|  |  |  |  |
| --- | --- | --- | --- |
| **Capstone Design Result Report** | | | |
| **Project Name** | **트렌스포머 기반 speech-to-text와 ChatGPT를 이용한 화법 교정 서비스** | | | |
| **Participatory Department** | 경북대학교 컴퓨터학부 | **Project Representative** | 김승호 교수 | |
| **Participants** | 참여인원 6명. 책임교수: 1명, 학부생: 5명 | | | |
| **Period** | 2023. 3. 8. ∼ 6. 19.(3months) | **Type** | The Middle | |
| **Propulsion Background** |  | | | |
| ○ 코로나19 사태가 최근에 진정되었으나, 그동안 오프라인에서의 의사소통이 줄어듦에 따라 공감대를 형성하는 화법 구사에 대한 어려움이 증가함.  ○ 최근 ChatGPT가 선풍적인 인기를 끌게 되었고, API를 이용한 서비스가 많은 관심을 받게 되었음.  ○ ChatGPT를 이용하여 여러 사람 간의 대화에서 화법의 문제점에 대해 알고, 이를 교정하여 대화를 하는 방식을 추진하게 되었음. | | | | |
| **Objectives and**  **Content** |  | | | |
| ○ 화자 간의 대화가 담겨 있는 음성 파일을 텍스트화(speech-to-text)  - 음성 파일에서 어느 화자가 발화하였는지 분리 시켜주는 모델 구현  - 해당 음성을 Text화 시켜주는 모델 구현  ○ 해당 대화 내용 Text를 API를 사용하여 ChatGPT에게 전달해, 화법의 문제점과 개선 대화 내용을 출력하는 프롬프트 설계 및 구현  ○ Text화 된 대화 내용과 문제점, 개선사항이 담겨 있는 UI 구성 | | | | |
| **Project Result** |  | | | |
| ○ 향후 더욱더 전문적인 서비스화의 기초 토대 마련  - 전문적인 대화법을 ChatGPT에게 학습을 시킨다면, 부부간의 건강한 화법교정 혹은 취업면접을 위한 화법교정 등과 같이 더욱더 전문화된 서비스 제공이 가능하다. | | | | |

**1. Background of Project Performance**

현재 가장 최신 기술 중의 하나인 생성 모델을 활용한 프로젝트를 진행하기로 조원들간의 합의 후 가장 어울리는 아이디어들에 대한 토의를 진행했다. 개발을 위한 개발이 아닌 정말로 사용할 수 있는 필요에 의한 개발을 하고자 했다. 이에 우리 사회가 겪은 코로나로 인한 영향에 대한 고민을 할 수 있었고, 코로나로 인한 언택트 사회에서 우리가 잃었던 것이 무엇인지 알 수 있었다. 사람 간의 만남이 적에 짐에 따라 의사소통으로 인한 사회적 공감대를 여러 매체나 풍자 개그 등에서 파악할 수 있었다.

이런 문제를 해결하기 위한 서비스에 대해 생각해보았다. 회의를 통해 다양한 아이디어들을 나눠볼 수 있었고, 그중 가장 필요성이 높고 사용가능성이 높아 보이는 아이디어를 선정하게 되었다. 생성 AI 기술을 활용해 화법을 교정해주는 서비스는 아래에 기술된 필요성으로 인해 개발 대상으로 선정되었다.

우리가 제시하는 화법 교정 서비스의 필요성은 다음과 같다.

* 녹음된 음성 파일을 통해 제공하는 교정 문맥이나 대화에서 필요에 맞지 않는 경우의 화법을 교정해준다. 이는 다른 사람에게 자신의 발화를 제공하는 것이 아니라 AI 서비스에게 제공하는 것이기에 교정자의 부담을 줄여줄 수 있다.
* 코로나로 인한 언택트 사회의 단절, 사회의 양극화로 인한 세대간 사용 언어 차이 등을 이해할 수 있게 교정된 화법의 이유를 제시하며 향후 자신이 갖추어야 할 화법의 지향성을 배울 수 있게 한다.
* 위 효과들로 인해 화법의 개선이 이뤄진다. 발전된 의사소통 능력을 통해 대인관계에 있어 필요한 자신감을 얻을 수 있고, 사용자의 삶의 질을 개선하고 궁극적의 사용자의 행복감을 고양 수 있다.

