화자의 음성을 보존한 AI 번역에 관한 연구
Research on AI Translation Preserving the Speaker's Voice

김소현 김민희 이동혁 채기웅 최지현 김대원
명지대학교 ICT융합대학 융합소프트웨어학부 데이터테크놀로지전공

요 약

본 연구는 거대 언어 모델(Large Language Model, LLM)을 기반으로 한 AI 번역과 기존의 통역이나 기계 번역 서비스에서 불가능했던 음성 변환(Voice Conversion, VC)을 적용한 서비스를 제안하다. 음성인식 기술인 Speech-to-Text(STT), Text-to-Speech(TTS), 거대 언어 모델(Large Language Model, LLM), 음성 변환(Voice Conversion, VC) 등의 다양한 기술을 결합하여 음성 번역 서비스를 구현하였다. 이 연구를 통해 음성 번역 서비스는 음성과 텍스트 간의 자연스러운 전환이 가능하며, 음성 변조를 통해 화자의 특징을 복원한 번역 음성을 생성할 수 있다. 또한, 이 서비스는 음성 번역 뿐만 아니라 교육적인 목적으로도 활용 가능하며, 실생활에 적용 가능한 형태의 모바일 어플리케이션으로 제작되었다.

1. 서 론

최근 인공지능(AI) 기술의 급속한 발전은 번역 분야에서 혁명적인 변화를 가져오고 있다. 특히, AI를 활용한 번역기 및 ChatGPT와 같은 언어 모델의 등장으로 언론은 통번역사의 역할이 줄어들거나 사라질 것으로 예측하고 있다[1]. 그러나 현실적으로 AI 번역은 아직 사용자 경험과 효율성 측면에서 한계를 가지고 있다. 구글 번역기, 파파고, 딥엘(DeeL)과 같은 대표적인 번역기는 음성 서비스를 제공한다. 실시간 통역에 가장 가까운 기술을 제공하고 있지만, 여전히 짧은 문장에 적합하며 음성 인식 과정에서 오류가 발생할 수 있다[2].

거대 언어 모델 기반의 서비스인 ChatGPT는 관용구와 속담을 포함한 함축적인 의미를 정확하게 해석하며, 이를 효과적으로 번역하는 능력을 갖추고 있다. 또한 정리되지 않은 구어체 발화에 대한 통역에서 문장 내 맥락을 파악하는 데 뛰어난 능력을 보여주어, 기존 기계 번역 서비스의 한계를 극복할 수 있는 가능성을 제시하고 있다[3].

따라서 본 연구는 거대 언어 모델(Large Language Model, LLM)을 기반으로 한 AI 번역과 기존의 통역이나 기계 번역 서비스에서 불가능했던 음성 변환(Voice Conversion, VC)을 적용한 서비스를 제안하다. 해당 서비스가 기존 기계 번역 서비스의 한계를 극복하고, 음성 변환을 통해 화자 간의 직접적인 의사소통 과정에서 더 큰 유대감을 형성할 수 있을 것으로 기대한다. 또한 의사소통 목적 외에도 발음 및 억양을 교정하는 교육적 목적으로도 활용할 수 있을 것으로 기대한다.

2. 선행연구

STT(Speech-to-Text)는 음성인식의 한 분야로서 사람의 음성 언어를 컴퓨터의 해석으로 문자 데이터로 변환하는 처리를 의미한다. 이러한 기술은 HCI(Human Computer Interaction), 텔레메틱스(Telemetics), 인공지능 비서, 챗봇(ChatBot) 등 다양한 분야에서 중요한 역할을 한다[4].

TTS(Text-to-Speech)는 STT와 반대로 입력으로 문자 데이터를 받아들여 음성 신호로 변환하는 처리를 의미한다. 딥러닝 기반 TTS 시스템은 텍스트에서 스펙트로그램을 생성하는 Text2Mel 과정과 스펙트로그램에서 음성 신호를 합성하는 보코더로 구성되어 있다[5].

거대 언어 모델(Large Language Model, LLM)은 방대한 양의 데이터를 기반으로 사전 학습된 초대형 딥 러닝 모델이다. 번역, 생성, 요약, 분류 등 다양한 자연어 처리 작업에 사용되며, 대용량의 모델을 지원하고 있다. 또한, 맞춤형 모델 학습을 통해 유연한 텍스트 생성이 가능한 장점이 있다[6].

