기술 동향 보고서

음성 인터페이스를 활용한 감성 인식과 감성 음성 합성에 대한 기술 동향

Technological Trends on Emotion Recognition and Emotional Speech Synthesis Using Speech Interface

한국방송통신대학교 대학원

정 보 과 학 과

우 은 주

2019년

- 목 차 -

- 그 림 목 차 -

- 표 목 차 -

제1장 서론

1.1 연구의 배경

사회가 발전하고 고도화됨에 따라 사회생활을 하면서 맺는 인간관계 속에서 정신적, 심리적 불안감을 느끼는 사람이 많아지고 있다. 경제발전 및 의료기술의 진전으로 고령 인구의 증가는 노인을 위한 의료 및 사회 복지 서비스라는 사회적 비용의 증가를 초래한다. 또한, 경쟁이 만연한 사회에서 생존을 위한 경쟁 압박은 스트레스, 우울증, 감정표현 불능, 자살 등 다양한 형태로 정서적, 심리적 문제를 발생시키므로 정서적·심리적 불안을 해소할 필요성이 크다.

이러한 사회에서 개인은 자기 자신의 감정을 표현, 감지, 인식하지 못하는 특징이 있어, 자신이 화가 나거나 불안해하는 감정이 정확히 어떤 감정인지 구별하지 못하는 경향이 커진다. 사회 전체적으로 정서적으로 불안, 우울 등의 심리적 상태로 고통받는 개개인의 정신 건강 개선을 위해 감성 치유 기술에 대한 수요는 확대될 전망이다.

생체 신호 인지 기술 및 기계학습, 딥러닝 등의 인공지능 기술의 발전으로 오늘날은 인간의 감정 상태를 파악하는 기술개발이 점차 일상생활에서 적용 가능한 방향으로 연구된다. 그 예로 얼굴의 표정이나 몸짓 등을 파악할 수 있는 영상 인식, 목소리의 강약과 크기 등을 감지할 수 있는 음성 인식, 맥박, 심전도 등의 변화를 감지하는 생체정보 인식 기술을 통해 인간의 내면에 있는 감정 상태 파악이 가능해진 것이다.

이와 같은 과학 기술의 발전으로 사용자의 감정을 읽어내고 이 정보에 기반한 개인화 서비스를 제공하는 기술은 미래의 기술과 산업의 새로운 돌파구를 마련하는 데에 큰 역할을 할 수 있다. 이 분야의 큰 잠재력과 시장성으로 인해 인공지능 및 음성 공학 기술이 발달한 국가들은 감정 인식 분야에 많은 관심을 가지고 지속적인 인공지능 기술개발에 있어서 필수적인 연구주제로 인적 · 물적 자원을 투자하고 있다[1].

인간의 감정 상태에 대한 데이터의 증가와 이러한 빅데이터와 인공지능 알고리즘을 활용한 새로운 비즈니스 솔루션 창출 기회의 가능성이 증대되면서 인간의 감성을 반영한 새로운 비즈니스 시장 창출의 기회가 많아짐에 따라 의료, 자동차, 교육, 판매, 광고 등의 다

양한 산업과의 융합을 통하여 일상생활에서 늘 사용하는 제품 및 서비스에 인간의 감성을 적용할 수 있다. 일례로 평상시에 소셜네트워크상의 사용 언어, 음성의 미묘한 변화에 대한 감성파악을 통해 우울증을 진단하고 자살을 사전에 방지하는 등 다양한 서비스 개발이 가능할 것이다.

1.2 연구의 목적 및 방향

정보 · 통신기술이 발달한 현대를 살아가는 많은 사람은 정서적, 심리적인 문제를 가지고 있고 이에 대해 사람의 감정을 읽고 이에 대응하는 기술은 육체적, 정신적 상처를 극복하며 살아가야 하는 현대인들에게 매우 중요한 문제이다. 따라서 본 과제는 감성 기반 연구라는 큰 틀에서 음성 인터페이스가 어떻게 사용되고 있는지를 중심으로 진행하였다.

