한국빅데이터학회지
제7권 제2호, 2022, pp. 91-102 https://doi.org/10.36498/kbigdt.2022.7.2.91

박물관 안내를 위한 시나리오 기반의 AI 음성 챗봇 시스템 구현

Implementation of Scenario-based AI Voice Chatbot System for Museum Guidance

정선우[1] · 최은성[2] · 안선규[3] · 강영진[4] · 정석찬[5†]

동의대학교 부산IT융합부품연구소[1], 동의대학교 대학원 인공지능학과, 부산IT융합부품연구소[2],
동의대학교 부산IT융합부품연구소[3], 동의대학교 인공지능그랜드ICT연구센터[4],
동의대학교 e비즈니스학과, 인공지능그랜드ICT연구센터, 부산IT융합부품연구소[5]

요 약

인공지능이 발전하면서 AI 챗봇 시스템의 활용이 활발히 이루어지고 있다. 그 예로 공공기관에서는 민원, 행정 분야에서 업무 보조, 전문지식 서비스 등으로 챗봇 활용 분야가 확대되고 있으며 민간기업은 대화형 고객 응대 서비스 등으로 챗봇을 활용하고 있다. 본 연구에서는 시나리오 기반의 AI 음성 챗봇 시스템을 제안하여 박물관의 운영 비용을 절감하고 관람객에게 양방향성 안내 서비스를 제공하고자 한다. 구현한 음성 챗봇 시스템은 실시간으로 특정 디렉터리를 감시하여 사용자의 음성을 감지하는 감시자 객체와 음성 파일이 생성되면 순차적으로 모델별 추론을 수행하여 AI의 응대 음성을 출력하는 이벤트 핸들러 객체로 구성되며, 스레드와 데크를 활용한 중복 방지 기능을 포함하여 단일 GPU 환경에서 추론 중에 GPU의 연산이 중복되지 않도록 한다.

■ 중심어 : 박물관, 음성 챗봇

Abstract

As artificial intelligence develops, AI chatbot systems are actively taking place. For example, in public institutions, the use of chatbots is expanding to work assistance and professional knowledge services in civil complaints and administration, and private companies are using chatbots for interactive customer response services. In this study, we propose a scenario-based AI voice chatbot system to reduce museum operating costs and provide interactive guidance services to visitors. The implemented voice chatbot system consists of a watcher object that detects the user's voice by monitoring a specific directory in real-time, and an event handler object that outputs AI's response voice by performing inference by model sequentially when a voice file is created. And Including a function to prevent duplication using thread and a deque, GPU operations are not duplicated during inference in a single GPU environment.

■ Keyword : AI Chatbot, museum

2022년 11월 17일 접수; 2022년 12월 08일 수정본 접수; 2022년 12월 08일 게재 확정.
* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 지역연계 첨단 CT실증(R2022020140, 1375027492)사업의 연구 결과로 수행되었음.
† 교신저자 (scjeong@deu.ac.kr)

Ⅰ. 서 론

인공지능의 발전과 더불어 전 분야에 걸쳐 AI 챗봇의 활용이 증가하고 있다. AI 챗봇은 딥러닝, 자연어 처리 등 인공지능 기반의 기술이며, 텍스트나 음성으로 사용자와 커뮤니케이션을 수행하여 인포테인먼트, 엔터테인먼트 외에도 다양한 비즈니스 영역에 걸쳐 사용될 수 있다. 그 예로 공공기관에서는 민원, 행정 분야에서 업무 보조, 전문지식 서비스 등으로 챗봇 활용 분야가 확대되고 있으며 민간기업은 대화형 고객 응대 서비스 등으로 챗봇을 활용하고 있다. 또한 2018년에 국립중앙박물관 및 국립나주박물관에서 AI 음성 챗봇이 탑재된 로봇 '큐아이'를 선보였다.

하지만 최근 국내 주요 박물관을 제외한 사립 박물관이나 비교적 규모가 작은 박물관은 첨단 디지털・인공지능 기술 도입 미흡, 코로나 팬데믹으로 인한 유례없는 장기휴관의 지속, 박물관의 전시 안내를 맡는 학예사나 전문인력의 부족 등의 문제를 겪고 있다.

이에 따라, 본 논문에서는 적은 양의 시나리오 데이터로도 박물관의 안내가 가능한 시나리오 기반의 AI 음성 챗봇 시스템 구현을 제안한다.

음성안내 챗봇 시스템을 위해 구현한 통합 인터페이스는 파일 감지 모듈이 무한루프로 동작하면서 특정 디렉터리에 사용자의 발화 음성이 담긴 음성 파일이 생성되었을 때의 이벤트를 감지하여 실시간으로 추론을 수행한다. 추론 과정은 앞서 말한 파일이 감지되면 음성인식(STT) 모델을 거쳐 음성 데이터를 텍스트로 전환하고, 문장 의도 분류 모델을 통해 들어온 텍스트에 대한 의도를 분류하고 사용자의 질의에 맞게 응대하는 텍스트를 내보낸다. 끝으로 응대 텍스트를 음성합성(TTS) 모델로 합성하여 AI의 응대 음성을 WAV 파일로 저장한다. 또한, 통합 인터페이스 구현 시 스레드(Thread)와 데크(Deque)를 활용한 중복 방지 기능을 구현하여 GPU가 1대인 경우에 연산이 꼬이지 않고 모델 추론을 수월하게 할 수 있도록 했다. 제안한 시스템은 목적 지향형 챗봇 시스템이므로 각 박물관의 성격에 맞게 시나리오만 변경하면 다양한 박물관에 활용가능하다.