음성 변환(Voice Conversion, VC)은 목표 스피커의 목소리를 복사하는 작업으로, 소스 스피커가 발음한 발화의 언어적 내용을 보존한다[7]. 최근에는 Diffusion을 적용한 음성 변환 모델이 개발되었다. 음성 변환을 실생활에서 사용하기 위해서는 소스 및 목표 스피커에 대한 참조 발화가 적은 경우에도 작동할 수 있는 모델이 필요하다.

따라서 본 연구에서는 선행연구를 기반으로 실생활에서 사용 가능한 음성 번역 서비스 형태의 모바일 어플리케이션을 제안한다.

3. 구조

구조에서는 프론트엔드, 알고리즘, 백엔드를 각 부분별로 자세하게 설명한다. 그림 1은 음성 번역 서비스의 대략적인 과정을 나타낸다. 한국어 발화로 생성된 음성데이터는 텍스트로 변환된다. 변환된 텍스트는 번역되어 다시 음성데이터로 변환되어 출력한다.

그림 1: 어플리케이션 개요.

3.1. 프론트엔드

프론트엔드에서는 모바일 어플리케이션의 페이지를 자세하게 설명한다.

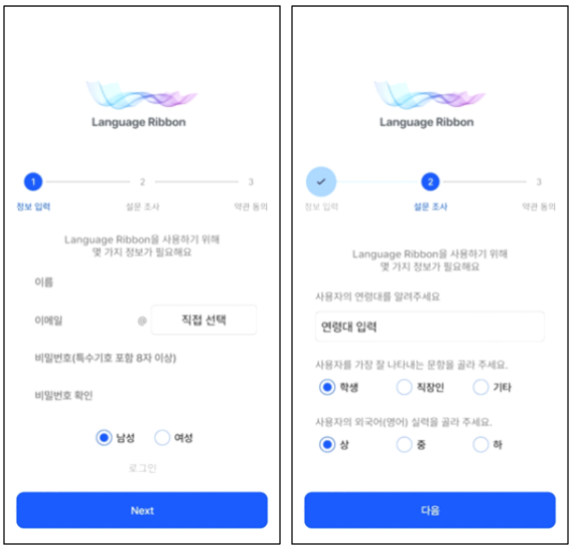

그림 2: 회원가입 페이지.

3.1.1. 회원가입 페이지

회원가입은 총 3 단계로 진행한다. 첫 번째는 사용자 정보 입력, 두 번째는 설문조사 응답, 세 번째는 약관 동의 단계이다. 각 단계에서는 모든 필수 정보를 입력해야만 회원가입 정보가 서버로 전송된다. 회원가입이 성공하면 사용자는 자동으로 로그인 페이지로 이동한다. 그러나 이메일 형식, 비밀번호 형식, 이용약관 미동의, 입력 미완료, 이미 가입된 아이디 등과 같은 회원가입 실패 시 다양한 오류 메시지가 화면에 나타난다. 또한, 회원가입 페이지에서 로그인 버튼을 클릭하면, 회원가입 과정을 거치지 않고 바로 로그인 페이지로 이동할 수 있다.

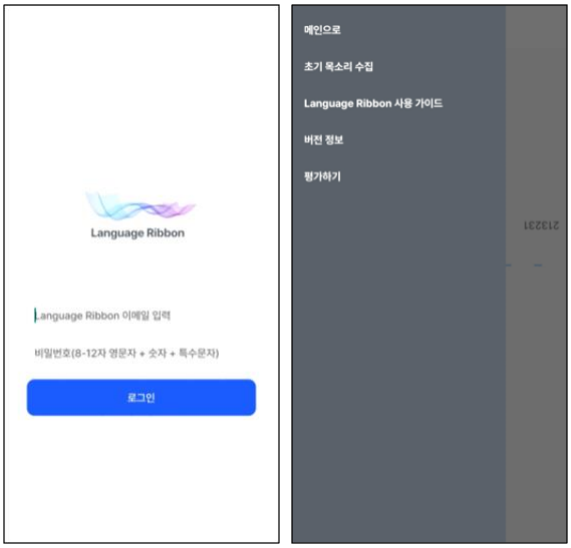

그림 3: 로그인 페이지 및 메뉴.

3.1.2. 로그인 페이지

로그인은 자체 로그인으로 이루어지며, 사용자는 회원가입 시 등록한 이메일과 비밀번호를 입력한 후 로그인이 가능하다. 클릭 시 "로그인 중"이라는 Toast 메시지가 화면에 나타나며, 이 정보가 서버에 Post 된다. 로그인이 성공하면 사용자는 메인 페이지로 이동하게 되고, 로그인이 실패할 경우 "로그인에 실패"라는 Toast 메시지가 표시된다. 또한, 서버로 Post 되는 데이터는 회원가입한 이메일과 비밀번호이다.

