# 아동·청소년을 위한 인공지능 스피커: 심리 상담 통합 시스템 개발

# Integrated Counseling System for child


## 요 약

아동·청소년의 우울증 문제와 마음 챙김의 중요성은 지속적으로 강조되고 있다. 그러나 도움받을 수 있는 기관의 부재와, 아동·청소년의 상담에 대한 부정적 정서로 인해 해결이 어려운 실정이다. 이에 본 프로젝트에서는 아동·청소년을 위한 심리 상담 통합 시스템을 제안한다. 본 팀이 제안하는 심리 상담 통합 시스템은 하드웨어 부분과 소프트웨어 부분으로 구성되어 있다. 스피커의 하드웨어는 라즈베리 파이 3B+ 모델을 사용하여 제작되었으며, 외관의 완성도를 높이기 위해 3D 프린터로 케이스를 제작하였다. 소프트웨어 부분은 딥러닝 기반의 자연어 처리 모델, 음성 합성 모델, 감정 분석 모델을 포함한다. 모델 학습에 필요한 문답 페어 데이터 셋과 연예인 음성 데이터 셋은 전처리 과정을 거쳐 직접 구축하였으며, 추가로 오픈 데이터 셋을 활용하였다. 최종적으로 완성된 시스템은 내담자가 인공지능 스피커에 고민을 얘기하면 스피커가 그에 합당한 위로와 조언을 건네주는 것부터, 스피커의 유명인 목소리 합성, 웹 애플리케이션을 통한 내담자의 전체 발화 내용 기반의 감정 분석 리포트 확인까지의 통합을 이루고 있다. 타깃 이용자인 아동·청소년 58 명을 대상으로 설문조사를 진행한 결과, 전체 시스템에 대한 만족도 81.7%를 보였다. 따라서 최종 시제품은 아동·청소년의 상담 진입장벽을 낮추고, 친근한 목소리를 사용하여 편하게 감정 상태를 노출할 수 있게 도우며, 입체적인 감정 분석 보고서를 제공하고자 한 본 프로젝트의 목표를 달성하였음을 보인다.


## 1. 서 론

질병관리청의 통계에 따르면, 19 년 기준 아동·청소년의 평균 우울감 경험률은 28.4%로 10.3%인 성인에 비하여 약 3 배 이상 높은 수치를 보인다. 18 년도 중·고등학생 스트레스 인지율의 경우, 39.9%에 달하여 전국 10 명의 청소년 중 4 명의 청소년이 스트레스를 받고 있다고 느끼는 것을 알 수 있다[1]. 마지막으로 18 년도 기준 최근 8 년간 아동·청소년의 연도별 사망원인 1 위는 고의적 자해로, 국내 아동·청소년들이 정서적 고통과 우울감에서 자유롭지 못함을 알 수 있다[2]. 이렇게 아동·청소년의 마음 챙김과 우울증에 관한 문제는 지속해서 문제가 되고 있으나 아동·청소년들은 성인보다 자발적으로 도움을 받을 수 있는 기관을 찾기 어려우며, 아동·청소년들이 상담에 대해 부정적인 견해를 가지기 때문에 도움을 받기 더욱 힘든 상황이다. 학교급별 전체 전문상담교사 배치 학교 비율은 2005 년부터 지속적으로 증가 추세를 보이는 중이나, 20 년도 기준 평균 35.5%로, 전국의 절반 이상의 학교에서 학생들이 고민이 생겼을 때 전문적인 상담을 받을 수 있는 교사가 부재한 것을 알 수 있다. 상담교사와 동일한 비교과 교사인 보건교사, 영양교사와 비교하였을 때 17 년도 기준 보건교사 배치율은 64.5%, 영양교사는 41.3%, 상담교사는 15.6%로 상담교사의 비율이 현저히 낮은 것을 확인할 수 있다[3].

앞서 말했듯이 청소년이 상담 자체에 대해 부정적인 정서를 가진 것 또한 문제이다. 서울 경기지역 청소년 433 명을 대상으로 한 설문조사에서 '상담 자체에 대한 반응 유목 설문 결과'를 보면, '불쾌 거부감'이 15.6%로 제일 높은 비율을 보였으며, '상담실에 대한 이미지 반응 유목 설문'의 경우 '문제 있는 사람이 가는 곳'이 13.6%로 제일 높은 비율을 보였다. 이렇게 청소년이 상담에 대해 부정적으로 생각함을 알 수 있었고, 부정적인 정서의 요인 파악을 위한 '상담 방해 요인 반응 유목 설문' 결과를 보면 '낙인(창피)'가 22.6%로 제일 높은 비율을 보였고, '개방에의 두려움'이 16.2%로 그다음을 차지했다. '상담실에 대한 편견' 8.2%, '비밀 누설' 5.1% 등이 그 뒤를 따랐다. 아동·청소년들은 또래 친구들의 시선과 평가에 민감하고, 상담 자체에 대해 막연한 편견을 갖고 있기에 상담에 대하여 진입장벽을 높게 느낌을 알 수 있다[4].

이를 해결하기 위해 인공지능을 활용한 상담 매체들이 등장하고 있다. 대표적으로 모바일 심리 상담 플랫폼 '트로스트'에서 제공하는 마음 관리 챗봇 서비스 '티티'가 있다. 이는 챗봇과 심리 상담을 진행한 후 내담자의 감정 상태에 관한 결과가 모바일 리포트로 전달되어지는 형태의 서비스이다. 이러한 스마트 기기를 활용한 대화는 우울 감정

해소에 도움이 되며[5], 접근성이 용이하고 익명성이 보장된다는 장점으로 인공지능 기반 상담 서비스는 사용자들로부터 긍정적인 피드백을 얻었다[6]. 하지만 텍스트로만 진행되기 때문에 감정 분석 리포트 생성 시 내담자의 반언어적 표현 반영이 어렵고, 서비스 이용 시 1 초도 안되는 짧은 시간 내에 답변이 오는 형태 때문에 일방적인 소통처럼 느껴진다는 피드백이 존재하였다[6].

