|  |
| --- |
| **1. 주제**  시각/청각 장애인을 위한 딥러닝 기반 다차원 음성 인지 시스템 제안  **분반, 팀, 학번, 이름**  가반, 6팀, 20251761, 마진성 |



|  |  |
| --- | --- |
| **2. 요약**  본 제안서는 시각/청각 장애인이 겪는 대화의 불평등을 해소하고, 딥러닝 기술을 활용하여 음성 데이터를 기반으로 얻을 수 있는 다 차원적인 맥락 정보를 파악하는 것을 목표로 한다.  이를 위해 크게 4가지 단계, 잡음 제거 및 특징 추출, 화자 식별 및 그룹화, 다차원 맥락 정보 추출, 통합 정보 전달을 중점으로 두어 적합한 모델을 제안함으로써 문제들을 해결한다.  단순한 음성 내용 전달을 넘어 화자의 성별, 실시간 감정 및 어조와 같은 비언어적 단서를 제공함으로써, 동일한 문장이라도 어조에 따라 의미가 이중적으로 이해될 수 있는 문장까지 정확히 해석하도록 할 수 있을 것이다. | **3. 대표 그림**  **대화 내용의 화자를**  **구분하고 감정을 매핑**  남성: “(화난 어조로)  정말 대단하다.”  여성: "(기쁜 어조로) 정말 대단하다!”  정말 대단하다.  정말 대단하다! |

|  |
| --- |
| **4. 서론**  **배경**  인간 사회에서 대화는 단순한 정보를 넘어 여러가지 의미를 가지는 중요한 요소이다. 대화는 음성적인 내용 이외에도 시각적인 표정, 몸짓이나 목소리의 높낮이 등의 어조와 같이 복합적인 요소를 통해서 맥락을 파악해야 하는 고차원적인 과정이다. 그러나 시각장애인과 청각장애인 같이 이러한 핵심 요소 중 하나를 이해하는데 어려움이 있는 경우, 대화의 모든 맥락을 완벽하게 인지하는 데 있어 불편을 겪을 수 있다.  실제로 한 청각장애인의 말에 따르면 소리만으로는 대화를 이해하기 어려워 다른 사람의 입 모양을 봐야 무슨 말인지 이해할 수 있는데, 전화 통화는 물론, 마주 앉은 사람도 얼굴을 바라보지 않고 대화를 나누기 때문에 대화의 장벽을 느낀다고 한다  따라서 본 프로젝트는 이러한 장애인의 대화 불평등을 해소하기 위해 딥러닝 기술을 활용하여 음성 데이터를 기반으로 얻을 수 있는 다 차원적인 맥락 정보를 파악하는 시스템을 구축하고자 한다.  **문제 정의**  다음은 프로젝트에서 해결해야 할 문제들이다.  첫번째는 음성 데이터의 낮은 품질 및 잡음 문제이다. 음성 데이터는 녹음 환경(실내, 길거리 등)에 따라 예측하기 어려운 잡음(배경 소음, 울림 등)이 포함될 수 있다. 이러한 잡음은 음성 데이터의 특징을 추출하는 데에 있어 정확도를 저하시키므로 데이터의 수집 환경을 보정하거나 잡음을 제거할 수 있는 기술이 필요하다.  두번째는 여러 화자의 혼재 문제이다. 일반적인 대화 환경에서는 여러 화자가 섞여 정해진 순서 없이 화자들이 돌아가며 대화하기 때문에 하나의 음성데이터를 개별 화자로 정확하게 분리하고 식별해야 한다.  세번째는 대화에서 얻을 수 있는 추가적인 정보들이다. 대화의 맥락을 완전히 파악하기 위해서는 누가 말했는지 구분하는 것 외에도 화자의 성별이나 발언에 담긴 실시간 감정 및 어조와 같은 비언어적 단서가 필수적이다. 예를 들어 "(기쁜 어조로) 정말 대단하다!"와 "(화난 어조로) 정말 대단하다." 와 같이 동일한 문장이라도 어조에 따라 의미가 이중적으로 이해될 수 있는 문장을 추가적인 정보 없이 해석하는 것은 어렵다.  **해결 방안**  다음은 문제에 대한 해결 방안들이다.  첫번째로 먼저 원시 음성 데이터로부터 인간의 청각 인지 특성을 모방하여 음성 신호에서 핵심적인 특징을 효율적으로 추출하는 MFCC(Mel-Frequency Cepstral Coefficients**)** 기법을 활용하여 불필요한 잡음을 제거한다. 다음으로 음성 구간만을 정확하게 분리하기 위해 VAD(Voice Activity Detection) 모델을 적용하여 비 음성 구간(잡음, 침묵)을 제거한다.  두번째로 화자 구분을 위해 음성 데이터와 같이 순서 있는 데이터에서특징을 효과적으로 학습하는 TDNN(Time Delay Neural Network) 모델을 사용하여 각 화자의 특징을 압축하고 벡터화 한 x-vector 임베딩을 수행한다. 이를 통해 각 화자들의 음성 간의 유사도를 측정할 수 있게 된다.  세번째로 화자의 성별, 감정 및 어조를 파악하기위해 주파수 패턴을 기반으로 높낮이나 세기 등의 정보를 통해 CNN-LSTM 모델을 학습한다. 이를 통해 단기+장기 모델의 장점으로 맥락을 파악할 수 있다. |