3.1.3. 메뉴

각 페이지는 메뉴를 통해 이동할 수 있다. 각 메뉴를 클릭하면 이전 Fragment 는 제거되고 해당 Fragment 가 생성된다. 로그아웃을 클릭하면 현재 로그인 세션이 종료되고 사용자는 회원가입 페이지로 자동 이동한다. 사이드바를 통한 메뉴 이동 시, 이전 페이지의 Fragment 는 화면에서 삭제되며, 새로운 Fragment 가 생성된다. 로그아웃을 클릭하면 현재 로그인 세션이 종료되어 로그아웃 상태가 되며, 사용자는 자동으로 회원가입 페이지로 이동하게 된다.

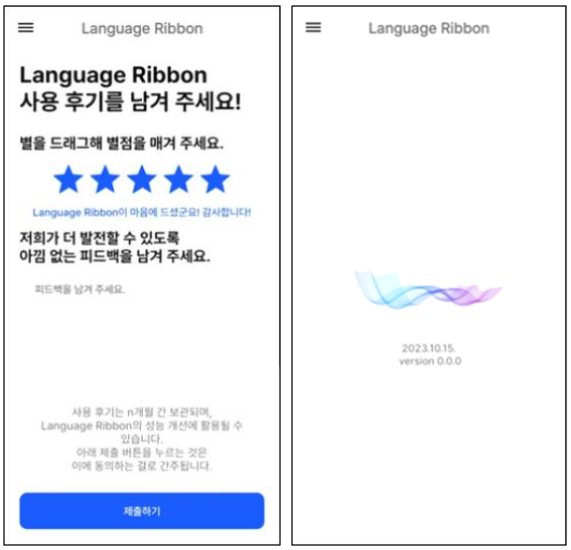

그림 4: 평가 페이지 및 버전 정보 페이지.

3.1.4. 평가 페이지 및 버전 정보 페이지

평가 페이지는 사용자가 어플리케이션을 체험한 후 별점과 텍스트로 평가한다. 별점은 총 5개로, 0.5단위로 입력이 가능하다.

버전 정보 페이지는 버전 정보를 확인할 수 있다.

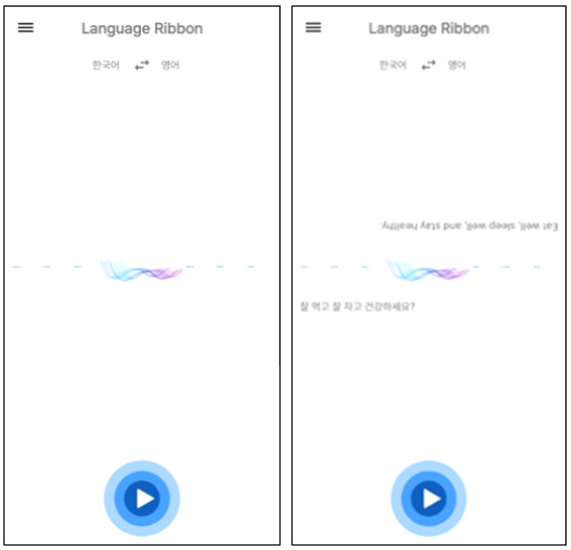

그림 5: 메인 페이지.

3.1.5. 메인 페이지

메인 페이지의 주요 기능은 사용자가 자신의 언어로 음성을 입력하면 타겟 언어로 번역된 음성이 출력되는 것이다. 초기 목소리 설정이 완료되지 않았을 경우, 녹음 버튼을 눌러도 초기 목소리 수집 페이지로 이동한다. 초기 목소리 설정을 완료한 후 녹음 버튼을 클릭하면 메인 기능을 사용할 수 있다. 녹음을 시작한 후 녹음 완료 버튼을 클릭하면 사용자 언어와 타겟 언어, 그리고 녹음된 음성 파일이 서버로 전송된다. 이후 화면에 "번역 생성 중"이라는 텍스트가 표시되며, 서버로부터 받은 값을 기반으로 음성이 재생되고 화면에는 해당 언어의 텍스트가 나타난다.

