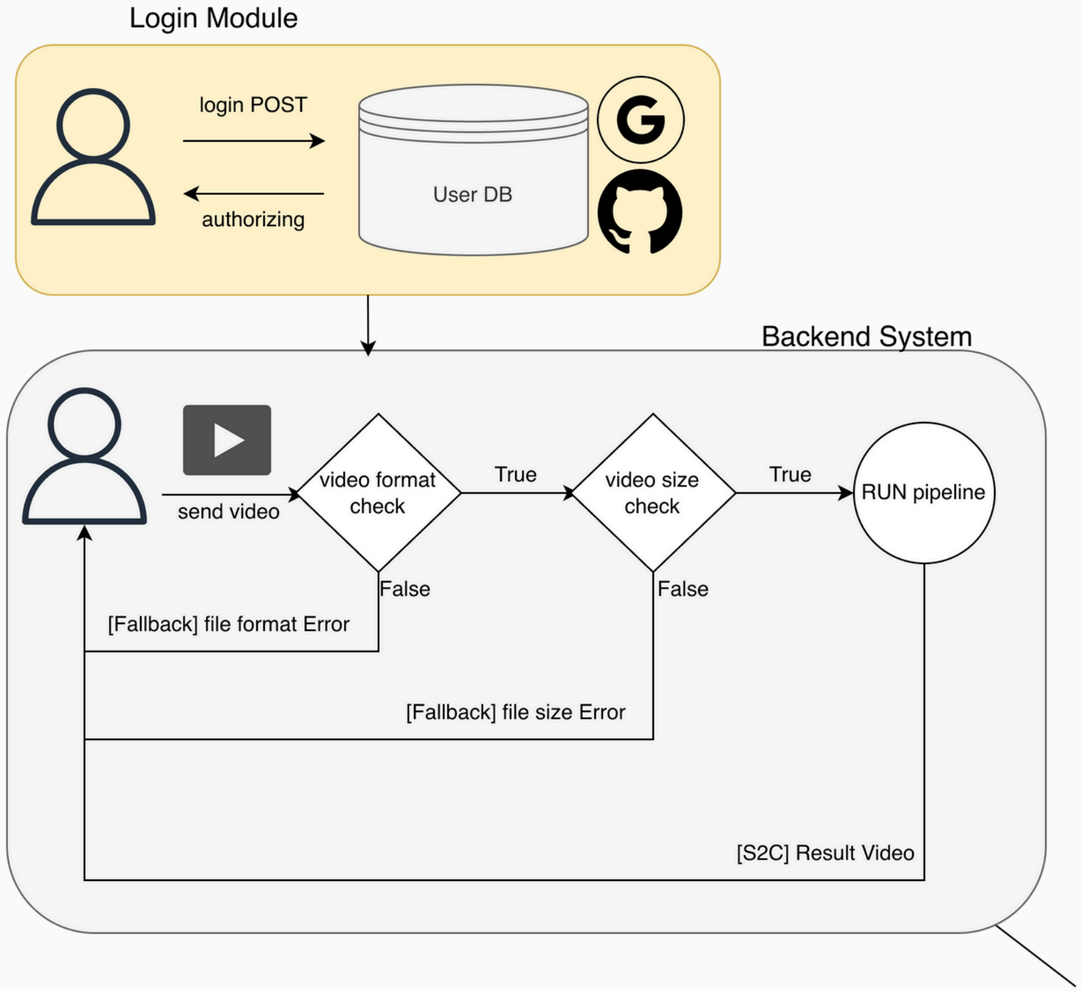
음색 · 운율 보존

화자 정체성 유지

세부 목표

목표 영역	세부 내역
파이프라인	STT-TTS 구현 영어 → 한국어 변환 파이프라인
음색 보존	3~10초 화자 샘플로 제로샷 스피커 임베딩/코덱 토큰 추정
운율 이전	피치, 에너지, 길이, 강세 곡선 추출 및 조건부 주입
입모양 변경	한국어 음성에 맞는 입모양으로 변경