**2. Project Objectives**

위 필요성을 바탕으로 우리는 음성을 텍스트로 만들고 그 텍스트를 ChatGPT를 활용해 교정된 화법으로 제시하는 기능을 구현하여야 한다. 이는 UI를 통해 사용자에게 제공된다. 이 간단한 요구사항을 통해 우리가 해야할 일들을 파악하고 전반적인 시스템의 구조를 디자인 했다. 이후 그 결과를 바탕으로 아래와 같은 과정을 거쳐 시스템을 개발하고자 했다.

첫째로 음성 파일을 다화자 분할할 수 있는 AI 모델, 그 분할된 화자를 바탕으로 음성을 텍스트로 전사해주는 모델을 알아보고 그 구조를 파악해 사용할 수 있어야 한다.

둘째로 UI를 활용한 서비스 제공 방식이기에, ChatGPT 서비스를 API 형식으로 활용하는 방법을 조사하고 우리가 전사한 텍스트를 넘겨주는 방법을 파악해야 한다. 이 과정을 통해 “프롬프트 디자인”으로 원하는 결과를 만들어 가야 한다는 것을 알 수 있었다. 이에 대해 조사한 결과 생성모델의 한계인 결과값이 무분별하게 나온다는 것을 확인하고 이를 개선할 수 있는 방법을 확인 해야한다.

셋째로 우리가 제공하는 서비스의 UI를 구성하고 그를 구현할 방법을 선정한다. 직관적이고 사용성이 좋은 것을 가장 중점으로 삼고 UI 구성을 진행해야 한다.

위와 같은 일련의 과정을 통해 최종적으로 화법을 교정해 주는 서비스를 만들어 내고자 하며 이는 우리 프로젝트의 목표로 제시할 수 있다.

**3. Project performance result**

본 프로젝트는 크게 2가지 파트로 나눌 수 있다. 첫 번째는 음성인식 및 화자별 발화전사, 두 번째는 ChatGPT 프롬프트 엔지니어링 및 User Interface 구현이다. 먼저 음성인식 및 화자별 발화전사 구현에 대해 설명을 하면, 화자 구분없이 speech-to-text만을 위해 OpenAI사의 transformer 기반 Whisper 모델을 사용하였다. Whisper 모델은 약 68만 시간의 음성파일을 사용하여 학습하였기 때문에, 다양한 언어의 음성발화를 텍스트로 변환이 가능하고, 무엇보다 소음, 엑센트, 사투리 등이 있어도 높은 정확도의 텍스트 변환능력을 보여준다.

그러나, 본 프로젝트를 위해서는 화자별로 발화를 전사를 해야하기 때문에 Whisper 모델만으로는 충분하지 않았다. 그래서 추가적인 화자구분 알고리즘이 필요했다. Whisper 모델은 '[시작시간-종료시간], “발화“ ‘ 형식으로 결과를 반환하는데, 예를 들어, [1.7, 5.4], “안녕하세요. 저는 홍길동입니다.” 와 같다. 이 예시는 “안녕하세요. 저는 홍길동입니다.”라는 발화가 음성파일의 1.7초에서 시작하여 5.4초에 종료됨을 의미한다.

Whisper로부터 주어진 발화 시작, 종료 시간을 바탕으로 음성파일을 분할한다. 그 분할된 음성파일들 각각을 librosa모델을 사용하여 로드를 한 뒤에 embedding을 한다. Embedding 개념은 자연어 처리에서 단어의 특징을 벡터화 하는 것에서 빌려온 것으로, 음성파일의 관점에서 해석한다면, 화자1과 화자2가 동일하게 “안녕”이라고 발화를 했을 때, 두 화자 모두 동일한 단어를 발화했지만 음성파일의 벡터는 달라야 한다. 즉, 어떤 단어를 발화했느냐가 아니라 화자의 아이덴티티에 따라서 음성파일의 벡터화가 진행되어야 한다. 정확도 높은 embedding을 위해서 Convolutional Neural Network(CNN) 기반의 음성파일 embedder을 사용했다. 그 후 Embedding 처리가 된 256차원의 벡터들에 대해서 Agglomerative clustering을 적용하여 화자의 아이덴티티별로 그룹화를 진행한다. 최종적으로는 앞서 언급했던 Whisper 모델로부터 받은 대화전사와 clustering의 결과를 조합하여, 화자별로 발화 전사가 된 결과를 ChatGPT로 넘겨준다. 화자별 발화 전사의 결과는 그림 1과 같다.