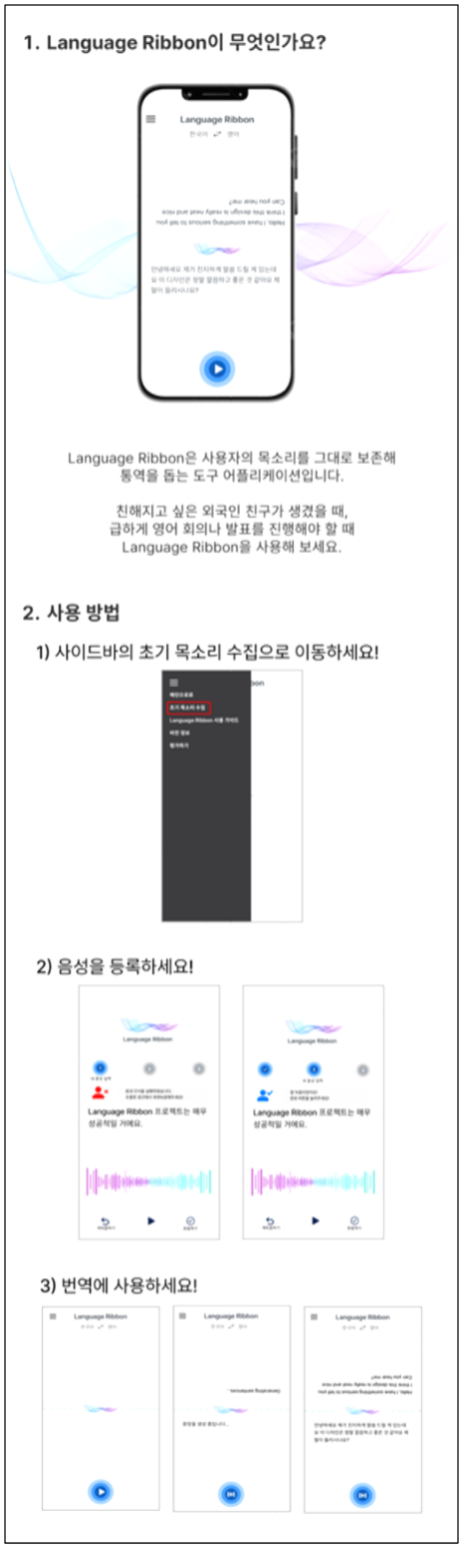

그림 6: 사용 가이드 페이지.

3.1.6. 사용자 가이드 페이지

사용 가이드를 확인할 수 있다.

그림 7: 초기 목소리 수집 페이지.

3.1.7. 초기 목소리 수집 페이지

초기 목소리 수집 페이지의 첫 번째 단계에서는 각 버튼과 사용 가이드가 상세히 설명되어 있다. 이후 단계에서는 한국어 음성녹음, 영어 음성녹음, 그리고 CER(Character Error Rate) 평가 점수를 확인하게 된다. 음성 녹음 페이지에서는 녹음 버튼을 눌러 녹음을 시작하고, 녹음 중단 버튼을 통해 녹음을 중단할 수 있다. 또한, 재녹음 버튼을 누르면 기존 음성이 삭제되어 다시 음성을 녹음할 수 있다.

CER은 자동 음성 인식 시스템의 성능을 측정하는 일반적인 메트릭으로서, 본 어플리케이션은 사용자의 이해를 위해 100-(CER 값 * 100)의 계산식을 사용하여 퍼센테이지로 표현한다. 이 수치가 81점 이상인 경우에는 푸른색으로 표시되며, 이는 높은 정확도를 나타내므로 다음 단계로 넘어갈 수 있다. 그렇지 않은 경우, Toast 메시지가 표시되어 다시 녹음이 필요함을 알려준다.

3.2. 알고리즘

알고리즘은 STT 모델을 통한 텍스트 생성, LLM을 통한 번역, TTS 모델을 통한 음성 생성과 VC 모델을 통한 음성 변환 등 총 네 가지의 과정을 거친다. 그림 8은 본 알고리즘에 사용된 모델을 과정 별로 나타낸다. 입력은 언어에 따라 다른 구조의 파이프라인으로 할당되어 음성 번역 및 변조된 결과로 출력된다.

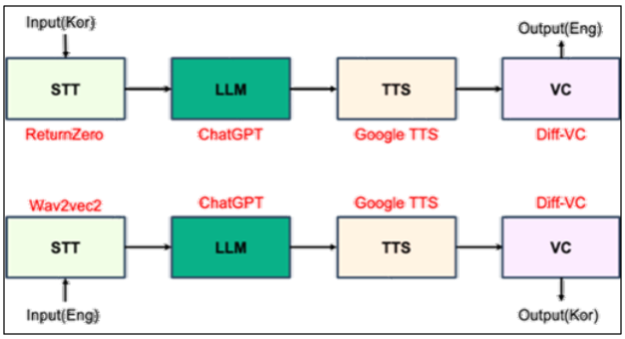

그림 8: 알고리즘 구조.