본 논문에서는 박물관을 주요 타겟으로 시나리오 기반의 AI 음성 안내 챗봇 시스템을 구성하였고, 통합시스템의 구성 요소로는 음성인식, 문장 의도 분류, 음성합성 기능 등이 있다.

Ⅱ. 관련 연구

박물관 관람객들은 작품이나 전시물 옆에 붙여진 짧은 글 외에는 전시물에 대해 정보를 얻기가 힘들다. 지금까지도 전문인력이 관람객들과 직접 동행하며 관람 동선을 알려주고 더 나아가 투어 안내를 제공하거나, 사용자가 특정 전시물 앞에 섰을 때 이를 감지하여 사전에 녹음된 음성 파일을 재생시켜 사용자에게 정보를 제공하는 등의 방식으로 안내 서비스를 제공하는 박물관이 존재한다. 그러나 전문인력이 동행하는 안내 서비스는 박물관 운영 시의 유지비용이 많이 들고, 녹음된 음성 파일은 관람자가 원하는 정보를 제공하기보다 일방적으로 정보를 제공하는 양방향성 서비스가 불가하다. 반면, AI 챗봇을 이용한 박물관 안내 시스템의 경우 운영 비용이 절감하고 사용자와 양방향성 의사소통 및 박물관 안내 서비스가 가능해진다는 이점이 있다.

현재까지 박물관의 안내 시스템 개선을 중심으로 많은 연구가 진행되고 있다. 예를 들어, 정현숙[1]이 제안한 도슨트 어플은 박물관 관람 시 안드로이드 기반의 모바일 도슨트 어플을 통해 문화재 관련 지식을 습득하고 교육 및 체험 프로그램을 이용하는 방법을 제안하였다. 또한 김종건[2]이 제안한 스마트 도슨트 챗봇은 관람객이 소지한 스마트폰을 통해 비대면으로 작품과 전시의 설명을 들을 수 있다. 하지만 이러한 시스템은 정상적인 모바일 환경이 필수적으로 수반되어야

한다. 그리고 관람객마다 각자 소지한 스마트폰의 성능이나 네트워크 환경에 따라 기능의 활용도가 떨어질 수 있다.

2.1 LAS 모델

자동 음성인식 딥러닝 모델을 사용하기 쉬운 형태로 구성한 OpenSpeech에서 제공하는 LAS (Listen, Attend and Spell)를 사용했다[3]. 기존의 음성인식 기술은 음향, 발음 및 언어모델 등을 개별적으로 학습하고 추론하여 최적화가 미흡하고 오류 전파로 인해 성능의 한계가 존재했으나, W Chan이 제안한 LAS는 종단 간 기계학습(End-To-End) 방식으로 분리되어 있던 모든 학습 과정을 하나로 통합한 모델이다. 즉 LAS 모델의 네트워크 구조는 딥러닝 모델이 특정 벡터에 주목하게 만들어 모델의 성능을 높이는 주의 집중 기법(Attention Mechanism)을 기반으로 하여, 들어오는 오디오 시퀀스 신호를 한 번에 한 문자씩 단어 시퀀스로 변환하는 시퀀스-투-시퀀스(Seq2Seq) 학습 구조를 가진다[4]. 시퀀스-투-시퀀스는 입력 시퀀스로부터 다른 도메인의 시퀀스를 출력하는 다양한 분야에서 사용되는 모델이다.

LAS 모델의 내부 구조는 입력된 음성에서 텍스트 정보를 추출하는 역할을 하는 Listener와 Listener가 압축한 정보를 바탕으로 문자 시퀀스를 생성하는 역할을 하는 Speller로 나뉜다. Listener의 경우 양방향 장단기 메모리(BLSTM) 3개 층을 겹겹이 쌓은 피라미드 RNN 구조로 구성된 딥러닝 모델이며, 이는 많은 양의 음성 프레임을 줄여주는 역할을 수행한다. 그리고 Speller 단계 이전에 주의 집중 기법을 이용하여 Speller가 문자 시퀀스를 생성할 때 입력 음성에서 어떤 부분에 집중할지 알려주는 역할을 하게 된다[5]. Speller는 순방향 장단기 메모리(LSTM) 2개 층으로 이루어져 있다.

2.2 KoBERT와 미세조정

KoBERT는 기존 BERT의 한국어 성능 한계를 극복하기 위해 SKTBrain에서 개발한 BERT의 한국어 버전이며, 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치(corpus)를 학습한 사전 훈련된 언어모델이다[6]. 최근 BERT의 파격적인 영향력으로 RoBERTa, ALBERT, BART 등 외에도 여러 BERT 모델이 사용되고 있지만, 그중에서도 KoBERT는 한국어에 대해 많은 사전 학습이 이루어져 있고 다중 분류를 쉽게 적용할 수 있기 때문에 사용자의 질의에 대한 의도를 분류하는 데에 훨씬 수월하게 접근할 수 있다.

미세조정(Fine Tuning)은 사전 훈련된 언어모델을 다른 특정 태스크에 활용하기 위한 기법이다. 본 논문에서는 문장 의도 분류 기술을 위해 KoBERT 모델에 추가적으로 네트워크를 구성하는 방식으로 미세조정을 적용한 모델을 사용했다. 해당 모델은 입력 문장의 의도를 분류하는 것을 목적으로 만들어졌기 때문에 분류기가 존재한다. 이 분류기는 질의 문장과 문장의 의도를 함께 라벨링한 텍스트 데이터를 학습시켜 입력 문장이 어느 의도 클래스에 속하는지 모델을 통해 분류하고 이후에는 분류된 의도 클래스 결과로 사전에 지정한 대답 리스트 중 하나를 무작위로 선택하여 챗봇이 응답하는 구조를 가진다[7].