A screenshot of a computer

Description automatically generated with low confidence

<그림1: 화자별 발화전사>

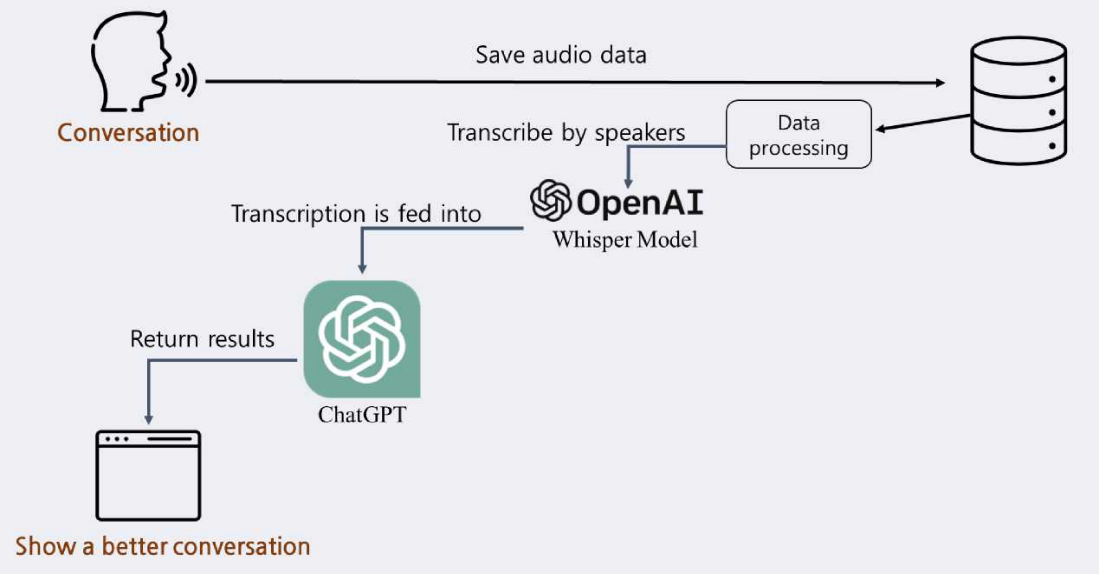
ChatGPT에서는 입력 받은 대화 내용을 바탕으로 개선된 대화내용을 추천해준다. 이 때 ChatGPT는 입력 프롬프트에 따라 결과의 질이 크게 달라지기 때문에 좋은 개선된 대화내용을 얻기 위해 다양한 프롬프트를 작성해 보았다. 그 결과 대화내용 교정 프롬프트 작성 시 ‘공격적인 화법 교정’, ‘맥락에 맞지 않는 내용 교정’ 등과 같이 교정할 부분에 대한 자세한 예시가 필요 하였고, ‘i-message’와 같은 구체적인 교정방법이 필요하였다. 그리고 ChatGPT의 결과를 웹에 업로드 하기 위해서는 일정한 형식이 필요하여 ‘문자열’, ‘Json’, ‘배열’ 등 다양한 형식으로 ChatGPT에게 반환해 달라고 프롬프트를 작성한 결과, 배열이 가장 결과가 좋아 배열 형식으로 웹으로 값을 전달하도록 하였다.