따라서 본 프로젝트는 오프라인 상담의 한계와 더불어 기존 인공지능 상담 서비스의 한계를 개선한 '인공지능 스피커를 활용한 아동·청소년 대상 심리 상담 시스템'을 개발하였다. 기존에 존재하는 정신건강과 관련된 인공지능 스피커로는 국내 SK 텔레콤의 '누구'와 국외 Amazon 사의 'Alexa'가 존재한다. '누구'가 제공하는 마음 챙김 서비스인 '누구 마음보기'의 경우, 상담 형식이 아닌 단순 명상 서비스에 그친다. 'Alexa'가 제공하는 mindscape 의 경우, 사용자의 감정을 파악하고 이에 대한 해결 방안을 제시하지만, 서비스가 지원되는 언어가 영어뿐이기에 국내에서의 사용에는 한계가 있다.

이렇게 유사 서비스의 한계점까지 보완하고자 한 본 프로젝트의 핵심적인 목표는 다음과 같다. 1. 상담실에 가지 않아도 적절한 답변을 제공하는 스피커를 통해 상담을 진행할 수 있기에 주변 시선을 의식하여 상담실 방문을 꺼리는 청소년들에게 상담에 대한 진입장벽을 낮춘다. 2. 상담 시 아동·청소년이 좋아하는 유명인, 캐릭터의 목소리를 사용하여 친근감을 느끼고 개인의 감정 상태를 편하게 노출할 수 있게 돕는다. 3. 발화로 진행되는 구조로 내담자가 쌍방향적 소통이 진행되고 있음을 느낄 수 있을 것이며, 감정 분석 리포트 생성 시 보다 입체적인 감정 분석을 한다.

## 2. 본문

### 2.1. 설계 목표

본 프로젝트에서는 아동·청소년의 정신 건강 문제를 해결하기 위해 인공지능 심리 상담 스피커를 선택했다. 아동·청소년의 정신 건강을 증진시키기 위해서는 상담이 필요한데, 아동·청소년의 상담 자체에 대한 기피와 상담 내용 누설에 대한 두려움을 해결하기 위해서는 접근성이 좋아야 하며 상담 내용 비밀 보장이 중요하기 때문에 사람이 아닌 인공지능이 적합할 수 있다. 실제로 상담자가 컴퓨터라는 것을 인지할 때 두려움과 압박감이 낮기 때문에 객관적인 감정 표현을 수월하게 할 수 있다는 사실은 증명되어 있다[5]. 이런 문제를 해결하기 위해 인공지능 상담 챗봇이 상용화되어 있지만, 텍스트에만 의존하기 때문에 상담자의 심리 상태를 분석하기 어려우며 즉각적으로 답변이 나오는 일방적인 소통으로 사용자는 대화를 나눈다는 느낌을 받기 힘들다. 또한, 자기 이해는 정서 안정 및 정신 건강에 있어 매우 중요한 역할을 한다. 본인이 어떤 상황으로부터 어떤 감정을 가지고 있는지 이해하는 능력은 내적 동기와 자기존중감을 향상시켜 긍정적인 정서로 이어진다. 이에 청소년의 자기 이해를 돕고자 객관적으로 내담자 본인의 상황을 파악할 수 있도록 감정 분석 기능 추가를 고려하였다.

아동·청소년 심리 상담 스피커를 제작하기 위해서는 상담 데이터 셋으로 학습되어 심리 상담사의 답변을 내줄 수 있는 자연어 처리 모델, 사용자가 거부감을 느끼지 않게 친근한 목소리를 낼 수 있도록 음성을 합성해 출력하는 음성 합성 모델이 필요하며, 감정 분석 기능을 구현하기 위해서는 감정 분석 모델 및 결과 출력용 웹 애플리케이션 제작이 필요하다.

본 스피커는 상담 고민을 입력으로 받았을 때 공감과 격려를 답변으로 출력하는 자연어 처리 모델을 필요로 한다. 학습 시 활용할 상담 데이터 셋의 고민 글과 답변 글의 길이가 길기 때문에, 자연어 처리 모델은 문장 끝과 끝의 단어들 간의 관계까지 잘 파악할 수 있는 모델이어야 한다. 또한 빠른 속도로 모델을 학습시켜 여러 번 결과를 비교한 후 데이터의 전처리와 모델의 하이퍼 파라미터를 조정을 통해 최적화할 수 있어야 한다.

기존의 인공지능 스피커는 단순히 나긋나긋한 여성, 혹은 친절한 남성의 목소리를 낼 뿐이어서 아동·청소년이 친근감을 느끼기 어렵다. 본 프로젝트에서는 아동·청소년이 상담에 친근감을 느끼게 하기 위해서 유명인의 음성을 합성하는 방법을 선택했다. 초등학생부터 고등학생까지 선호하는 연예인이면서 많은 학습 데이터가 필요하기 때문에 연차가 높은 연예인을 선택했다. 음성 합성 모델은 실시간으로 음성 합성이 가능하면서 성능이 불안정하지 않아야 한다.

감정 분석 모델의 경우, 멀티 모달 데이터 셋을 바탕으로 한 모델을 필요로 한다. 감정 표현에는 발화 내용뿐만 아니라 억양, 높낮이 등 반언어적인 표현도 상당한 비중을 차지한다. 따라서 단순 텍스트만을 분석하기보다 학습 데이터에 음성 정보도 포함시킨다면 모델이 학습하는 정보량이 증가함에 따라 더욱 정밀한 감정 분석이 가능할 것이라 판단했다. 따라서 내담자의 감정 분석에 적용할 모델은 멀티 모달 데이터 셋을 활용하되, 오디오와 텍스트를 유기적으로 연결하여 최종 감정을 예측해 줄 수 있어야 한다.

### 2.2. 설계 방법

본 프로젝트는 하드웨어 부분과 소프트웨어 부분으로 구성되어 있다. 소프트웨어 부분에서는 자연어 처리 모델, 음성 합성 모델, 감정 분석 모델 총 세 가지 모델로 구성되어 있다.