2.3 타코트론2와 웨이브글로우

인간의 귀는 컴퓨터와 달리, 주파수가 높아질수록 소리의 구분이 어려워진다. 이를 위해, 멜-스케일 함수를 이용해 인간의 청각과 비슷한 형태로 주파수를 변형한 후, 스펙트로그램에 반영하는 것을 멜-스펙트로그램이라고 한다. 수식 (1)은 멜-스케일 함수의 식이다[8].

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

본 음성합성 기술은 크게 텍스트로부터 멜-스펙트로그램을 생성하는 모델인 타코트론2와 생성된 멜-스펙트로그램으로부터 음성을 합성하는 모델인 웨이브글로우로 나뉜다. 구글이 제안한 타코트론2 모델은 주의 집중 기법 기반의 시퀀스-투-시퀀스 딥러닝 구조로 이루어져 있으며, 문장-음성 쌍으로 이루어진 데이터만으로 별도의 작업 없이 학습이 가능한 종단 간 기계학습 방식의 모델이다[9]. 그리고 NVIDIA에서 제안한 웨이브글로우는 가우스 확률분포로부터 오디오 샘플을 생성하고 멜-스펙트로그램에서 고품질 음성을 합성할 수 있는 흐름 기반의 신경망으로 이루어져 있다[10].

타코트론2 모델의 내부 구조는 인코더, 디코더에 주의 집중 기법을 적용한 모듈로 이루어져 있으며, 인코더는 입력 문자를 일련 길이의 벡터로 변환하고, 주의 집중 기법으로 매 시점 변환된 벡터로부터 시간 순서에 맞게 정보를 추출하여 이를 디코더로 전달한다. 마지막으로 디코더는 이전 단계에서 얻은 정보를 바탕으로 멜-스펙트로그램을 생성한다.

웨이브글로우는 학습하는 동안 일련의 흐름을 통해 데이터셋 분포를 구형 가우스 분포로 변환하는 방법을 학습하고, 흐름의 한 단계는 수정된 WaveNet 아키텍처로 구성된다. 또한 추론하는 동안 네트워크가 반전되고 타코트론2로부터 생성된 멜-스펙트로그램으로부터 가우스 분포를 통해 오디오 샘플이 생성되는 방식으로 구성된다.

Ⅲ. 통합 인터페이스를 위한 AI 모델 구축

통합 인터페이스 구현을 위해 2.1~2.3에서 소개한 모델을 기반으로 자체 구축한 음성 데이터셋과 시나리오 데이터셋을 통해 딥러닝을 수행하였고 학습 완료된 AI 모델로 테스트를 수행하였다. 학습 및 테스트 환경은 Windows 10 운영체제와 NVIDIA RTX 3080 Ti GPU 1대로 구성된다.

3.1 음성인식 모델 구축

음성인식 모델은 사용자가 발화한 음성을 인식하여 텍스트로 변환하는 데 그 목적을 두고 있다. 고품질의 종단 간 기계학습이 가능한 음성인식 시스템을 구축하기 위해서는 다양한 자연발화 현상을 처리할 수 있는 대규모의 자연발화 말뭉치를 수집해야 한다. 이에 따라, 한국의 AI 통합 플랫폼인 AI Hub에서 제공하는 한국어 자유발화 음성 데이터셋 중 600,000개의 데이터로 구성하였다. 해당 데이터셋의 구축 분량은 1,000시간이며, 전체 화자는 총 2,000명으로 발화자의 성별 비율은 남성 923명(46%), 여성 1,077명(54%)으로 구성된다. 최종 음성 데이터 포맷은 샘플링레이트 16KHz(16,000Hz), 16bits headerless linear PCM 파일, 텍스트 전사문 등이 있다[11].

또한 학습 전에 음성 데이터 확인을 수월하게 할 수 있도록 기존 파일 형식인 pcm에서 wav로 변환하였고, AI 모델의 학습과 검증을 위해 한국어 자유발화 음성 데이터셋을 8:2의 비율로 학습 데이터 480,000개와 검증 데이터 120,000개로 나누었다. 교육 및 평가를 위한 전사문 데이터는 텍스트 전처리 모듈로 잡음과 간투어 표현 등을 위해 사용된 레이블 ‘b/’, ‘n/’, ‘/’, ‘,’, ‘*’, ‘+’ 등과 제공된 철자전사와 발음전사 중 철자전사를 삭제하여 전처리를 수행했다. 또한 전사문 텍스트의 모든 음절에 대한 라벨링을 수행하여 전사문 텍스트에 1번씩만 출현한 문자는 삭제하였다. 학습은 71번의 에포크와 64 배치사이즈로 설정하여 수행하였고, 대부분의 하이퍼파라미터는 LAS 모델의 기본 설정으로 학습하였다. 표 1은 AI Hub에서 확보한 전사문 텍스트 원본과 전처리를 수행한 후의 예시이다.

〈표 1〉 AI Hub의 데이터셋에 대한 전처리 예시

Raw Data	전처리 후 Data
너 혹시 (컴퓨터/컴퓨타)에 대해 뭐 잘 알아?	너 혹시 컴퓨터에 대해 뭐 잘 알아?
진짜 맛있어. l/내가 요즘에 가장 좋아하는 과자야. b/	진짜 맛있어. 내가 요즘에 가장 좋아하는 과자야.
어/ 나+ 나는 작년에 제주도를 두 번이나 갔거든?	어 나 나는 작년에 제주도를 두 번이나 갔거든?
맞아. 그러니까* 드라마로도 나오고 영화로도 나오는 거지.	맞아. 그러니까 드라마로도 나오고 영화로도 나오는 거지.
어/ 자세히 보면은 걔가 제일 요행을 바래.	어 자세히 보면은 걔가 제일 요행을 바래.