음성 인터페이스는 직관적이고 편리해 키보드나 마우스 같은 별도의 조작이 필요한 장치 없이도 누구나 사용할 수 있는 인터페이스이지만 음성 인식 및 음성 합성은 일반 사람들이 쉽게 이해하기에는 그 기술의 진입 장벽이 높고 그와 관련한 일반적인 교양서도 거의 출판되지 않은 실정이다. 음성은 인간-기계 상호작용에 있어 강력한 도구가 될 수 있으므로 음성 인터페이스에 관한 기술 동향을 감성 인식과 감정이 녹아 있는 음성 합성을 주제로 각각 2장과 3장으로 나누어 정리하였다. 부록으로 감성 음성 합성에 사용되는 W3C 표준 규약인 SSML의 응용 사례를 도입하여 서비스하고 있는 기업별로 정리하였고 SSML의 사용법을 보기 쉽도록 간략하게 서술하였다. 감성 음성기술의 응용 분야로 음성 인터페이스의 로봇공학 분야의 응용 사례를 4장에 정리하였다. 본 보고서에 기재한 난해한 기술 용어는 각주로 정리하였다.

1.3 음성 공학 용어정리

본 보고서에는 통상적으로 자주 쓰이지 않는 음성 공학 분야의 용어가 많이 나오므로 아래의 <표 1-1>에 자주 사용된 용어를 한글과 영어 모두 함께 기재하여 정리하였다.

<표 1-1> 용어

한글 용어	영문 용어	의미
음성	speech	사람의 발음 기관을 통해 내는 구체적이고 물리적인 소리. 발화자와 발화시에 따라 다르게 나는 소리로서 자음과 모음으로 나뉘는 성질이 있다. 유의어) 말소리
음성 인식	speech recognition	사람이 발성한 음성의 의미 내용을 컴퓨터 따위를 사용하여 자동으로 인식하는 것 사람이 말하는 음성 언어를 컴퓨터가 해석해 그 내용을 문자 데이터로 전환하는 처리를 말한다. STT(=Speech To Text)라고도 한다. 키보드 대신 문자를 입력하는 방식으로 주목을 받고 있다. 로봇, 텔레매틱스 등 음성으로 기기제어, 정보검색이 필요한 경우에 응용된다.
음성 합성	speech synthesis	컴퓨터를 이용하여 사람의 말소리를 기계적으로 합성하는 일. 음성 인식과 함께 번역 기계, 로봇 제조 기술 따위에 쓴다. 모델로 선정된 한 사람의 말소리를 녹음하여 일정한 음성 단위로 분할한 다음, 부호를 붙여 합성기에 입력하였다가 지시에 따라 필요한 음성 단위만을 다시 합쳐 말소리를 인위로 만들어내는 기술이다. TTS(=Text-to-Speech)라고도 한다.
음성 변환	voice conversion	음성 인식, 음성 합성에서 어떤 사람의 음성을 다른 사람의 음성으로 변환하는 것 변환할 화자의 음성 특징을 추출한 후, 그 특징을 기초로 해서 음성 스펙트럼이나 운율의 특징을 수정하는 방법을 이용한다.
원시화자	source speaker	음성변환기술에서 음성을 변환시키고자 하는 원래 발화했던 화자
목표화자	target speaker	음성변환기술에서 변환하고픈 음성을 발화한 화자
오디오 스타일변환	audio style transfer	영상처리에서의 화풍변환기법을 오디오에 적용한 것. style로 지정한 오디오의 특성을 원시 content에 입혀서 하나의 오디오 파일에 두 개의 특성이 모두 나타나도록 하는 것
병렬 말뭉치	parallel corpus	연구 등의 목적으로 구축한 음성 데이터 중 여러 명의 화자가 동일한 대본 내용을 낭독하는 코퍼스
비병렬 말뭉치	non-parallel corpus	연구 등의 목적으로 구축한 음성 데이터 중 여러 명의 화자가 다른 대본 내용을 낭독하는 코퍼스