시스템 구조도



- Login Module과 Backend System에 대해 다룸
- 해당 시스템은 **유저를 관리**하고, 서비스의 **인프라를 관리**함
- 소셜 로그인 모듈, 인증/인가 모듈, 사용자 관리 모듈, 파일 업로드 관리 모듈이 해당 구조 내에서 이루어짐

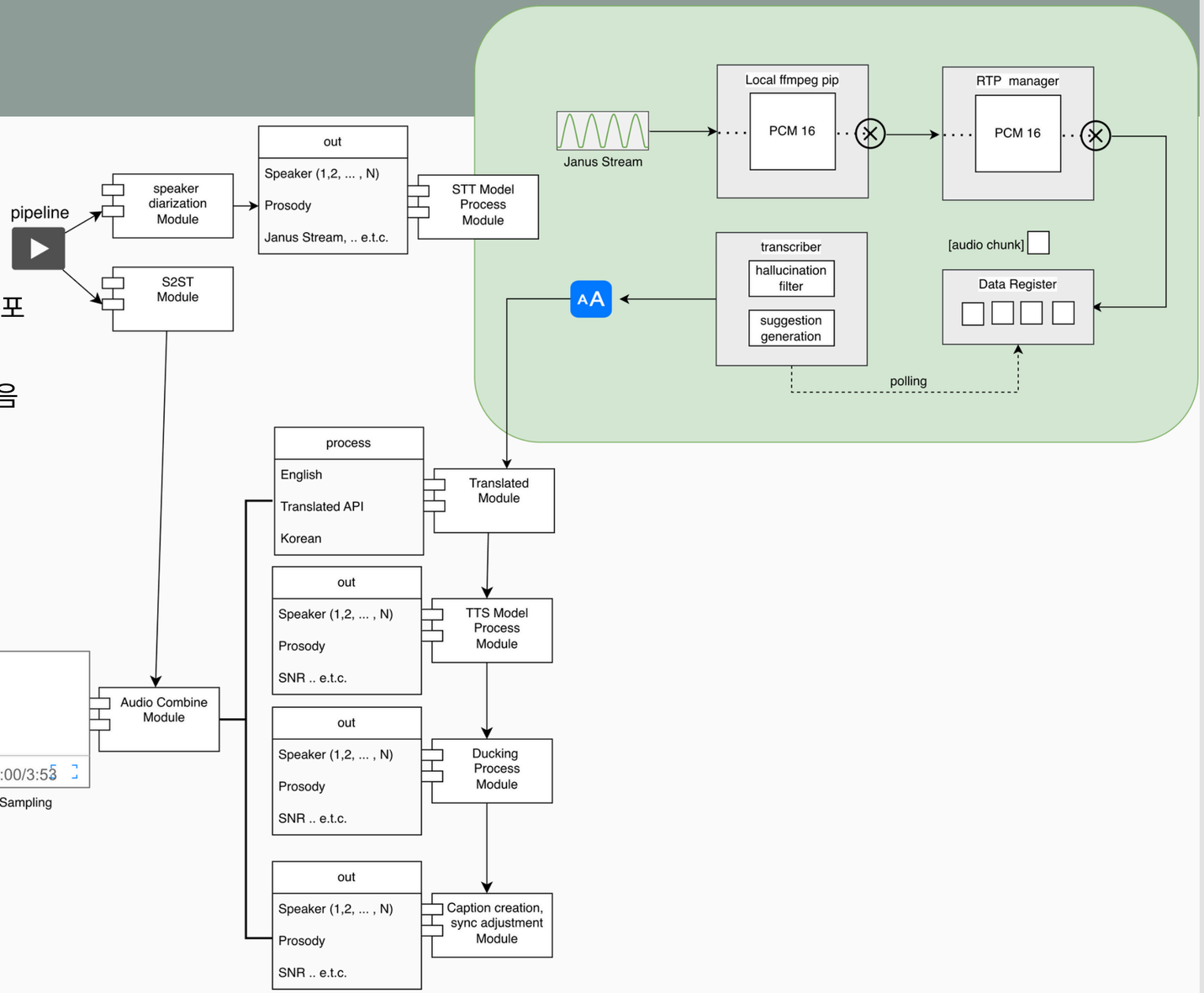
시스템 구조도

개요

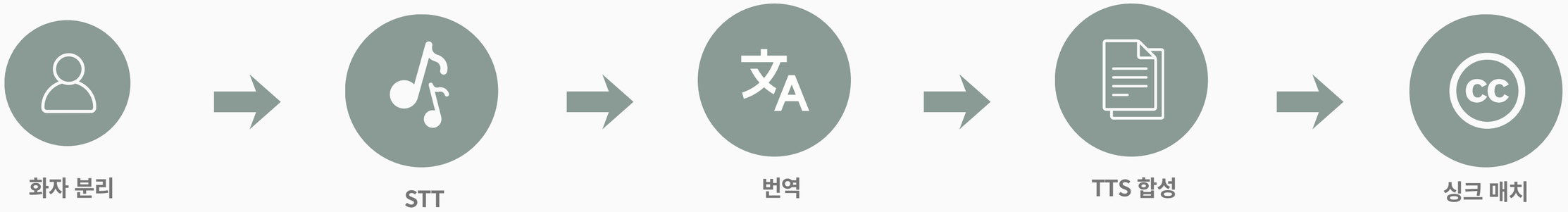
- Backend System에서 파일 입력 시 전체 파이프라인 실행
- **화자 분리 모듈(Speaker Diarization)** 과 **S2ST 모듈**이 병렬(포크) 실행
- 분리된 화자는 **STT → 번역 → TTS** 과정을 통해 텍스트 변환 및 음성 합성
- 싱크 매치를 통해서 입모양을 변경한 영상 합성
- 최종적으로 **Audio Combine Module**이 교차 음성을 융합
- **Video Sampling**을 통해 Backend로 결과 전송