학습이 완료된 LAS 모델의 음성 인식률을 검증하기 위하여 평가 데이터셋을 입력 데이터로 하고, CER 값을 이용하여 성능 평가를 진행했다. 600,000개 데이터를 학습하는 동안 총 71번의 에포크가 끝날 때마다 120,000개의 검증 데이터셋에 대한 평가를 수행하였다.

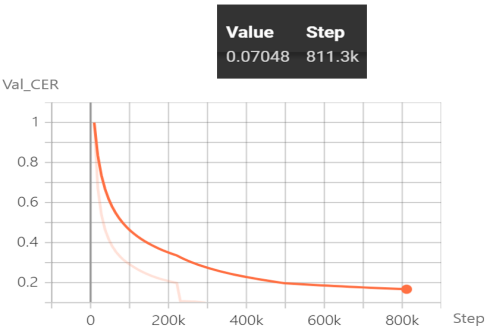
음성인식 분야에서는 오류를 판정하는 기준이 언어마다 상이하다. 영어의 경우 단어 단위로 띄어쓰기가 되기 때문에 단어의 기준이 비교적 명확하다. 한국어의 경우 단어 뒤에 조사를 사용하기 때문에 본 논문에서는 단어를 기준으로 하는 단어 오류율(WER : Word Error Rate)이 아닌 음절 오류율이(CER : Character Error Rate) 평가지표로서 적합하다고 판단되어 OpenSpeech 내부 음절 오류율 계산 모듈을 활용하여 평가지표로 사용하였다.

CER은 각 문장의 공백을 제거한 후, 음절 단위의 삭제, 삽입, 대체 개수를 계산하여 인식률을 산출한다. 즉, 원래 문장과 예측 문장 간에 다른 음절이 얼마나 있는가를 평가하는 방식이다. 아래의 수식 (2)는 CER 값에 대한 계산 방법을 의미한다[12].

$$CER(\%) = 100 \times (1 - \frac{S + D + I}{N}) \quad (2)$$

수식 (2)의 S는 음성 인식된 텍스트에 잘못 대체된 음절 수, D는 잘못 삭제된 음절 수, I는 잘못 추가된 음절 수, N은 원본 텍스트의 음절 수를 의미한다. 또한 인식률(정확도)은 음성 인식된 텍스트와 원본 텍스트의 일치하는 정도로 산정하기 때문에 1에서 CER 값을 빼고 100(%)을 곱하여 계산한다.

모델의 결과는 CER 값의 평균값으로 계산했을 때 0.07048으로 약 93%의 음성 인식률을 보였다. 아래의 그림 1은 학습이 진행되는 동안의 CER 값 변화량을 그래프로 나타낸 것이고 X축은 반복 스텝, Y축은 검증 데이터셋에 평가를 진행하여 도출된 CER 값을 의미한다.

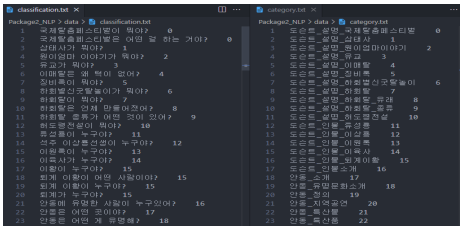


〈그림 1〉 평가 데이터셋에 대한 CER 값 변화량

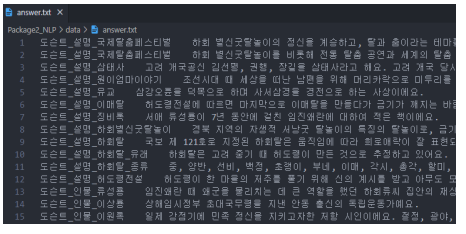
3.2 문장 의도 분류 모델 구축

문장 의도 분류 모델의 데이터는 경북 안동시에 위치한 유교랜드의 시나리오 데이터셋을 직접 구축하여 사용했다. 유교랜드 주변 관광지에 대한 소개와 유교랜드 내의 전시물에 관한 설명이 담겨 있는 100개의 질의-응답 시나리오에서 124개의 의도 카테고리를 추출하여 생성한 데이터셋이며, 구성은 질의 문장과 의도가 정숫값으로 라벨링된 텍스트 파일(질의-의도 데이터), 의도별로 특정 정숫값을 부여한 텍스트 파일(의도 데이터), 의도별로 대응하는 대답 문장이 적힌 텍스트 파일(응대 데이터) 등으로 이루어진다. 여기서 학습

에 직접적으로 사용되는 데이터는 질의-의도 데이터이며, 나머지 의도 데이터와 응대 데이터 파일은 학습 이후 추론 과정에서 사용된다. 또한 질의-의도 데이터 파일에는 124개의 의도가 라벨링되어 있으므로 학습에 사용되는 의도 클래스도 동일하게 124개로 구성된다. 따라서 모델은 문장이 입력 데이터로 들어오면 124개 의도 중 하나로 분류하게 되고, 의도 데이터와 응대 데이터 파일을 통해 사전에 설정한 의도에 맞는 챗봇의 응대 텍스트 중 하나를 무작위로 선정하여 출력값으로 내보내게 된다. 그림 2와 그림 3은 질의-의도 데이터 파일, 의도 데이터 파일, 응대 데이터 파일 등 문장 의도 분류 모델에 사용된 데이터의 예시이다.