3.2.1. STT 모델을 통한 텍스트 생성

첫 번째 단계는 음성을 텍스트로 변환하는 과정이다. 각 입력 음성의 언어에 따라 다른 STT 모델을 선택한다. 한국어 음성에는 국내 STT 서비스 기업인 리턴제로사의 모델을, 영어 음성에는 PyTorch에서 제공하는 미리 학습된 Wav2Vec 2.0을 활용한다. Wav2Vec 2.0은 영어 음성에 탁월한 인식률을 보이지만, 한국어 음성에 대해서는 낮은 인식률과 알파벳만을 출력하는 한계가 있다. 따라서 한국어 음성은 높은 인식률과 한글 출력을 위해 한국어 특화 STT 모델인 리턴제로사의 모델을

활용한다. STT 모델은 음성 입력을 받아 해당 언어의 텍스트를 생성하며, 이 텍스트는 다음 모델로 전달된다.

3.2.2. LLM을 통한 번역

두 번째 단계에서는 이전 단계에서 생성된 텍스트를 번역하는 과정이다. 가장 유명한 LLM 어플리케이션 중 하나인 ChatGPT의 GPT-4를 활용한다. 선행연구에서 보았듯이 GPT-4는 이전 맥락을 고려하여 대화가 자연스럽게 이어지도록 한다. 또한 GPT-3.5 보다 높은 정확도와 빠른 속도를 제공한다[8]. LLM은 이전 모델의 출력 텍스트를 입력으로 받아 해당 입력을 번역하고 새로운 텍스트를 생성하여 다음 모델에 전달한다.

3.2.3. TTS 모델을 통한 음성 생성

세 번째 단계는 텍스트를 음성으로 변환하는 과정이다. 대표적인 TTS 모델인 FastSpeech 2.0은 Tacotron2에 비해 빠른 추론 속도를 보여준다[9]. 그러나 FastSpeech 2.0은 한글 처리가 어렵다는 한계가 있다. 따라서 본 단계에서는 한국어와 영어 음성 생성이 모두 가능한 구글사의 모델을 활용한다. TTS 모델은 이전 모델에서 출력된 텍스트를 입력으로 받아 해당 입력에 대한 음성을 생성하며, 이 음성을 다음 모델로 전달한다.

3.2.4. VC 모델을 통한 음성 변환

마지막 단계는 VC 모델을 통한 음성 변환이다. 기존의 VC 모델들은 많은 초기 음성을 요구하고 훈련해야 한다. 하지만 실제 발화 환경에서 사용 가능하려면 one-shot 형태의 빠른 추론 속도를 가진 모델이어야만 한다. 따라서 이 조건을 만족하는 오픈소스 모델인 Diff-VC를 VC 모델로 채택하였다. Diff-VC 는 모델의 인코더(Encoder)와 강력한 확산 기반 디코더(Decoder)는 도메인이 다른 이전에 본 적 없는 화자에 대해서도 화자 유사성과 음성 자연스러움 면에서 우수한 결과를 달성하였다[7]. VC 모델은 이전 단계에서 생성된 음성과 어플리케이션 초기설정에서 얻은 음성을 입력으로 받아, 이전 음성의 내용을 유지한 채 어플리케이션 초기설정에서 받은 음성의 특징을 복원하여 새로운 음성을 출력한다.

3.3. 백엔드

백엔드에서는 어플리케이션의 회원 관리 기능, 초기 목소리 수집, 음성 통역 및 변조 기능을 구현한다.

3.3.1. 회원 관리

Language Ribbon 어플리케이션의 사용자 정보 관리 기능은 회원 가입, 로그인, 로그아웃 기능으로 구성되어 있다.

회원 가입 기능은 signup 함수를 통해 구현되었다. 이미 로그인한 사용자가 이 함수를 호출하면 메인 페이지로 리다이렉트한다. POST 방식으로 요청이 들어올 경우, Django의 SignupForm을 사용하여 회원가입 폼을 처리한다. 폼의 유효성 검사를 통과하면 사용자 정보를 저장하고 사용자 프로필을 생성한다. 그 후, 사용자의 ID와 함께 회원가입이 성공적으로 완료되었음을 나타내는 메시지를 JSON 형태로 반환한다. 만약 폼의 유효성 검사를 통과하지 못하면, 에러 메시지를 반환한다. 사용자의 프로필 정보는 UserProfile 모델을 통해 관리되며, 이 모델은 Django의 User 모델과 일대일 관계를 가진다. 사용자의 이름, 성별, 나이, 직업, 영어 능력 수준, 그리고 사용자의 초기 목소리 정보 등을 포함한다. 초기 목소리 정보는 아래에서 설명하는 초기 목소리 수집 기능에서 사용된다.