|  |
| --- |
| **5. 본론**  시스템 개요    **음성 전처리 및 특징 추출 (VAD / MFCC):** 잡음과 침묵을 제거(VAD)하고 인간의 청각 특성을 모방한 핵심 음향 특징(MFCC)을 추출하여 후속 작업의 정확도를 위한 기반을 마련한다.  **화자 벡터화 및 그룹화 (TDNN & x-vector):** 다수 화자가 혼재된 대화에서 TDNN 기반의 x-vector 임베딩을 통해 화자 고유의 특징을 추출하고, 이를 기반으로 동일 화자를 그룹화하여 누가 말했는지를 식별한다.  **성별 및 감정 분석 (CNN-LSTM):** 분리 및 그룹화된 개별 화자의 음성에 대해 성별 정보와 감정, 어조를 분석하여 대화의 추가적인 맥락을 추출한다.  **통합 정보 전달 (ASR/TTS API):** ASR(자동 음성 인식) API를 활용하여 대화 내용을 텍스트로 변환하고 성별/감정 등의 추가적인 정보들을 합하여 TTS(Text-to-Speech) 출력을 한다.  **1. 데이터 수집 및 전처리**  AI-Hub 등 공개 데이터셋 플랫폼을 통해 다양한 환경(잡음, 울림)에서의 음성 데이터셋을 확보하고 Silero VAD 등의 모델을 통해 음성 구간만을 분리하여 librosa 라이브러리를 통해 MFCC 특징을 추출한다.  **2. 화자 분리/식별**  데이터셋에 라벨링되어 구분된 화자를 통해 TDNN 기반으로 x-vector 임베딩 모델을 학습시키고 유사도를 계산하여 화자별로 분리한다.  **3. 성별 및 감정/어조 분석**  성별 및 감정 레이블이 포함된 데이터셋을 통해 CNN-LSTM 모델을 설계하고 학습시킨다.  **4. 통합 정보 전달**  [Google Speech-to-Text API](https://search.naver.com/p/crd/rd?m=1&px=444&py=1041&sx=444&sy=459&vw=1438&vh=790&p=jnagRwqo1e8ssNDmkVdssssssjs-106868&q=%EA%B5%AC%EA%B8%80+asr&ie=utf8&rev=1&f=nexearch&w=nexearch&s=HeOIp7AFleKDRPbbOgvW2A%3D%3D&time=1760645983874&abt=%5B%7B%22eid%22%3A%22NEWS-CROP-IMG%22%2C%22vid%22%3A%223%22%7D%2C%7B%22eid%22%3A%22PWL-SAS-RPRV%22%2C%22vid%22%3A%221%22%7D%2C%7B%22eid%22%3A%22NNS-RANKING%22%2C%22vid%22%3A%2254%22%7D%5D&a=web_gen*w.link&i=a00000fa_62ee01df346f14a429754a7e%5ef&r=1&u=https%3A%2F%2Fwww.audiotype.org%2Fspeech-to-text%2Fgoogle%2F&ssc=tab.nx.all&bw=672&bh=830&bx=285&by=60&tid=layout&bid=web%2Fprs_template_v2_web_basic_desk.ts&cr=2) 등을 통해 대화 내용을 텍스트로 변경하고 앞서 얻은 정보들을 포함시켜 TTS와 함께 통합한다. |

|  |
| --- |
| **6. 결론**  본 제안서는 딥러닝 기술을 활용하여 시각/청각 장애인이 겪는 대화의 불평등을 해소하고 다 차원적인 맥락을 인지할 수 있도록 돕는 시스템 구축을 목표로 했다.  이 목표 달성을 위해 크게 네 가지 단계로 잡음 해결, 화자 식별 및 분리, 다차원 맥락, 그리고 정보 통합에 중점을 두었다. 구체적으로는 음성 전처리를 위한 VAD/MFCC, 화자 식별을 위한 TDNN 기반 x-vector 임베딩, 감정 및 어조 분석을 위한 CNN-LSTM 모델을 제안하였다.  **향후 할일**  실시간 처리 환경: 실제 대화 환경에 적용할 수 있도록 플랫폼을 구축하고 음성 데이터의 실시간 최적화를 통해 딜레이를 최소화해야 할 것이다.  사용자 피드백: 시각/청각 장애인을 대상으로 하는 만큼 사용성 및 정보 전달 방식에 대한 피드백을 수집하고 UI등을 개선하여 대화 불평등 해소에 기여할 수 있도록 해야 할 것이다. |

**7. 출처**

[1]입 모양 안 보고 대화... AR이 열어준 청각 장애인의 ‘새로운 세상’.(2022).

<https://imnews.imbc.com/replay/2022/nwdesk/article/6393871_35744.html>

[2]조민수, 권철홍. (2021). x-vector를 이용한 다화자 음성합성 시스템. 문화기술의 융합, 7(4), 675-681.

[3] 서경빈, 이동욱, 김남호, 최광미. (2024-06-19). CNN-GRU모델을 이용한 음성데이터 감정 분석 연구. 한국통신학회 학술대회논문집, 제주.

[4] 다카시마 료이치. (2023). *파이썬으로 배우는 음성인식* (정권우, 역) (pp. 248-250). n.p.: 비제이퍼블릭.