보코더	vocoder	인간의 목소리를 분석・합성하여 소리를 재생산하는 프로그램. 또는 그러한 기계 보코더는 오디오 신호의 특성 요소를 캡처한 다음 이러한 특성 신호를 사용하여 다른 오디오 신호에 영향을 주는 오디오 프로세서이다. 보코더 효과의 기반 기술은 처음에 음성을 합성하기 위해 사용되었다.
보코더를 사용하지 않은	vocoder-free	보코더를 사용하지 않고 음성 변환을 수행한
스펙트로그램	spectrogram	음파 분석기에 의한 음파의 스펙트럼을 사진으로 찍은 것
기본 주파수	fundamental frequency	음원인 성대의 진동수를 의미하고 음향적으로는 F_0 값으로 측정됨
피치	pitch	음의 높낮이를 가리키는 음악용어이나 음성에서는 음성의 높낮이를 의미 음악에서는 보통 A4, B3과 같은 표기법을 사용하거나 Hz로 표기한다. 악보에서는 음높이를 줄과 칸으로 표현한다. 음의 진동수가 클수록 음높이는 높아지며, 진동수가 2배가 되면 한 옥타브 높은음이 된다.

제2장 감성 음성 인식 기술

2.1 감성 음성 인식의 개요

인간의 표면적인 감정은 얼굴의 표정이나 행동, 그리고 음성 인식으로 어느 정도 인식이 가능하다. 하지만, 내면적인 감성 인식은 생체 신호를 이용하면 인식이 가능할 수는 있지만 쉽지 않은 작업이다. 따라서 단일 신호에만 의존하지 않고, 영상, 음성, 생체 신호 등 멀티모달 신호를 사용하는 것이 감성 인식의 정확도를 높이는데 이바지할 수 있으므로 멀티모달 신호를 이용한 딥러닝 알고리즘을 효과적으로 설계하고 학습시킬 수 있다면 소셜 로봇 같은 기계와 인간의 상호작용이 좀 더 원활해질 수 있을 것이다[1]. <그림 2-1>은 감성 인식 및 분석 기술의 대표적인 유형인 영상, 음성, 생체정보 인식을 보여준다.

표정·이미지 음성 생체정보

_출처: Microsoft Azure 홈페이지, 전파신문(2016.11.5.), Lukasz Piwek 외(2016) 참조

<그림 2-1> 감성 인식 및 분석 기술의 유형

이 중에서도 음성에 녹아든 감성 인식은 음성의 강약, 높이, 속도, 크기 등의 특징을 파악 · 인식하고 이를 분석하여 감정변화의 상태를 추정하는 기술로 사용자의 음성을 활용하여 감성을 인지하는 기술은 음성 자체를 인식하여 텍스트로 변환하는 기술단계에서 점차 발전되었고, 현재에는 사용자의 감정 상태를 인식하고자 하는 연구를 많이 시도하는 추세이다.

현재 단어는 무시하고 오직 운율적 정보에 초점을 맞추어 감성을 인식하는 기술이 자동차 제조업에서도 활발히 연구가 진행 중이다. 운전자는 음성으로 CD 재생기나 히터, 환풍기를 작동시킬 수 있는데, 감성 인식 기술을 통해 음성명령에서 운전자의 감정이 화가 나 있는지, 졸린 상태인지, 아니면 실망한 상태인지를 인지하여 운전자의 기분에 따라 정보 제공 방식을 바꾸는 것이다. 예를 들어 운전자가 지루하거나 졸린 상태에서는 자동차가 볼륨을 높여 CD 재생기를 작동시키게 되고, 운전자가 화가 나 있거나 흥분 상태에 있다면 보다 침착한 방식으로 정보를 제공하게 된다.