로그인 기능은 login 함수를 통해 구현되었다. 이미 로그인한 사용자가 이 함수를 호출하면 메인 페이지로 리다이렉트한다. POST 방식으로 요청이 들어올 경우, Django의 AuthenticationForm을 사용하여 로그인 폼을 처리한다. 폼의 유효성 검사를 통과하면 사용자의 프로필 정보를 조회하고 이를 JSON 형태로 반환한다. 만약 폼의 유효성 검사를 통과하지 못하면, 에러 메시지와 함께 400 상태 코드를 반환한다.

로그아웃 기능은 logout 함수를 통해 구현되었다. 로그인한 상태의 사용자가 이 함수를 호출하면 사용자를 로그아웃시키고, 로그아웃이 성공적으로 이루어졌음을 나타내는 메시지를 JSON 형태로 반환한 후, 로그인 페이지로 리다이렉트한다.

이렇게 구현된 사용자 정보 관리 기능을 통해, 사용자는 자신의 정보를 안전하게 관리하며, Language Ribbon 어플리케이션의 음성 번역 기능을 원활하게 이용할 수 있다.

3.3.2. 초기 목소리 수집

Language Ribbon 어플리케이션은 음성 통역 및 변조 기능을 제공하기 위해, 사용자의 초기 목소리를 수집하고 처리하는 기능을 제공한다. uploadvoice 함수를 통해 구현하였다.

uploadvoice 함수는 사용자로부터 한국어 혹은 영어로 된 음성 데이터를 받아 처리한다. 요청 방식이 POST가 아닌 경우, 'audio' 파일이 누락된 경우, 'lang' 값이 'kr' 혹은 'en'이 아닌 경우에는 잘못된 요청임을 알리는 메시지를 JSON 형태로 반환한다.

'lang' 값이 'kr'인 경우, 사용자로부터 받은 한국어 음성을 텍스트로 변환한 후, 변환된 텍스트와 원래 스크립트를 비교하여 CER(Character Error Rate)을 계산한다. 계산된 CER이 0.3 이하인 경우, 음성이 잘 녹음되었다고 판단하여 DB에 저장한다.

'lang' 값이 'en'인 경우에도 동일한 프로세스를 거친다. 사용자로부터 받은 영어 음성을 텍스트로 변환하고, 변환된 텍스트와 원래 스크립트를 비교하여 CER을 계산한 후, CER이 0.3 이하인 경우 DB에 저장한다.

3.3.3. 음성 통역 및 변조

음성 통역 기능은 translate_to_voice 함수로, 음성 변조 기능은 Diff-VC 추론서버를 통해 구현되었다.

3.3.3.1. translate_to_voice 함수

이 함수는 음성을 입력 받아 텍스트로 변환한 후, 그 텍스트를 다른 언어로 번역하고, 번역된 텍스트를 다시 음성으로 변환한 뒤 VC 모델(Diff-VC)의 추론서버를 거쳐 사용자의 목소리로 음성을 변조시키는 기능을 제공한다. POST 요청이 들어온 경우, 'lang'과 'target-lang' 파라미터를 받아온다. 이 파라미터는 각각 원본 언어와 대상 언어를 나타낸다. 원본 언어가 영어이고 대상 언어가 한국어인 경우, 원본 언어가 한국어이고 대상 언어가 영어인 경우 모두 사용자가 업로드한 음성 파일을 텍스트로 변환하고, 그 텍스트를 영어로 번역한 후, 번역된 텍스트를 다시 음성으로 변환한다.