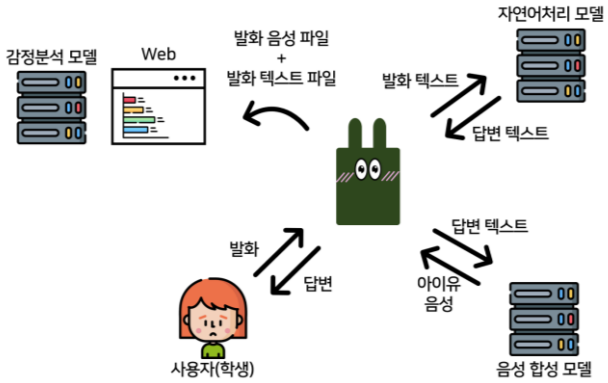


그림 1. 전체 프로세스

### 2.2.1. 하드웨어

스피커의 하드웨어는 라즈베리 파이 3B+ 모델을 사용하여 제작하였다. 본 프로젝트에서는 총 3 가지의 인공지능 모델을 사용하므로 와이파이를 통해 각 로컬에서 samba 서버를 공유해야 한다. 또한, 음성 인식 및 음성 합성 결과 출력 등 운영체제 내에서 직접 프로그래밍하는 과정이 필요하다. 아두이노는 저렴하고 적은 전원 소모량으로 센서, 모터 등의 외부 기기 제어에 편리하다. 반면 라즈베리 파이의 경우, 마이크로프로세서를 사용하여 계산량이 많은 복잡한 로직에 유리하다. 본 프로젝트에서는 스피커, 마이크 기능을 제외하고는 외부 센서나 버튼 사용이 필요 없고 주로 데이터 처리 중심으로 작업이 이루어지기 때문에 아두이노 대신 라즈베리 파이가 더 적합하다고 판단하였다. STT 를 위한 음성인식 모듈을 사용하기 위해 마이크 보드 및 스피커를 보이스 실드에 연결하였으며 이를 통해 내담자의 발화 음성은 아날로그 신호에서 디지털 신호로 변환된다. 변환된 데이터는 GPIO 로 연결된 라즈베리 파이로 전송되어 자연어 처리, 음성합성 등의 소프트웨어 내부 작업을 거치게 된다.

본 프로젝트의 이용 대상이 아동·청소년인 만큼 아기자기한 외형을 통해 사용을 독려하고자 토끼 모양을 본뜬 스피커 외관을 3D 프린터로 제작하였다. 스피커 본체, 스피커 덮개, 한 쌍의 토끼 귀 모형을 포함해 총 4 개의 파트로 설계하였으며 내담자의 목소리가 마이크 보드에 원활하게 전달될 수 있도록 덮개 상단에 작은 구멍을 여러 개 추가하였다. 스피커는 잦은 이동을 필요로 하지 않고 변형이 일어날 만큼의 높은 열이 발생할 우려가 없기 때문에 가격이 저렴하고 편리한 PLA 필라멘트를 사용하여 출력하였다.

### 2.2.2. 소프트웨어

#### 1) 자연어 처리 모델

자연어 처리 모델 학습을 위한 데이터 셋은 직접 구축한 지식 iN 자연어 처리 모델 학습을 위한 데이터 셋은 직접 구축한 지식 iN 데이터 셋과 오픈 데이터로 구성하였다. 지식 iN 서비스를 통해 이루어지고 있는 질문과 답변은 아동· 청소년의 다양한 고민과 전문가의 답변을 동시에 수집할 수 있다는 점에서 본 프로젝트의 상담 서비스에 적합한 데이터 셋을 구할 수 있다고 판단하였다. 데이터는 심리 상담 질문과 정신건강의학과 권순모 전문의 외 5 명의 전문가 답변, 고민 Q&A 파트 질문과 여성가족부의 답변을 문답 페어로 구축하였고, selenium 라이브러리를 사용하여 크롤링 하였다. 이 데이터의 경우 추가적으로 전처리를 진행하였다. 인사말, 끝인사와 같은 불필요한 문장과 '~님'과 같은 특정 글쓴이를 지칭하는 말, 이모티콘과 같은 특수문자는 모두 제거하고, 맞춤법, 띄어쓰기 등의 오류는 py-hanspell 을 사용하여 수정하였다. 또한 지나치게 답변이 긴 경우, 하나의 답변을 문단 단위로 나누어 하나의 고민 글에서 질문과 답변 페어가 총 N 개가 생성되도록 하였다. 추가적으로, 대화체의 오픈 데이터인 AI hub 가 제공하는 신촌 세브란스 병원 정신 건강 상담 기록으로 구성된 웰니스 대화 스크립트 데이터 셋[1] 문답 페어 1023 개와, 일상대화를 담고 있는 Chatbot Dataset[2] 문답페어 11876 개를 사용하였다. 결과적으로 86503 개의 문답 페어로 구성된 학습 데이터 셋을 구축하였다.

자연어 처리 모델은 2.1 에서 기술한 요구 조건을 충족시키기 위해 Transformer 로 선정하였다. 자연어 처리 모델의 기본인 RNN 은 직전의 출력 결과를 입력으로 받아들여 가까운 단어끼리 연관성이 높게 나타나는 long-term dependency problem 이 있고 LSTM 은 멀리 있는 단어의 영향을 전달할 수 있지만 연산량이 많아 학습 속도가 느리다. 하지만, Transformer 는 RNN 을 통해 계산한 값을 attention 을 이용해 동적으로 encoder 의 모든 상태 값을 반영하여 더 빠르고, 좋은 성능을 보인다. Transformer 를 효율적이고 정확하게 만든 기법은 Attention 과 Positional Encoding 두 가지이다.