다른 예로는 통화하는 상대방의 속마음 또는 감정을 예측 분석하는 감정 분석 서비스가 국내 한 통신사에 의해 개발 중이다. 이 서비스는 기쁨, 슬픔, 짜증, 분노 등 18가지 감정에 대해 분류가 가능하며 현재 인식률은 71.7% 정도이다. 이외에도 마스터 액션과 음성 파형을 특징으로 추출하여 사람의 감정을 인식하는 캐릭터가 일본에서 최근 개발되어 화제가 되었다. 이러한 시도는 개인의 감정을 학습하는 기술이 얼마나 잠재적인 시장성이 높은지를 보여주는 중요한 단서이다[2].

성대의 길이와 특성과 같은 물리적인 개인차와 말하는 습성 등으로부터 나타나는 개인별 음성의 특징에 대한 데이터 구축 및 분석을 통한 감성 인식이 중요하므로 이에 음성신호 중 인공지능 알고리즘을 활용하여 감성 인식에 효율적인 특징을 선별하여 음성패턴을 분석하여 화자의 감성을 인식 및 분류하는 연구가 활발하게 진행되고 있다. 이를 위해 분석 기법으로 RNN(Recurrent Neural Network) 등의 딥러닝 기법을 사용하고 있다.

한동안 음성기반 감성 인식 연구는 작은 시간 영역 단위의 실시간 감성 인식에 집중되어 있었다[3]. 실시간 감성 인식 연구는 크게 두 가지로서 새로운 특징을 추출하거나 분류 방법론을 다르게 해 정확도를 개선하는 연구들로 나눌 수 있다.

특징 추출 연구로는 개인마다 발성의 특징이 다르므로 그러한 특징을 찾는 것을 목표로 한다. 예를 들면, 구간이 아닌 순간의 감정을 인식하는 기술이 있다[4]. 분류 방법론 중 하나인 계층적 분류 방법론은 분류기들을 여러 개 사용하여 음성에서 비슷한 감정의 인자를 나누어 분류하는 방식으로[5] 비교적 정확도는 높지만 짧은 음성만 인지가 가능한 단점이 있다. 한편 남성과 여성의 학습 모델을 각각 생성한 후 입력 음성을 성별로 구별한 후 성별에 맞는 학습 모델과 비교하는 방식으로 인지하는 방법도 있다[6].

최근에는 딥러닝을 이용한 음성기반 감성 인식 기술들이 주류이다. 이지원 등은 일반화 오류를 보완할 수 있는 다중 작업 기반 합성곱 신경망을 이용한 음성 감성 인식 시스템을

제안하였다[7]. 즉, 이 연구는 감정 분류만을 수행하던 기존 신경망의 기능을 확장해 성별, 감정 활성도, 긍정도 정보를 활용한 다중 작업 기반 신경망 학습을 통해 감성 인식의 성능을 높이고자 하였다.

한편, 강소연 등은 베이지안 로지스틱 회귀(Bayesian logistic regression)를 랜덤 포레스트(random forest)로 대체하여 종래 베이지안 기법들보다 정량적인 감성 인식 성능을 향상시킨 연구를 수행하였다[8].

2.2 감성음성 데이터베이스

감성음성 인식기의 평가에서 고려해야 할 중요한 문제는 성능 평가에 사용된 데이터베이스이다. 저품질 데이터베이스를 사용하는 경우 감성음성 합성 결과가 좋지 않을 수 있고 데이터베이스 설계는 감성음성 분류 작업에 매우 중요하다. 어떤 감성음성 데이터베이스에서 분류 작업은 발화 문장의 강세를 감지하는 것일 수도 있고, 그 감정이 유아에서부터 성인에 이르기까지 기준이 매우 다양할 수 있다. 감성 분류 작업은 데이터베이스에 포함된 감정의 수와 유형에 따라 정의할 수 있다.

독일, 미국 등 감성 인식 연구를 지속해서 수행한 국가의 연구자들은 실시간 음성기반 감성 인식에서 가장 기초가 되는 감정유발 시나리오에 근거하여 감정 기반 음성 데이터베이스를 구축해 왔고, 전 세계적으로 동 분야를 연구하는 연구자들의 기술개발에 활용되고 있다 [9-12].