〈그림 2〉 좌 – 질의–의도 데이터 파일 우 – 의도 데이터 파일



〈그림 3〉 응대 데이터 파일

문장 의도 분류 모델은 다중 클래스 분류를 위해 소프트맥스 회귀(Softmax Regression) 함수를 통한 성능 평가를 진행했다. 소프트맥스 함수는 보통 인공신경망의 출력층에서 사용되는 활성화 함수이지만, 본 의도 분류 모델에서 뚜렷한 평가 지표의 부재로 인해 소프트맥스 값으로 대체하였다.

소프트맥스 함수는 선택해야 하는 선택지의 총개수를 k라고 할 때, k 차원의 벡터를 입력받아 각 클래스에 대한 확률을 추정한다. 즉, 시나리오에서 의도의 개수가 총 100개라고 하면 입력 데이터로 문장 하나가 들어왔을 때, 100차원의 벡터를 입력받아 모든 의도에 대해 해당 문장이 분

〈표 2〉 검증 결과

순번	질의 문장	예측한 의도	Softmax Value
1	유교랜드는 뭐하는 곳이야?	유교랜드_설명	0.9443
2	유교랜드 주소가 어떻게 돼?	유교랜드_위치	0.4498
3	주변에 가볼만한 곳 있니?	주변관광_소개	0.9805
4	근처 맛집 추천해줄래?	주변관광_맛집	0.9367
5	안동은 뭐가 유명해?	안동_유명문화소개	0.9101
6	안녕 반가워.	도슨트_인사	0.9597
7	독립기념관은 어디야?	주변관광_독립운동기념관_위치	0.4172
8	독립기념관은 여기서 얼마나 걸려?	주변관광_독립운동기념관_이동시간	0.8295
9	하회마을은 여기서 얼마나 걸려?	주변관광_하회마을_이동시간	0.5165
10	하회탈에 대해 알려줄래?	도슨트_설명_하회탈	0.3947
11	너는 누구니?	도슨트_자기소개	0.9928
12	군자마을이 어디야?	주변관광_군자마을_위치	0.7530
13	이황이 어떤 일을 했어?	도슨트_인물_퇴계이황	0.9688
14	독립운동기념관은 뭐야?	주변관광_독립운동기념관_설명	0.7303
15	맛집 추천해줄래?	주변관광_맛집	0.9874
16	유교는 뭐야?	도슨트_설명_유교	0.7785
17	국제탈춤페스티벌에 대해 알려줘.	도슨트_설명_국제탈춤페스티벌	0.9887
18	석주 이상룡선생에 대해 알려줘.	도슨트_인물_이상룡	0.9527
19	안동은 어떤게 유명하니?	안동_유명문화소개	0.9638
20	안동 특산물 알려줘.	안동_특산물	0.9491

류될 확률 100개를 계산하고 그중 가장 확률이 높은 의도와 그 확률값을 출력하는 방식이다[13].

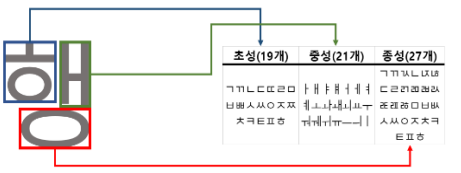
해당 모델은 반복 실험을 통한 소프트맥스 값으로 평가를 진행하였다. 검증 데이터는 질의-응답 NLP 모델 학습 시에 사용한 학습 데이터셋 중 일부 문장들의 조사 및 단어를 변경하거나 평서문을 의문문으로 또는 의문문을 평서문의 형식으로 변형하는 등의 방식으로 검증 데이터를 구축하였다. 반복 실험은 총 20개의 검증 문장 데이터에 대한 의도를 예측 분류하여 도출된 각 소프트맥스 값들을 산술평균값으로 계산하였고, 그 결과 20개 문장에 대한 평균 소프트맥스 값은 0.8202로써 약 82%의 의도 분류 정확도를 보였다. 아래 표 2는 20개 문장에 대한 검증 결과를 나타낸 것이다.

검증 결과를 살펴보면 20개 문장에 대한 의도가 올바르게 분류되었음을 볼 수 있다. 또한 소프트맥스 확률값이 다른 문장들에 비해 비교적 낮은 2번, 7번, 10번 문장 3개가 존재했지만, 의도는 올바르게 예측하였다. 이에 따라 의도 클래스에 대한 세분화가 필요한 것으로 보인다.

3.3 음성합성 모델 구축

음성합성 모델은 음성인식 모델과는 다르게 학습 데이터로서 화자 한 명의 음성이 필요하다. 따라서, 음성합성 모델의 학습 데이터는 위메이크보이스의 TTS 프로그램을 사용하여 텍스트를 음성데이터로 직접 변환하여 구축했다. 이 데이터셋은 여성 성우 한 명의 목소리로 약 12시간 분량의 음성 녹음한 파일 12,853개와 전사문 스크립트로 구성된다. 최종 음성 데이터 포맷은 샘플링레이트 44.1KHz(44,100Hz), WAV 파일, 텍스트 전사문 등이 있다. 음성 데이터셋의 학습 시 44.1KHz의 샘플링레이트로 구성된 음성데이터를 학습하면 그만큼 더 많은 데이터를 학습해야 하므로 학습의 속도에 영향을 끼쳐 샘플링레이트를 절반으로 재조정하여 모든 음성 파일을 22,050Hz로 맞추었다.