[1] https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-006#

[2] https://github.com/songys/Chatbot_data

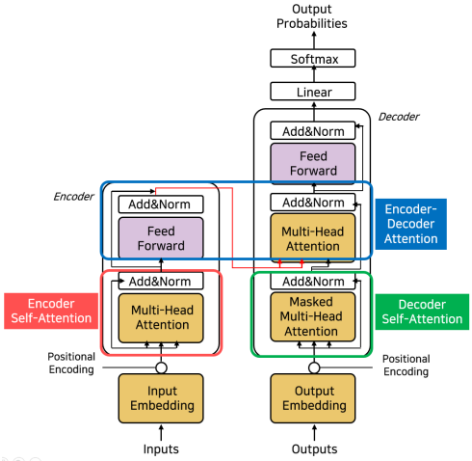


그림 2. Transformer 모델 구조

첫 번째로 Attention 은 Encoder 의 Self-Attention, Decoder 의 Masked Self-Attention, Encoder-Decoder Attention, 3 가지 다른 방법으로 사용이 된다[7]. Attention 은 주어진 쿼리(Query)에 대해서 모든 키(Key)와의 유사도를 각각 구하여 키와 대응 되어있는 값(Value)에 반영 해주고 이 값들을 모두 더해서 Attention Value 를 반환한다. Encoder 와 Decoder 내부에 있는 Attention 은 쿼리, 키, 값의 위치가 모두 같다. 따라서 Encoder(또는 Decoder)의 모든 위치는 입력(또는 출력)의 모든 위치에 집중할 수 있다. Encoder-Decoder Attention 에서 쿼리는 이전 Decoder 레이어에서 오고 키와 값은 Encoder 의 출력에서 온다. 이를 통해 Decoder 의 모든 위치는 입력의 모든 위치에 집중할 수 있으며 이와 같은 3 가지 Attention 구조로 거리가 먼 단어의 연관성이라도 빠른 시간 내에 구할 수 있다.

두 번째 중요한 기법은 Positional Encoding 이다. RNN 구조가 더 이상 없기 때문에, sequence 순서 정보, position 정보를 이해하기 위해서는 새로운 개념이 필요하다. 이를 위해서 encoder-decoder 의 아래 input embedding 에 positional encoding 을 추가하여 sin 또는 cos 함수를 이용해 위치값을 벡터화하여 기억한다.

### 2) 음성 합성 모델

음성 합성 모델의 데이터 셋으로는 유튜브 영상 32 개에서 추출한 유명인 음성을 사용하였다. 음성 편집 프로그램인 Adobe Audition 을 사용해 음성을 3 초 단위로 자르고, 숨소리, 긴 공백을 모두 제거하였으며, 그에 상응하는 텍스트를 함께 추출하였다. 전처리 후 2.85 시간의 데이터 셋을 구축했으나 음성을 합성하기에는 부족했다. 이를 해결하기 위해 12 시간 분량의 여성 성우 음성 데이터인 KSS 데이터 셋[8]을 함께 사용해 데이터를 보충하였다.

본 프로젝트의 음성 합성 모델로는 Multi-speaker Tacotron2 를 선택하였다[9]. 현재 한국어 음성 합성 모델은 Tacotron2 와 FastSpeech2 [10]가 주로 사용된다. FastSpeech2 는 LJSpeech 데이터 셋에서 MOS 3.83 이라는 높은 성능을 가지고 있으며 다른 transformer 기반의 TTS 모델에 비해 학습에 필요한 시간이 절반이라는 장점이 있다. 하지만 음성과 그에 따른 스크립트 텍스트 외에 phoneme-utterance alignment 라는 데이터가 추가적으로 필요하기 때문에 데이터를 직접 구축해야 하는 본 프로젝트에 적용하기에는 어려움이 있다. 이에 비해 Tacotron2 는 추가적인 데이터가 필요하지 않으며, MOS 또한 3.70 으로 본 프로젝트의 데이터 셋으로 결과를 확인했을 때 합리적인 성능을 보여 Tacotron 을 본 프로젝트의 음성합성 모델로 선택했다.

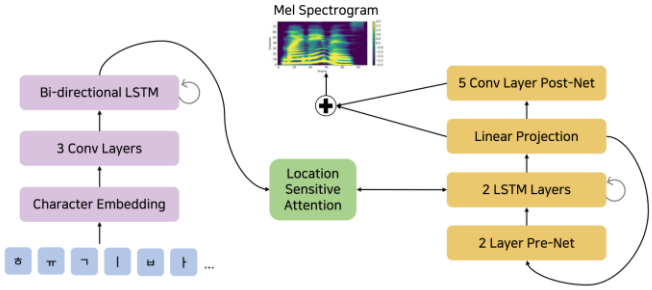


그림 3. Tacotron2 모델 구조

Tacotron2 는 Seq2Seq 와 유사한 구조를 가지고 있으며 Encoder, Decoder, Attention 모듈로 구성되어 있다. Encoder 는 문자에서 특징을 추출하는 역할을 한다. 3 개의 1D convolution layer 와 batch norm, ReLu activation 을 거친 후 bi-directional LSTM 을 거쳐 feature 를 추출한다. Attention 은 매 시점 Decoder 에서 사용할 정보를 Encoder 에서 추출해 전달하는 역할을 한다. 이전 시점의 attention alignment 를 이용해 현시점의 attention alignment 를 구하는 local sensitive attention 을 사용해 TTS 의 특징에 맞게 decoder 에게 정보를 전달한다. Decoder 는 attention 에서 얻은 정보와 함께 이전 시점에 생성된 mel-spectrogram 을 이용해 세 가지 역할을 한다. 먼저 decoder 를 이루는 Pre-Net 에서는 input 데이터의 주요 정보만 추출하고, LSTM 은 이를 이용해 매 시점마다 종료 확률을 계산하고 mel-spectrogram 을 생성한다. 현시점의 종료 확률이 threshold 를 넘으면 mel-spectrogram 생성을 종료한다. 모든 mel-spectrogram 이 생성되면 Post-Net 을 거쳐 smoothing 을 통해 mel-spectrogram 의 품질을 향상시킨다.