UI부분은 음성파일 또는 유튜브 링크를 입력 받는 부분, 입력 받은 오디오를 재생하는 부분, 전사된 대화내용 및 ChatGPT가 제안하는 개선된 화법을 출력하는 부분으로 구성되어 있다. 음성기록에서는 화자(Speaker 0, 1, 2..)와 발화 시간, 전사한 발화내용이 출력된다. ChatGPT가 제안하는 개선된 화법에는 음성 기록 부분에서 교정이 필요한 부분만 출력되며 화자, 발화시간, 이유, 개선된 발화 내용이 출력된다. 결과는 다음의 그림 2와 같다.



<그림2: ChatGPT가 제안하는 화법>

전체 작동 방식은 다음의 그림 3과 같다. 먼저 사용자로부터 음성파일 혹은 유튜브 링크를 받으면 전처리 단계를 거친 후 WAV 파일 형식으로 변환한다. Whisper 모델을 사용하여 음성파일의 발화전사 및 발화시작-종료 시간을 얻는다. Whisper로부터 받은 발화시작-종료 시간에 맞게 음성파일을 segmentation한다. 앞서 언급한 방식을 통해, 화자별 발화전사를 한 후 이것을 화법교정에 대해서 사전 학습이 된 ChatGPT에게 프롬프트와 함께 넘겨주어 화법 교정 내용을 받아 사용자에게 그 결과를 보여준다.



<그림 3: 전체 시스템 다이어그램>

**4. Expected effectiveness and utilization plan**

1번 항목에 제시한 필요성들을 필두로 인해 사용자가 위 서비스를 사용하게 된다면 얻을 수 있는 기대 효과이다. 사용자의 발화에서 문제점을 찾아 그 근본적인 원인을 파악해주고 개선된 화법을 제시하며 사용자가 단순히 이 제시된 교정을 받아드리는게 아니라, 이해하고 원초적인 문제를 해결할 수 있게 한다.

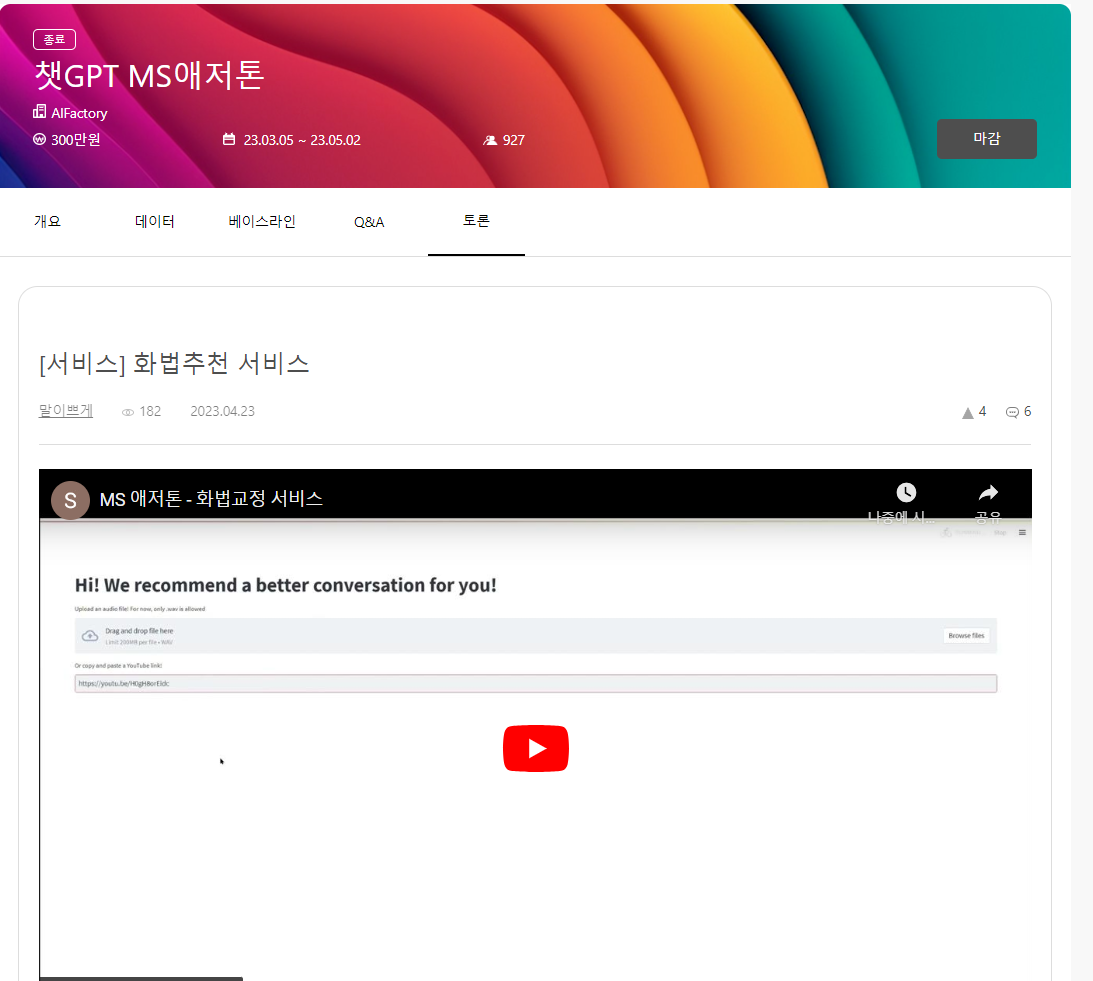
이런 교정 과정을 통해 발전된 의사소통 능력을 발전시킬 수 있다. 따라서 자연스럽고 문제 없는 의사소통을 통해 대인관계에 있어 필요한 자신감을 얻을 수 있다. 결과적으로 서비스의 사용을 통해 사용자의 삶의 질을 개선하고 궁극적으론 사용자의 행복감을 고양 수 있다는 점이 사용자가 얻을 수 있는 기대효과이다.