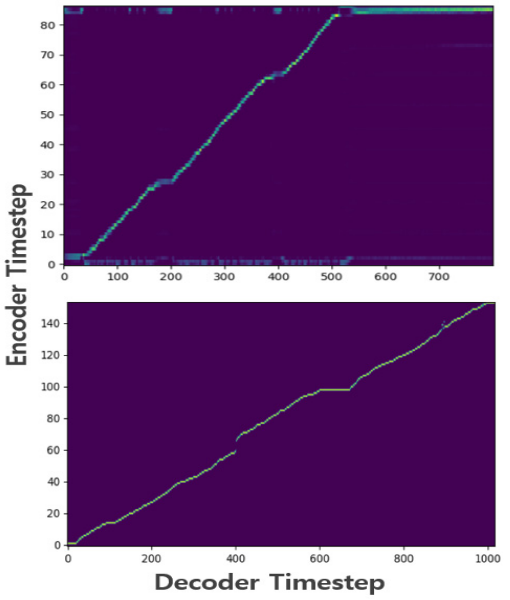
음성합성 모델은 타코트론2와 웨이브글로우 등의 2개 모델을 통해 추론 해야 하므로 각각 딥러닝을 수행해야 한다. 학습하기 이전에 전사문 텍스트에 대하여 숫자, 특수문자, 영어 등을 한글로 변환하고 음절 단위의 문자를 초성, 중성, 종성 등의 음소 단위로 변환하는 작업으로 전처리를 수행하였다. 한글의 음절은 19개의 자음으로 구성된 초성, 21개의 모음으로 구성된 중성, 27개의 자음으로 구성된 종성으로 이루어져 있으므로 그 경우의 수가 무려 10,773개이므로 데이터가 적다면 학습 데이터에 많은 음절이 포함되지 못하여 새로운 음절에 모델이 대응하지 못하는 현상이 발생한다. 이를 방지하고자 음소를 기본 입력 단위로 하여 모델 학습을 수행했다. 그림 4는 한글의 음소 단위를 분류해놓은 것이다.



〈그림 4〉 한글의 음소 단위

음성합성 모델의 경우, 타코트론2와 웨이브글로우 모델을 학습 데이터 12,000개로 하여 각각 800,000번, 602,000번의 반복 스텝(Step) 동안 500번의 에포크와 16의 배치사이즈로 학습을 진행하였다.

학습이 완료되고 합성음의 품질을 확인하기 위한 수단으로 오디오 정렬(Alignment) 그래프를 출력하여 확인하였다. 그래프의 X축은 디코더의 시간 흐름(Decoder Timestep), Y축은 인코더의 시간 흐름(Encoder Timestep)을 의미한다. 즉, 인코더의 시간 흐름은 입력 문자를 단계별로 읽고 상태 벡터를 출력하는 순차적 흐름이고, 디코더의 시간 흐름은 모든 상태 벡터를 읽고 오디오 프레임을 단계별로 생성하는 흐름이다.



〈그림 5〉 음성을 정렬하는 오디오 정렬 그래프

따라서, 오디오를 정렬한 그래프가 그림 5의 하단에 있는 그래프처럼 우상향으로 일직선을 그리면서 선이 선명한 것은 입력 문자에 대해 순서대로 초점을 맞추어 올바른 오디오 프레임을 생성하여 높은 품질의 음성을 합성하고 있는 것을 의미한다.

또한 그림 5에서 상단의 그래프는 「내일은 오전부터 늦은 오후 사이에 전국 대부분 지역에서 소나기가 내리겠습니다.」라는 문장으로 학습 완료된 모델에서의 정렬 그래프를 생성한 것으로, 이는 우상향으로 일직선을 그리고는 있지만 X축(Decoder Timestep)이 0에서 500일 때는 그래프 상단에 작은 점들이 출현하였고, 500 이후부터는 우상향이 아닌 수평선을 그린다. 여기서 수평선이 그어져 있는 부분은 컴퓨터가 학습 시 사용한 음성 데이터에서 음성발화가 끝나는 부분을 제대로 인지하지 못한 채로 학습이 계속 진행되어 나타나는 것이며, 작은 점 모양은 잡음을 뜻한다. 따라서, 잡음 처리방안과 음성발화의 종료 시점을 정확하게 구분하기 위한 고찰이 필요하다.

Ⅳ. 통합 인터페이스

본 논문에서 제안한 AI 음성 안내 챗봇 시스템은 앞서 생성한 AI 모델들을 API 형태로 모델을 불러오고 모든 추론 과정을 거치는 통합 인터페이스로 구현한다. 그 과정은 사용자의 발화 음성 데이터가 입력되면 이를 감지하여 LAS 모델을 거쳐 텍스트 데이터로 전환되고, 대용량의 한국어 말뭉치 언어모델인 KoBERT에 미세조정을 적용한 모델을 통해 입력 텍스트에 대한 의도를 분류하고 사전에 설정한 의도별 응대 리스트 중 하나를 무작위로 출력한다. 끝으로, 타코트론2와 웨이브글로우 모델에서 응대 텍스트를 입력 데이터로 하여 순차적 추론을 진행한 후 AI의 응대 음성을 내보낸다. 그림 6은 통합시스템의 인터페이스 개요도를 도식화한 것이다.