<표 2-1> 감성음성 데이터베이스

데이터베이스	국가	언어	감정범주
Berlin Emoitonal Database(EMO-DB)	독일	독일어	Anger, Happiness, Sadness, Fear, Disgust, Boredom, Neutral
Danish emotional Database	덴마크	덴마크어	Anger, Joy, Sadness, surprise, neutral
IITKGP: SEHSC	인도	힌두어	Anger, Happiness, Sadness, Fear, Disgust, Boredom, Neutral, Sarcastic, Surprise
IEMOCAP	미국	영어	Happiness, Anger, Sadness, Neutral, Disgust, Fear, Excitement, Surprise

<표 2-1>은 대표적인 감성음성 데이터베이스를 보여준다. Berlin Emotional Speech DB[9]는 독일 베를린 공과대학 연구팀에서 훈련된 연기자로부터 구축한 7가지의 감정에 대한 총 800개의 문장으로 구성된 독일어 데이터베이스이다.

Danish emotional Database[10]는 덴마크 커뮤니케이션 센터에서 구축한 감정 데이터베이스로, 훈련하지 않은 사람들의 음성을 대상으로 구축하였다.

인도의 Vel 공과대학 연구팀의 IITKGP: SEHSC[11]는 배우의 발화를 이용한 데이터베이스(훈련된 연기자)와 실제 상황에 가까운 인공적인 상황(콜센터, 환자와 의사와의 대화)을 녹음한 데이터베이스로 8개 감정에 대하여 12,000개의 문장으로 구성되어 있다.

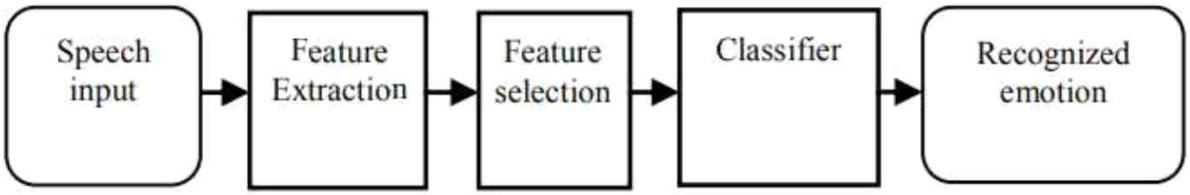
미국의 USC 공과대학 연구팀은 10명의 훈련된 연기자를 대상으로 시나리오 기반, 상황 기반의 대화를 통하여 8개 감정으로 구성된 감성음성 데이터베이스 IEMOCAP[12]를 구축하였다.

이렇게 수집된 데이터베이스를 활용하여 기계학습, 특히, 딥러닝 기반의 알고리즘을 활용하여 감정을 인식하는 연구가 지속해서 이루어지고 있다. 독일 에를랑겐대 연구팀은 Berlin Emotional Speech DB를 이용하여, 운율, 묵음길이, 되풀이, 수정정보 분석을 통한 7가지의 감정(Happiness, Sadness, Anger, Boredom, Disgust, Fear, Neutral)을 인식하는 기술을 연구하였으며 평균 80%의 인식 성공률을 보였다. 미국 USC대 연구팀은 Berlin Emotional Speech DB와 일반인 문장녹음 데이터베이스를 이용하여, 운율, 어휘, 화법 분석을 통한 4가지 감정(Happiness, Sadness, Anger, Neutral) 인식 기술을 연구하였으며 평균 80%의 인식 성공률을 보였다.