다양한 방면에서 활용되는 ChatGPT기술을 사용한다는 점도 중요하다. 현재 생성 모델은 다양한 우려가 있는 기술이다. 하지만 특정 기능을 위한 도구로 생성 모델을 사용하며, 이때 이 기능은 사람을 위한 것일 때 좋은 방향으로 기술을 사용하는 방향성의 제시가 될 수 있다. 이러한 관점에서 우리가 제공하는 서비스는 좋은 방향성의 사례를 잘 보여줄 수 있다는 점이 중요한 기대 효과 중 하나이다.

이러한 화법 교정은 현재 학원 등에서 전문적으로 봐주는 사람이 있을 정도로 현재 사회에 꼭 필요한 서비스라고 할 수 있다. 점점 많은 사람들이 이 사회적 문제에 대한 공감을 할 수록 이를 해결하기 위해 스스로 위 서비스를 사용할 사람들이 많아질 것이라 예상된다. 우리는 합리적인 비용으로 이 서비스를 제공할 수 있다. 이를 통해 사용자는 비용을 아끼고 사회 전반적으로 의사소통 능력을 배양해 문제에 대한 전반적인 개선에 기여할 수 있다.

**5. Main product**

* AIFactory에서 주관하는 챗GPT MS 에저톤에 참가하여 ‘화법 추천 서비스’라는 주제로  
  참가함. (https://aifactory.space/competition/2290/discussion/264)



<그림 4: 공모전 참가>

**6. Participating Personnel(Detail)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Professor** | **Affiliation** | **Department of Computer Science at the college of IT** | | **Name** | **김승호** |
| **Participants**  **(Students)** | **Affiliated**  **Department** | **Degree**  **(Gender)** | **Student Number** | **Name** | **Assigned Task** |
| **컴퓨터학부** | **학부(남)** | **2019117923** | **배완기** | **화자별 음성인식 모델 구현** |
| **컴퓨터학부** | **학부(남)** | **2018115065** | **강창명** | **화법교정 관련 논문 리딩 및 프롬프트 엔지니어링** |
| **컴퓨터학부** | **학부(남)** | **2018114449** | **김동현** | **UI 구현 및 프롬프트 엔지니어링** |
| **컴퓨터학부** | **학부(남)** | **2018116698** | **최성혁** | **화법교정 관련 논문 리딩 및 UI 구현** |
| **컴퓨터학부** | **학부(남)** | **2018115122** | **김기욱** | **화자별 음성인식 모델 구현** |