처리 플로우

입력 → 화자 분리 → 음성 인식 → 번역 →합성 → 후처리 → 융합 → 전송



AI 파이프라인



화자 분리
영상 내 다중 화자를
개별 분리하여
각 발화자의
음색과 특성을 식별

STT
영어 음성을
텍스트로 알맞게
변환

번역
영어 음성을 텍스트로
변환한 후 한국어로
문맥에 맞게 번역

TTS 합성
원화자의 음색과
운율을 보존하여
한국어 음성으로 합성

싱크 매치
음성과 정확히 싱크된
입모양 생성

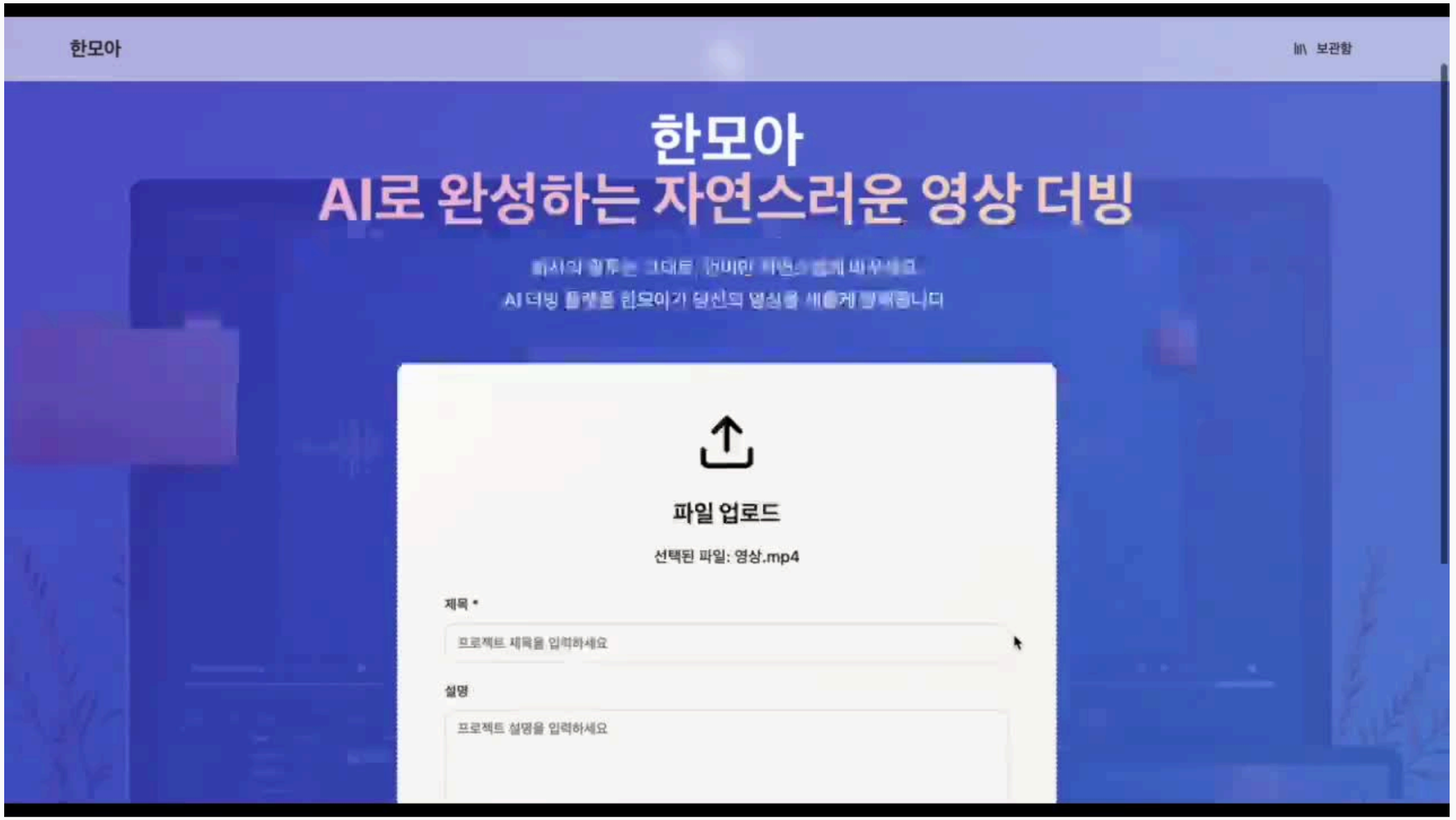
AI 파이프라인 – 상세 표

단계	입력	처리	출력	사용 기술
STT	전체 영어 음성	음성 인식	영어 텍스트 타임 스탬프	**WhisperX** Neural MT
번역	영어 텍스트 타임스탬프	번역 API 사용	한국어 텍스트	**papago API**
TTS 합성	한국어 텍스트 화자 음성 샘플	음색/운율 주입 음성 합성	한국어 음성 화자별 음색 보존	**Zonos**
싱크 매치	원본 BGM 합성된 음성	볼륨 자동 조절	믹싱된 오디오 BGM 보존	**Wav2Lip**
자막 처리	합성 음성 번역 텍스트	타임코드 정렬 자막 포맷팅	한국어 자막 SRT/VTT 파일	**Alignment** Subtitle Processing

더빙 전 영상

Only TTS

시연 영상


한모아
보관함
한모아
AI로 완성하는 자연스러운 영상 더빙
AI 더빙 플랫폼 한모아가 당신의 영상을
파일 업로드
선택된 파일: 영상.mp4
제목 *
프로젝트 제목을 입력하세요
설명
프로젝트 설명을 입력하세요

싱크 매치

질문과 답변

감사합니다