3.3.3.2. Diff-VC 추론서버를 통한 음성 합성 및 최종 번역 음성 제공

Diffusion 음성 변조를 위해, 백엔드에서 S3(Storage Server)로부터 사용자 초기음성과 TTS API를 거쳐 출력된 번역 음성을 Diff-VC 추론서버로 전송한다. 이때, Diff-VC 추론서버는 입력 음성을 WAV 형식으로 받으므로, TTS API가 반환한 MP3 형식의 음성파일을 pydub 모듈을 이용해 WAV형식으로 변환한다. Diff-VC 서버는 입력 음성에 해당하는 'file1'과 목표 음성에 해당하는 'file2'를 POST 방식의 form 형태로 입력 받는다. Diff-VC 서버로 file1 파라미터에 번역된 음성을 넣고, file2 파라미터에 S3로부터 가져온 사용자 초기음성을 넣어 Diff-VC 서버로 전송하게 되면, 서버 내부에서 1차적으로 입력 음성과 목표 음성의 잡음을 제거하고, 이후 훈련된 가중치 파일(.pt)를 이용에 사전에 정의한 추론 함수를 이용해 GPU상에서 추론(Inference)을 진행하여 입력 음성에 대해 목표음성의 스타일을 반영한 음성파일을 생성한다. 이후 생성된 음성을 streaming 방식으로 응답하여 백엔드 서버에 최종 결과 음성파일을 전달한다. 최종적으로 백엔드 서버는 응답 받은 파일을 처음 요청한 사용자에게 자신의 목소리로 변조된 음성파일을 전달하게 되는데, 이때 텍스트 번역 결과도 함께 제공하기 위해 응답 헤더에 "X-Json-Response" 헤더를 추가하고, 헤더의 값으로 텍스트 번역 결과와 응답메시지를 담은 JSON 데이터를 담아 사용자에게 함께 전송한다. 이때 'X-Json-Response' 헤더의 값은 직렬화 과정 중 base64 인코딩을 거쳐 반환된다. 결과적으로 사용자는 사용자의 목소리로 번역된 음성과 'X-Json-Response' 헤더에 해당 번역 텍스트를 응답 받게 되고, Language Ribbon 어플리케이션이 'X-Json-Response'의 디코딩된 JSON문자열과 최종 번역 음성을 어플리케이션 상에 적절하게 표시한다.

4. 실험

본 어플리케이션은 모델 검증과 어플리케이션 검증을 각각 진행하였다. 모델 검증은 MOS, 어플리케이션 검증은 사용성 평가로 진행하였다.

4.1. MOS(Mean opinion score)

평가자에게 주어진 MOS의 범위는 1에서 5까지이며, 높은 점수일수록 음성 품질이 높다는 것을 의미한다. 평가자는 총 세가지 음성을 평가한다. 세 가지 음성 중 Ground-truth는 원본 화자의 음성을 의미하고, 번역 제외는 원본 화자의 음성을 번역을 제외한 파이프라인의 입력으로 넣어 출력한 음성을 나타낸다. 마지막으로 본 어플리케이션은 번역이 적용된 음성이다. 우리는 본 어플리케이션이 출력한 음성의 점수가 나머지 두 음성의 점수와 큰 차이가 없다면 성능이 좋은 것으로 판단하였다. 데이터 셋은 20대 남녀의 한국어와 영어 발화 음성이다. 평가자는 총 18명으로 명지대학교에서 영어 교양을 가르치는 교수 4명, 영어영문학과 및 영미권 거주 경험이 있는 학생 6명과 타 전공 학생 7명으로 구성되어 있다.

	영어		한국어		전체	
	자연스러움	유사도	자연스러움	유사도	자연스러움	유사도
Ground truth	3.65	4.63	4.82	4.95	4.23	4.79
번역 제외	3.36	2.8	**2.41**	**2.27**	2.88	2.54
본 어플리케이션	**3.38**	**2.82**	2.11	2.1	2.75	2.46

표 1: MOS 결과.

번역 제외 음성과 본 어플리케이션 음성은 실제 화자의 음성과 유의미한 차이가 있음을 알 수 있다. 한국어 음성은 번역 제외 음성이 본 어플리케이션 음성보다 더 높은 점수를 획득하였다. 하지만, 영어 음성은 본 어플리케이션 음성이 번역 제외 음성보다 더 높은 점수를 획득하였다. 이를 통해 번역 과정이 유의미한 차이를 발생시키지 않음을 알 수 있다.

4.2. 사용성 평가

총 10가지의 문항으로 구성되어 있으며, MOS 검사와 마찬가지로 1점에서 5점으로 평가한다. 어플리케이션 파일을 전달하여 평가자가 직접 모바일 환경에서 사용한다. 회원가입부터 초기 목소리 수집, 통역 기능을 스스로 이용한다. 평가자는 10명의 명지대학교 학생이며 각 평가 점수의 평균을 반영한다.