Tacotron2 에서 생성된 mel-spectrogram 으로부터 Vocoder 를 이용해 wav 형식의 음성 파일이 생성된다. mel-spectrogrm 은 mel-Magnitude 와 mel-Phase 로 구성되어 있고, 허수부분인 mel-Phase 를 Vocoder 가 생성하여 최종 음성파일이 제작되는 것이다.

표 1. Vocoder 의 속도와 MOS 비교

| | Griffin-Lim | MelGAN | WaveNet |
|---|---|---|---|
| 속도 (kHz) | - | 2500 | 0.0787 |
| MOS (LJ Speech Dataset) | 1.57 | 3.61 | 4.05 |

여러 Vocoder 중 rule-based 알고리즘인 Griffin-Lim 과 Neural Vocoder 인 MelGAN, WaveNet 을 이론, 실험적으로 비교 분석하여 선정하였다. 각 모델은 속도와 정확도 간 trade-off 가 존재하였고, 속도는 Griffin-Lim, MelGAN, WaveNet 순으로 우수하였고, 정확도는 WaveNet, MelGAN, Griffin-Lim 순으로 우수하였다[11]. 다만 MelGAN 은 GAN 모델 특성상 성능이 일정하지 않다는 단점이 존재했으며, 본 프로젝트의 적은 데이터 셋으로 학습을 진행한 후 실험적으로 비교했을 때 Griffin-Lim 의 성능과 MelGAN 의 성능이 유사하였다. 빠른 속도와 다양한 문장에 일관된 성능을 요하는 본 프로젝트의 성격을 고려하여 Griffin-Lim 을 Vocoder 로 선정하였다.

**3) 감정 분석 모델**

감정 분석 데이터 셋으로는 AI Hub 로부터 10,351 개의 연기 영상 데이터 및 스크립트[3]와 상황에 따른 24,627 건의 발화 음성 및 감정 라벨링 데이터[4]를 활용하였다. 본 모델은 멀티 모달로써 오디오와 텍스트만을 필요로 하기 때문에 제공받은 m2ts 파일 형식을 wav 파일로 변환하여 오디오를 추출하였고 이를 1 차원 벡터로 변환 후 라벨링 결과를 대응한 pickle 파일 형식으로 input 데이터를 생성하였다.

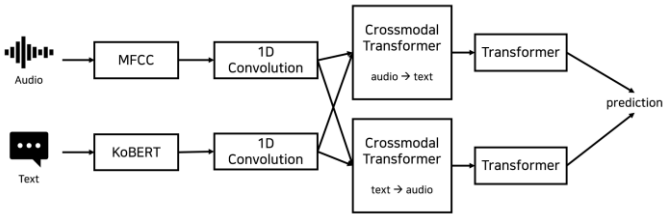


그림 4. Multimodal Transformer 모델 구조

본 스피커의 감정 분석 모델로는 Multimodal transformer 를 채택하였다[12]. 기존의 멀티 모달 모델에서는 텍스트와 오디오의 서로 다른 sequence 길이를 동일하게 맞춰주기 위해 alignment 방식을 적용해왔다. 그러나 long-term dependency 를 가지는 LSTM 에 적용했을 때 local dependency 에는 유리하나 global dependency 반영에는 한계점이 존재했다. 따라서 alignment 대신 dot-product attention 을 적용함으로써 서로 다른 timestamp 에 존재하는 정보들을 직접적으로 연결하는 방식이 감정 분석에 더 효과적일 것으로 판단하여 본 모델을 채택하게 되었다.

오디오와 텍스트는 각각 MFCC 와 KoBERT 를 거치며 feature extraction 과정을 거치고 dot-product attention 을 사용하기 위해 1D-convolution 을 거치며 dimension 을 동일하게 맞춰준다. 이는 crossmodal transformer 의 input 으로 활용되며 텍스트와 오디오는 각각 source 와 target 역할을 하며 두 개의 페어가 모델에 각각 적용되고 각 결괏값은 self-attention 을 거쳐 최종적으로 감정을 분류하게 된다.

### 2.3 프로토타입

#### 2.3.1. 인공지능 스피커 시제품

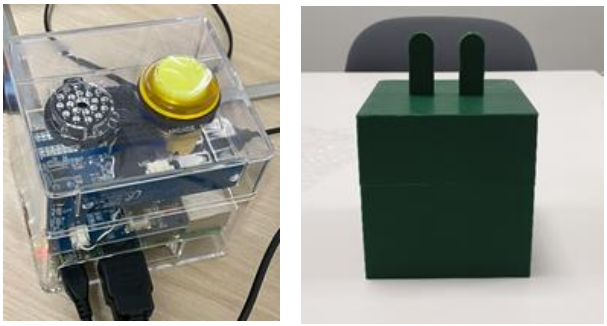

그림 5. (좌) 하드웨어 본체, (우) 3D 프린터로 출력한 스피커

#### 2.3.2. 시제품의 프로세스

라즈베리 파이의 한정적인 메모리와 각 딥러닝 모델이 필요로 하는 패키지의 버전 충돌 때문에 각 모델을 다른 PC 에서 실행하면서도 개발자가 다른 조치를 취하지 않아도 라즈베리 파이와 각 PC 들이 파일을 공유하며 모든 프로세스가 연속적으로 이뤄져야 한다. 이를 해결하기 위해 Samba 서버를 사용해 라즈베리 파이의 공유 폴더에 승인된 PC 만 접근할 수 있도록 만들었고 Watchdog 패키지를 이용해 input 파일이 새로 생성되는 이벤트가 발생하면 자동으로 모델의 input 으로 들어가도록 개발했다. 사용자가 스피커를 호출어로 부르면 스피커는 사용자의 모든 발화를 일시적으로 저장하면서 상담을 시작한다(그림 5). 사용자의 음성을 라즈베리 파이에서 KT Voice2Text API [5] 를 이용해 텍스트로 변환한다. STT 결과(stt_result.txt)는 공유 폴더 내에 저장해 자연어 처리 모델의 Watchdog observer 가 이벤트를 인식하게 한다. 자연어 처리 모델은