〈그림 6〉 AI 한국어 음성 챗봇 시스템 인터페이스 구성도

4.1 통합 인터페이스 구현

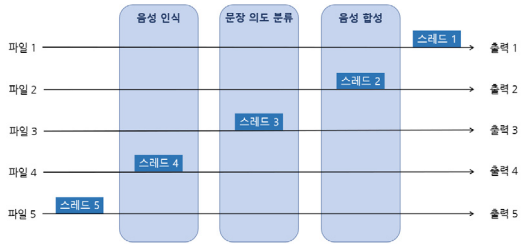
통합 인터페이스는 Python 언어를 기반으로 구축하였으며, 파이썬 스크립트 파일인 'Interface_Thread.py'를 통해 본 시스템 전체를 동작시킨다. 해당 스크립트를 실행하면 추론 수행에 필요한 각 라이브러리를 불러오고 학습된 각 모델(음성인식 모델, 문장 의도 분류 모델, 음성합성 모델)을 API의 형태로 불러온다. 이후, 파이썬의 Watchdog를 활용해 감시자 객체(Observer)와 이벤트 핸들러(Event Handler) 객체를 생성한다. 감시자 객체는 무한루프로 구성되어 지정한 디렉터리 내부에서 발생하는 이벤트를 실시간으로 감지하는 역할을 하고, 이벤트 핸들러 객체는 파일의 확장자가 wav인 파일이 생성되면 해당 이벤트를 활용하여 모델별로 순차적 추론을 수행한 후 wav 파일 형태로 결과 디렉터리에 저장한다.

이 시스템은 프로세스 내에서 실제로 작업을 수행하는 주체를 의미하는 스레드와 앞과 뒤에서 데이터를 처리할 수 있는 데크를 활용한 중복 방지 기능을 구현하여 GPU가 1대인 경우에 추론 중 GPU 연산이 중복되지 않도록 구현했다. 그림 7은 스레드와 데크를 활용해서 구현한 중복 방지 기능을 도식화하여 나타낸 것으로 예를 들어, 음성 파일 5개가 동시에 감지되면 5개의 스레드가 생성되며 스레드별로 카운트를 부여한다. 동작 방식은 스레드 1이 음성인식을 시작할 때 데크에 스레드 1 카운트를 저장하고 음성인식이 끝나면 스레드 1 카운트를 데크에서 삭제한다. 이 때, 대기하고 있던 스레드 2가 곧바로 음성인식을 시작

〈표 3〉 통합 인터페이스 테스트 결과

순번	질의 음성	음성인식	출력 응대 음성	추론 소요 시간
1	이름이 뭐야?	이름이 뭐야?	안녕하세요. 저는 유교랜드의 안내요원을 맡고있는 아무개입니다.	1.41초
2	이매탈은 왜 턱이 없나요?	이메탈은 왜 턱이 없나요?	허도령 전설에 따르면 마지막으로 이매탈을 만들다가 금기가 깨지는 바람에 미처 완성하지 못하고 죽어 미완성인 채로 남게 되었다고 해요.	1.21초
3	국학진흥원이 어떤 곳인지 설명해줘.	국학 지능원이 어떤 곳인지 설명해줘.	유교 기록문화재들을 안전하고 과학적으로 보존하는 곳이에요.	1.11초
4	안녕?	안녕?	네 안녕하세요. 반가워요.	0.96초
5	여기서 하회마을까지 얼마나 걸려?	여기서 하회마을까지 얼마나 걸려?	여기서 하회마을은 차로 40분 정도 걸려요.	0.88초
6	안동에 볼만한 공연이 어떤 게 있어?	안동에 볼만한 공연이 어떤 게 있어?	안동 지역공연으로는 퇴계연가, 석주 이상룡, 원이엄마, 삼태사 등이 매년 상연되고 있어요.	0.88초
7	안동은 어떤 곳이야?	안동은 어떤 곳이야?	한국 정신문화의 수도로, 우리나라 유일의 지역학인 안동학이 존재하는 곳이에요. 일제강점기 당시 전국에서 가장 많은 독립운동가를 배출한 곳이기도 해요.	0.76초
8	퇴계 이황이 누구야?	퇴게 이황이 누구야?	퇴계는 이황 선생님의 호이며, 조선 전기 성균관 대사성, 대제학, 지경연 등을 역임한 학자에요.	0.89초
9	유교가 뭐야?	유교가 뭐야?	삼강오륜을 덕목으로 하며 사서삼경을 경전으로 하는 사상이에요.	0.89초
10	국제 탈춤페스티벌이 뭐야?	국제 탈춤페스티벌이 뭐야?	하회 별신굿탈놀이의 정신을 계승하고, 탈과 춤이라는 테마를 바탕으로 열리는 가을 축제예요.	1.05초

하는 방식이다. 따라서, 스레드별로 고유 카운트를 부여함으로 인해 GPU 연산이 서로 중복되지 않고 추론을 수행할 수 있다.



〈그림 7〉 스레드와 데크를 활용한 중복 방지 기능

4.2 통합 인터페이스 테스트 결과

통합 인터페이스의 동작을 테스트하기 위해 5개의 음성 파일을 입력 데이터로 하여 모든 모델의 추론 과정을 수행하고 최종적으로 정확한 응대 음성을 출력하는가를 확인했다. 통합 인터페이스의 테스트 결과를 나타낸 표 3을 살펴보면, 음성인식 결과로 2번에서 이매탈을 이메탈, 3번에서 국학진흥원을 국학 지능원, 8번에서 퇴계를 퇴게 등 한국어 음운 규칙을 구사하지 못하여 잘못 음성 인식하였으나, 10개 문장에 대해 의도를 정확하게 분류하여 올바른 응대 음성을 출력했음을 알 수 있다. 이는, 음성인식 결과가 기존 문장과 비교해 틀린 부분이 있더라도 대용량의 한국어 말뭉치 언어모델인 KoBERT를 활용하여 문장 의도 분류 모델을 학습했기 때문에 틀린 단어를 유추할 수 있기 때문이다.