2.3 감성음성 인식의 특징

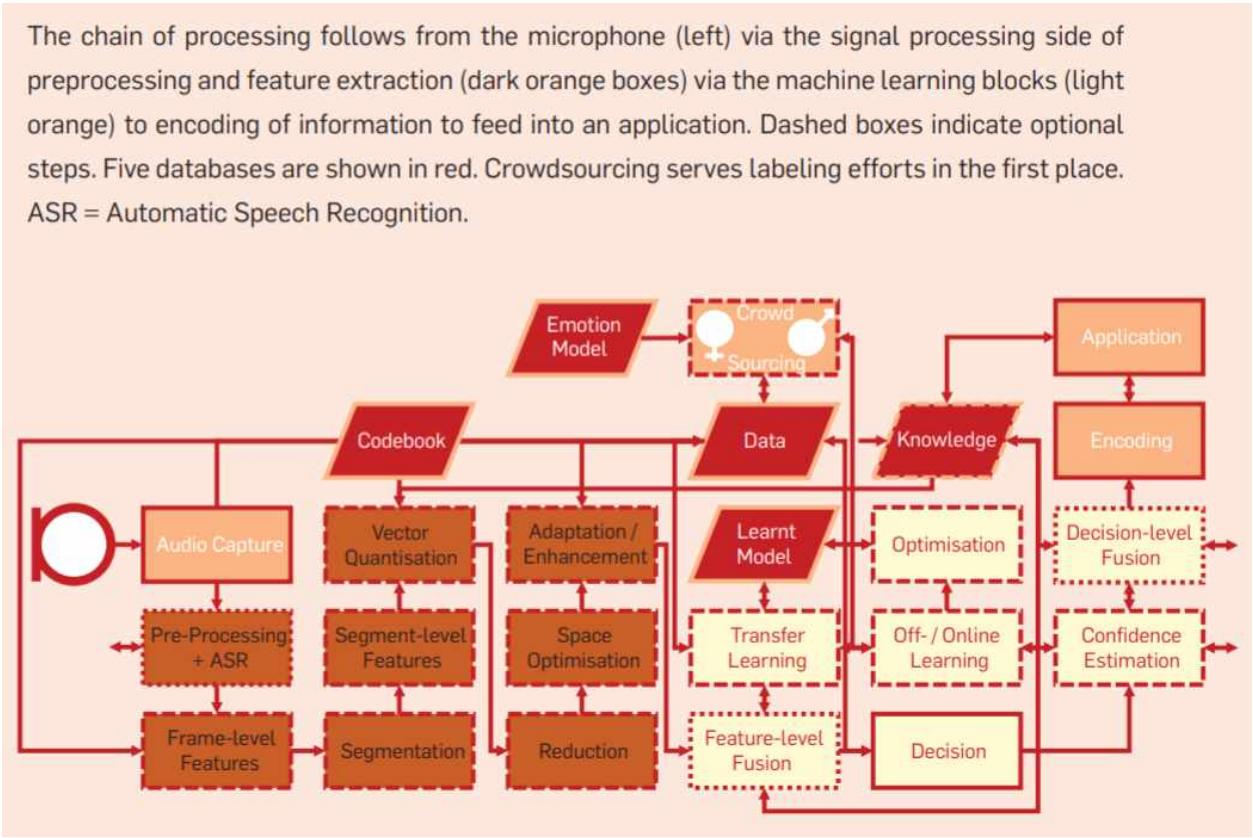
2.3.1 감성음성 인식의 개요

음성 신호 기반 감성 인식시스템은 전반적으로 <그림 2-2>와 같이 Speech input, Feature extraction, Feature selection, Classification, Recognized emotional output의 총 5가지 주요 모듈로 구성이 된다[13].



<그림 2-2> 음성 신호 기반 감성 인식시스템의 구조도

현재 기술 수준에서 구현 가능한 감성음성 인식 엔진의 전체적인 구조는 <그림 2-3>과 같이 개략적으로 나타낼 수 있다[14].



<그림 2-3> 감성음성 인식 엔진의 구조도

감성음성 인식에서 인식률을 평가할 때, 가장 중요한 요소는 입력값으로 들어가는 음성 신호이다. 감성이 들어간 음성 신호로 만들어진 데이터베이스의 음성 신호들이 얼마나 자연스러운지가 전체 엔진의 인식 성능에 큰 영향을 준다.