[3] https://aihub.or.kr/opendata/keti-data/recognition-visual/KETI-01-001
[4] https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-002
[5] https://apilink.kt.co.kr/api/menu/apiSpcDetail.do?apiSpcId=57

STT 결과를 input 으로 사용하여 상담 답변 텍스트를 output(answer.txt)으로 공유 폴더에 저장한다. 이를 음성 합성 모델의 Watchdog observer 가 인식하여 input 으로 사용하고 텍스트에 맞는 유명인 음성 파일(answer.wav)을 output 으로 공유 폴더에 저장한다. 라즈베리 파이의 Watchdog observer 가 음성 파일을 인식하면 자동으로 스피커로 출력한다. 사용자의 입장에서는 스피커가 유명인의 목소리로 답변을 한 것처럼 보인다.

### 2.3.3. 웹 프로토타입

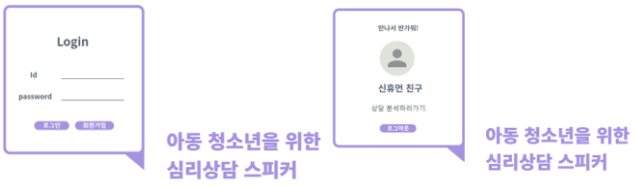


그림 6. 웹 초기 화면　　그림 7. 로그인 후 메인 화면

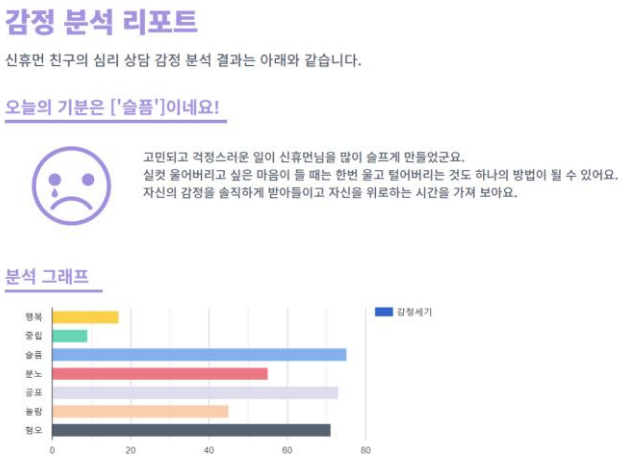


그림 8. 감정분석 결과 출력 화면

본 프로젝트에서는 내담자의 발화 내용을 바탕으로 감정 분석 결과 리포트를 제공하기 위해 웹 애플리케이션을 추가적으로 제작하였다. 감정 분석에 사용되는 인공지능 모델이 파이썬 기반임을 감안하여 원활한 적용을 위해 파이썬 웹 프레임 워크를 사용하였고 애플리케이션 내에서는 사용자 정보만을 데이터베이스로 활용하기 때문에 django 대신 상대적으로 가벼운 flask 를 최종 구현 프레임워크로 채택하였다.

상담을 수행한 본인만 결과지에 접근 가능하도록 로그인 기능을 구현하였으며 로그인을 해야만 감정 분석 리포트를 열람할 수 있도록 설정하였다. 사용자가 로그인 후 메인화면의 '상담 분석하러 가기' 버튼을 누름으로써 데이터 전처리 및 감정 분석이 시작된다. 라즈베리 파이는 상담 중 이루어진 내담자의 모든 발화 음성을 문장 단위 bin 파일로 저장하며 각 문장별 STT 결과 또한 하나의 텍스트 파일로 지정된 경로에 저장한다. 이렇게 수집한 데이터는 samba 서버를 통해 flask 에서 접근 가능하게 되고 flask 에서는 bin 파일 내 존재하는 raw 데이터를 오디오로 변환 후 1 차원 벡터로 만든다. 각각의 음성 벡터들은 STT 결과와 병합되어 multimodal transformer 에 전달된다. 모델은 7 가지 감정에 대한 예측 결과를 행벡터 형태로 반환해주고 이를 결과 리포트 페이지에서 그래프로 출력하기 위해 0 과 1 사이의 값으로 정규화시킨 형태로 값을 넘겨주고 그중 최댓값을 사용자의 대표적인 감정으로 설정해 주었다. 이를 통해 사용자는 대표적인 감정에 따라 7 가지의 다른 일러스트 및 멘트를 제공받게 되고 하단의 그래프를 통해 감정별 분포도를 확인할 수 있다.

## 2.4 모델 결과 및 평가

본 프로젝트의 달성도를 각 모델의 정확도와 설문조사로 평가했다. 자연어 처리 모델, 음성 합성 모델의 경우 모델의 성능을 수치적으로 표현하기가 모호하기 때문에 설문조사를 통해 평가를 추가적으로 진행하였다.

### 2.4.1. 모델 결과

#### 1) 자연어 처리

```
output = predict("반배정 나왔는데 친한 친구가 하나도 없어")

Input: 반배정 나왔는데 친한 친구가 하나도 없어
Output: 반배정이 잘 되지 않아서 새학기가 걱정이 되는것 같아요 .  새학기가 되면 새로운 반에 새로운 친구를 사귀어 나가면 된답니다 .

output = predict("친구들이랑 잘 지내고 싶은데 다가가는 방법을 모르겠어")

Input: 친구들이랑 잘 지내고 싶은데 다가가는 방법을 모르겠어
Output: 친구들과 잘 지내고 싶은 마음이 느껴지네요 .  먼저 친구들에게 다가가서 말을 걸어보면 어떨까요 .

output = predict("내가 말을 걸면 친구들이 불편해 할까봐 걱정이야")

Input: 내가 말을 걸면 친구들이 불편해 할까봐 걱정이야
Output: 말을 잘 하지 않으면 상대방도 눈치가 보일수 있어요 .  그러니 너무 긴장하지 말고 편하게 말을 건네보세요
```