또한, 통합 인터페이스는 스레드와 데크를 통한 중복 방지 기능의 구현으로 한 대의 GPU에서도 사용자의 음성이 입력되고 AI의 응대 음성이 출력되는 시간이 평균 1.004초 소요되는 것을 확인하였다.

Ⅵ. 결 론

본 논문에서는 사용자가 음성을 발화하면 이를 감지하여 인식하고, 사용자가 말하는 바의 의도를 분류하여 올바른 응대 음성을 다시 출력하는 박물관의 안내를 위한 시나리오 기반의 AI 음성 챗봇 시스템을 제안하였다. 이를 위해 음성합성 모델의 음성데이터와 문장 의도 분류 모델의 시나리오 데이터셋을 자체적으로 구축하고 전처리하여 AI 모델을 학습하고 검증하였다. 이를 통해 생성된 각 AI 모델을 모듈별로 통합한 통합 인터페이스를 구현하였고, 이 시스템은 중복 방지 기능이 포함되어 단일 GPU 환경에서도 추론 과정에 있어 GPU 연산이 중복되지 않고 빠른 속도로 추론할 수 있다. 하지만, 통합 인터페이스의 테스트 결과로 음성인식으로 도출된 문장에서 한글의 음운 규칙을 정확하게 구사하지 못하는 부분을 위해 맞춤법 교정 기능을 구현하여 정의할 필요가 있다.

향후 연구에서는 문장 의도 분류 모델에서 소프트맥스 값이 낮게 나오는 부분이 존재하여 의도 클래스를 더 세분화하는 방안으로 성능을 높일 예정이며 또한, 음성합성 모델의 잡음 처리 기능을 향상시키고 음성이 끝나는 시점을 정확하게 판단하기 위해 오디오가 끝나는 부분에 침묵 패딩을 추가하고 파일 전체에 트리밍을 수행하는 등의 방식으로 고도화할 예정이다.

참 고 문 헌

[1] 정현숙, 이기길, 이대경, 김정민, “조선대학교 박물관 모바일 도슨트 어플 설계 및 구현”, 융합정보논문지, 제8권, 제5호, pp.121-129, 2018.

[2] 김종건, 허정윤, “비대면 전시의 사용자 경험 개선을 위한 스마트 도슨트 챗봇”, 한국HCI학회 학술대회, 제21권, 제1호, pp.184-187, 2021.

[3] https://github.com/openspeech-team/openspeech
[4] William Chan, Navdeep Jaitly, Quoc V. Le and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4960-4964, 2016.
[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, “Attention Is All You Need”, Advances in neural information processing systems 30, 2017.
[6] https://sktelecom.github.io/project/kobert/
[7] https://github.com/nawnoes/WellnessConversation-LanguageModel
[8] https://en.wikipedia.org/wiki/Mel_scale
[9] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis and Yonghui Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4779-4783, 2018.
[10] Ryan Prenger, Rafael Valle and Bryan Catanzaro, “WaveGlow: A Flow-based Generative Network for Speech Synthesis”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3617-3621, 2018.
[11] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=123
[12] 민소연, 이광형, 이동선, 류동엽, “한국어 특성 기반의 STT 엔진 정확도를 위한 정량적 평가방법 연구”, 한국산학기술학회논문지, 제21권, 제7호, pp.699-707, 2020.
[13] 김영원, 이수진, “소프트맥스 함수 특성을 활용한 침입탐지 모델의 공격 트래픽 분류성능향상방안”, 융합보안논문지, 제20권, 제4호, pp.81-90, 2020.

저 자 소 개

정 선 우(Sun-Woo Jung)
·2021년 2월 : 동의대학교 생산정보기술공학과 (공학사)
·2022년 3월～현재 : 동의대학교 부산IT융합부품연구소 연구원
<관심분야> : 인공지능, 빅데이터, 자연어처리

최 은 성(Eun-Sung Choi)
·2017년 2월 : 동아대학교 전자공학과 / 스마트그리드 연계전공(공학사)
·2022년 3월～현재 : 동의대학교 인공지능학과(석사과정)
·2021년 7월～현재 : 동의대학교 부산IT융합부품연구소 연구원
<관심분야> : 인공지능, IoT, 자동계측

안 선 규(Seon-Gyu An)
·2005년 2월 : 동의대학교 정보통신공학과 (공학사)
·2012년 7월～현재 : 동의대학교 부산IT융합부품연구소 선임연구원
<관심분야> : 인공지능, IoT, 자동계측

강 영 진(Young-Jin Kang)

·2013년 8월 : 동서대학교 정보통신학과 (공학사)
·2020년 8월 : 동서대학교 유비쿼터스IT (공학석사, 박사)
·2021년 3월~2022년 2월 : 동서대학교 소프트웨어 융합대학 초빙교수
·2022년 3월~현재 : 동의대학교 인공지능그랜드ICT연구센터 연구교수
<관심분야> : 인공지능, 암호이론

정 석 찬(Seok-Chan Jeong)

·1987년 2월 : 부산대학교 기계설계학과 (공학사)
·1993년 3월 : 오사카부립대학 경영공학과 (공학석사, 박사)
·1993년 2월~1999년 2월: 한국전자통신연구원 선임연구원
·1999년 3월~현재 : 동의대학교 e비즈니스학과 교수
·2019년 1월~현재 부산IT융합부품연구소 소장
·2020년 7월~현재 인공지능그랜드ICT연구센터 센터장
<관심분야> : 정보시스템, IoT 융합, 빅데이터, 클라우드, 블록체인, 인공지능