1. 이 어플리케이션을 자주 사용하고 싶을 것 같다.	4.7
2. 이 어플리케이션이 불필요하게 복잡하다고 생각했다.	1.5
3. 이 어플리케이션을 사용하려면 기술 지원자(안내 가이드, 도우미 등)의 도움이 필요할 것 같다.	1.8
4. 이 어플리케이션이 사용하기 쉽다고 생각했다.	4.7
5. 이 어플리케이션을 사용하는 것이 매우 불편하다고 느꼈다.	1.3
6. 이 어플리케이션을 시작하기 전에 많은 것을 배워야 했다.	1.3
7. 이 어플리케이션의 다양한 기능이 잘 통합되어 있다고 생각했다.	4.3
8. 이 어플리케이션이 불안정하다고 생각했다.	1.6
9. 대부분의 사람들이 이 어플리케이션을 매우 빠르게 배울 수 있을 것이라고 생각한다.	4.5
10. 이 어플리케이션을 사용하는 데 매우 자신감이 있다.	4.7

표 2: 사용성 검사 문항 및 결과.

본 어플리케이션의 점수가 높을수록 긍정적으로 해석할 수 있는 문항으로는 1번, 3번, 7번, 9번, 10번이

다. 특히 자주 사용하고 싶다, 사용하기 쉽다, 사용하는데 자신감이 있다는 문항에서 최고 점수를 얻었다. 반대로 점수가 낮을수록 긍정적으로 해석할 수 있는 문항에서는 어플리케이션을 사용하기 불편하다는 문항과 어플리케이션을 시작하기 전에 많을 것을 배워야 한다는 문항이 최저 점수를 얻었다. 따라서 어플리케이션이 사용하기 편하고 많은 숙련도를 요구하지 않는 것을 알 수 있다.

5. 결론 및 한계

본 연구에서는 현실적이고 유용한 형태의 음성 번역 서비스를 제안하였다. 거대 언어 모델과 음성 변환 기술의 통합으로 사용자들에게 우수한 번역 경험을 제공하는 데 성공하였다. 음성과 텍스트 간의 자연스러운 전환과 화자 특징 복원을 통해 번역된 음성의 품질을 높일 수 있었다. 이 서비스는 일상 대화나 교육적인 목적으로도 활용 가능할 것으로 기대된다. 그러나, 한국어와 영어의 언어적 구조의 차이로 인해 한국어 음성의 자연스러운 처리에 어려움이 있었다. 이를 극복하기 위해 한국어의 언어적 특성을 고려한 음성 변조 모델을 훈련시키는 것이 필요하며, 이를 통해 더 나은 성능을 기대할 수 있다.

참고문헌

[1] 김유리. "챗 GPT 로 사라질 위기 직업군 1 위, 번역가, 통역사, 세무사, 회계사 4 위," Tax Times, Apr. 12, 2023. [Online]. Available: http://www.taxtimes.co.kr/news/article.html?no=259163

[2] 배문정. "통역 사용자를 대상으로 한 통역 모드 선호도 조사: 인간 통역, AI 통역, 자막 비교," 번역학연구, vol. 24, no. 3, pp. 591-614, 2023.

[3] 박미정. "생성형 AI 와 기계번역 - 챗 GPT 번역을 통한 한일통역교육 고찰" 통번역학연구 27, no.3 (2023) : 27-56.doi: 10.22844/its.2023.27.3.27

[4] 민소연, 이광형, 이동선 and 류동엽. "한국어 특성 기반의 STT 엔진 정확도를 위한 정량적 평가방법 연구" 한국산학기술학회논문지 21, no.7 (2020) : 699-707.doi: 10.5762/KAIS.2020.21.7.699

[5] 권철홍. "한국어 TTS 시스템에서 딥러닝 기반 최첨단 보코더 기술 성능 비교" 문화기술의 융합 6, no.2 (2020) : 509-514.

[6] 임철홍. "LLM(Large Language Model) 속성과 성능 연관성 연구" 정보화연구 20, no.3 (2023) : 257-266.doi: 10.22865/jita.2023.20.3.257

[7] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, “Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme,” *arXiv.org*, Aug. 04, 2022. https://arxiv.org/abs/2109.13821 (accessed Dec. 11, 2023).

[8] OpenAI, “GPT-4 Technical Report,” *arXiv:2303.08774 [cs]*, Mar. 2023, Available: https://arxiv.org/abs/2303.08774

[9] 권세영, 맹지연, 백예슬, and 김영국, "개인화된 TTS 구현을 위한 음성합성 딥러닝 모델 비교," in Proceedings of KIIT Conference, 2021, pp. 759-762.