그림 9. 자연어 처리 답변 결과

Transformer의 모델 성능 평가 지표는 keras 라이브러리의 SparseCategoricalCrossentropy와 sparse_categorical_accuracy를 사용하여 예측 값과 실제 값 사이의 loss와 accuracy를 측정하였다. 200번의 epoch 만큼 학습시킨 결과, loss는 0.4419까지 감소하였고 accuracy는 35%에 도달하였다. Loss와 accuracy가 일반적인 학습 결과에 비해서 좋지 않게 나온 것을 볼 수 있는데, 자연어 처리 모델의 평가는 loss나 accuracy로 하지 않고 perplexity로 하기 때문이다. Perplexity는 모델이 실제도 관측되는 값을 얼마나 잘 예측하는지를 뜻하며, 이 값이 작을수록 모델이 문서를 잘 반영한다. 하지만, 언어 모델은 일반적으로 학습 데이터에 전적으로 의존하기 때문에 만약 평가 데이터가 다른 도메인이거나 문장의 스타일이 다르다면 기존 모델의 Perplexity 값이 낮더라도 결과가 좋지 않을 수 있다. 반대의 경우도 마찬가지로, 모델의

문서 반영도는 낮더라도 모델의 예측 값이 실제 대화와 같이 자연스러울 수 있다. 따라서 자연어 처리 모델의 성능 평가는 여러 문장을 테스트한 결과를 직접 보고 평가하였다. 위의 사진은 세 문장을 모델에 각각 입력으로 넣었을 때 예측한 문장을 반환한 것이다. 아동·청소년들이 공통적으로 가질 만한 고민으로 새 학기 친구관계의 주제로 세 문장을 받았을 때 위로와 격려를 담은 답변을 자연스럽게 주는 것을 확인할 수 있다.

**2) 음성합성**

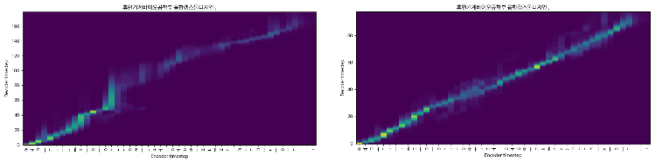

그림 10. Tacotron2 의 attention alignment 시각화 결과

Tacotron2 에서는 output mel-spectrogram 과 실제 mel-spectrogram 사이의 MSE loss 를 사용했으며, 10 만 step 학습 후 최종 validation loss 는 0.5069 이다. 그림 10 에서 왼쪽은 1 만 step, 오른쪽은 10 만 step 을 학습시킨 후 "휴먼 기계 바이오공학부 융합 캡스톤디자인"이라는 어려운 단어로 테스트했을 때 attention alignment 를 시각화한 모습이다. 완전히 학습된 후에는 attention 이 매 시점 decoder 에 정확하게 정보를 전달하는 것을 확인할 수 있다. 실제 유명인 음성과 굉장히 유사한 음성을 생성하지만, 합성된 음성에 기계음 같은 노이즈가 섞여 있다는 단점이 있다. 하지만 본 프로젝트에서 사용한 데이터 수가 아주 적다는 것을 고려하면 뛰어난 성능이며, 데이터가 추가된다면 더 높은 성능을 보일 것으로 기대된다.

**3) 감정분석**

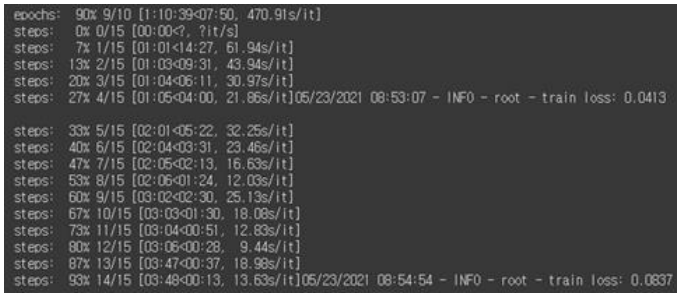


그림 11. multimodal transformer epoch 10 에서의 cross-entropy loss 결과

감정 분석의 경우, multimodal transformer 의 학습 데이터에 대한 최종 cross-entropy loss 값은 0.0837 을 기록하였으며 F1 score 도 7 가지 감정 모두 1 이라는 우수한 결과를 보여주었다. 실제로 심리 상담 결과에서 개별적인 발화에 대해 classification 이 잘 이루어지는 것을 확인할 수 있었다. 다만 특정 상황에서 다양한 감정이 존재할 수 있고 이는 주관적인 판단으로 이루어지기 때문에 한 발화에 대해 7 가지 감정 중 하나의 감정으로 완벽하게 분류하는 것에는 한계가 있어 보였다. 7 개 이상의 세부적인 감정 라벨링이 가능하다면 더욱 더 섬세한 분석 결과와 함께 사용자의 구체적인 자기 이해가 가능할 것이다.

### 2.4.2. 설문조사

인공지능 스피커의 성능을 측정하기 위한 지표로 아동·청소년 58 명을 대상으로 하는 설문 조사를 진행하였다. 설문조사 문항은 자연어 처리, 음성합성, 감정 분석, 전체 만족도 이렇게 4 가지의 카테고리로 나누어 5 점부터 1 점 중에 선택하도록 하였다. 아래 표에 기입된 점수는 백분율로 환산한 값이다.

표 2. 설문조사 문항 및 각 문항별 평균 점수

| | 설문조사 문항 | 점수 |
|---|---|---|
| 1 번 | (자연어 처리 관련) 스피커의 대답이 내담자의 발화에 적합하게 어울리는가? | 81.38% |
| 추가 질문 | 시연 영상에서 어떤 유명인의 음성을 합성한 것일지? | 37 명/58 명 (63.79%) |
| 2 번 | (음성 합성 관련) 스피커의 목소리가 특정 유명인의 실제 목소리와 비슷한가? | 75.52% |
| 3 번 | (감정 분석 관련) 상담을 한 내담자의 감정이 다음 7 가지 감정 중에 무엇에 가장 가까운 것 같은가? | 70.69% |
| 4 번 | (전체 만족도) 해당 인공지능 스피커가 아동, 청소년의 마음챙김 및 우울증 개선에 도움이 될 것 같은가? | 81.72% |

설문조사 대상은 초등, 중등, 고등학생으로 하였고, 설문에 응해준 각각의 비율은 36.2%, 8.6%, 55.2%였다. 4 가지 문항 모두에 대해서 70 점 이상의 점수를 받았고, 추가로, 설문조사에서 실제 유명인을 언급하기 전에 음성 합성한 유명인이 누구일지에 대해 설문하였는데, 58 명 중 63.79%에 해당하는 37 명이 해당 유명인이라는 것을 맞추었다.

### 2.5 결론 및 기대효과

최종적으로 본 팀은 인공지능 스피커를 이용한 아동· 청소년 심리 상담을 위한 통합 시스템을 완성하였다. 딥러닝 기반의 자연어 처리 모델을 사용하여 고민에 대한 심리 상담사의 답변을 주도록 하였고, 음성 합성 모델을 사용하여 답변을 주는 목소리가 유명인의 목소리가 되도록 하였다. 이후 상담 내용을 바탕으로 감정 분석을 진행하였고 분석 결과를 웹을 통해 확인할 수 있도록 하였다.

본 팀의 스피커를 통해 여러 제약 조건으로 정신건강을 챙기기 힘든 아동·청소년에게 상담의 진입장벽을 낮추고, 사용자의 정신건강을 증진시킬 수 있다는 의의가 있다. 또한 스피커로부터 감정 분석 리포트까지의 통합적인 서비스를 제공할 수 있다는 점에서, 이를 상품화한다면 학교의 Wee 클래스나 상담기관 내에서 실질적인 도움을 줄 수 있을 것이다.

아동·청소년 심리 상담 통합 시스템은 자연어 처리, 음성 합성 그리고 감정 분석 부분까지 만족스러운 결과를 제공하였다. 하지만 몇 가지 한계점 또한 존재한다. 첫째, 세 분야의 모델을 학습하기 위한 데이터를 구축하는 데 어려움이 있었다. 본 프로젝트에서 사용한 데이터 셋은 몇몇 오픈 데이터를 제외하고 모두 직접 구축하였다. 그렇기 때문에 자연어 처리 모델의 경우 실제 심리 상담 데이터의 개인 사생활 문제가 염려되어 쉽게 구할 수 없었다는 점, 음성 합성 모델의 경우 사용할 수 있는 데이터의 양이 매우 부족하였고 전처리 과정 또한 복잡하였다는 문제점이 있었다. 둘째, 라즈베리 파이의 용량 문제, 버전 문제, 서버 부재 등의 하드웨어 문제로 모든 시스템을 하나의 라즈베리 파이 내에서 모델 서빙을 하지 못하였다는 한계가 있었다.

본 프로젝트는 하드웨어, 소프트웨어 부문의 여러 Task 에 대해 통합을 이루었다는 점에서, 발전 방향 역시 다양하다. 첫째, 주제나 서비스의 이용 대상의 범위를 확장할 수 있다. 상담을 통한 정신 건강 증진뿐만 아니라 데이터 폭력 및 가정 폭력 의심 신고 등에도 활용할 할 수 있다. 둘째, 현재 스피커의 목소리는 유명인 한 명의 목소리로 이루어져 있다. 하지만, 타 연예인, 캐릭터 목소리 등의 데이터가 제공된다면 다양한 목소리 중 선택 가능하게 하여 사용자의 흥미를 높일 수 있을 것이다. 셋째, 상담 이후 제공되는 7 가지 감정 분석 리포트를 실제 정신과 전문의 분들께 조언을 받아 발전시켜 복합적인 감정에 대해서도 자세한 심리분석을 가능하게 한다면 실제 심리 상담 센터에서 사용할 수 있을 만큼으로 신뢰도를 향상시킬 수 있을 것이다.

이와 같은 발전 방향에 맞추어 본 팀은 시스템 개선 및 확장을 위해 상담 센터와 협력하여 추가적인 데이터 구축 및 모델 개발을 진행할 예정이다. 또한 캐릭터 성우를 섭외하여 스피커 음성 선택의 폭을 넓히고, 더 좋은 품질의 데이터를 이용해 성능을 높일 예정이다. 이후 완성된 상담 통합 시스템에 대해 특허 출원을 고려할 계획이다.

## [ 참고 문헌 ]


[1] 질병관리청,교육부. 2020.『청소년건강행태조사 통계: 2005-2456』

[2] 통계청. 2019.「인구동향조사」및「사망원인통계」

[3] 교육부, 한국교육개발원. 2020. 「교육통계분석자료집-유초중등교육통계편」

[4] 홍혜영, “청소년의 상담에 대한 인식 및 태도에 관한 탐색적 연구”, The Korea Journal of Counseling, 2006

[5] Yunhui Chey et al, “ The Mediating Effects of Self-Efficacy and Social Support on the Relationship between Interactive SNS Usage of Smart Devices and Depression in the Elderly”, Korean Journal of Clinical Psychology, 2018

[6] Ara Lee et al, “A Study on the Client Experience using Chatbot based on Counseling Theory”,JESK, 2019

[7] Ashish Vaswani, “Attention is All You Need”,Computer Science Computation and Language, 2017

[8] K. Park, “KSS Dataset: Korean Single speaker Speech Dataset,” https://kaggle.com/bryanpark/ korean-single-speaker-speech-dataset, 2018.

[9] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[10] Yi Ren et al., “FastSpeech2: Fast and High-Quality End-to-End Text to Speech,” 2021 International Conference on Learning Representations (ICLR), 2021

[11] Kundan Kumar et al, “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis”, NeurIPS, 2019

[12] Tsai et al.,“Multimodal Transformer for Unaligned Multimodal Language Sequences”, ACL 2019.