2.3.2 구간 특징과 전역적 특징

감성 인식시스템의 설계에서 중요한 문제는 다른 감정을 효율적으로 특성화하는 적절한 기능을 추출하는 것이다. 패턴 인식 기술은 감성 음성 인식 영역에 의존적이기 때문에 적절한 기능 선택은 분류 성능에 큰 영향을 미칠 수 있다.

특징 추출에서 4가지 문제를 고려해야 한다. 첫째 문제는 특징 추출에 사용되는 분석 영역이다. 일부 연구자들은 음성 신호를 프레임이라고 하는 작은 간격으로 구간별로 특징 벡터를 추출하는 일반적인 기법을 따르지만, 다른 연구자들은 전체 음성 데이터에서 전체 영역의 특징을 추출하는 것을 선호하기도 한다. 또 다른 중요한 문제는 특징 추출에 가장 적합한 유형이 무엇인지에 대해서다. 여기서 말하는 유형이란 예를 들면 음성의 피치(pitch), 에너지(energy), 제로 크로싱(zero crossing) 등이다. 세 번째 문제는 분류기의 전체 성능에 대한 사후 필터링 및 무음 제거와 같은 일반적인 음성 처리의 효과가 무엇인지에 대한 것이다. 마지막 문제는 감정을 모델링 하기 위해 음향 특징을 사용할 수 있는지 또는 언어, 담화 정보 또는 얼굴 특징과 같은 다른 유형의 특징과 결합할 필요가 있는지다.

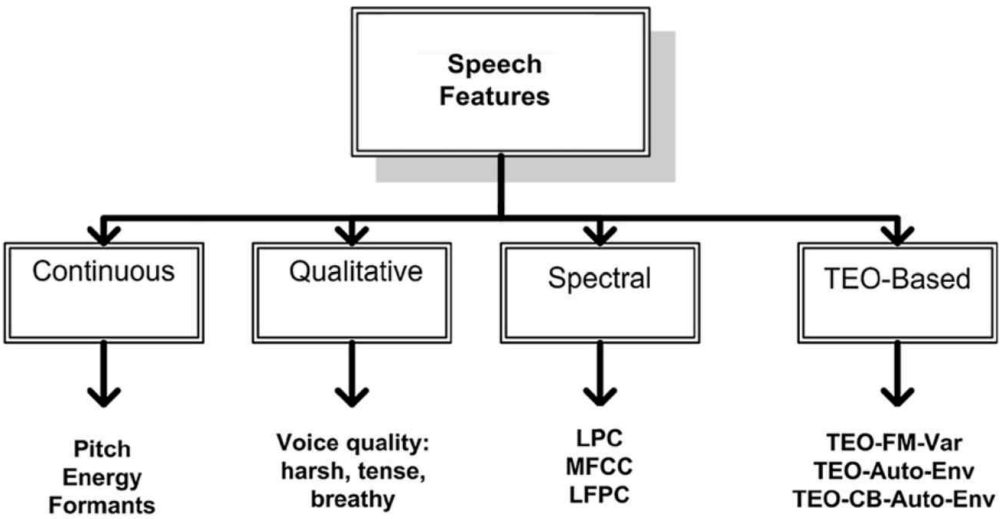
음성 신호는 고정된 것이 아니므로 일반적으로 음성은 음성 신호를 작은 것으로 나누는 프레임이라는 음성 단편들을 이용하여 처리한다. 각 프레임 내에서 신호는 거의 정지된 것으로 간주하며 피치 및 에너지와 같은 운율적인 음성 특징(prosodic speech feature)은 각 프레임에서 추출되어 구간 특징 (local feature) 이라고 한다. 반면에 전역적 특징은 발화에서 추출된 모든 음성 특징의 통계로 계산된다. 감성음성 인식에 더 적합한 구간 및 전역적 특징에 대해 의견이 분분했다.

대다수 연구자는 분류 정확도 및 분류 시간 측면에서 전역적 특징이 구간 특징보다 우수하다는 데 동의했다[15-16]. 전역적 특징은 구간 특징보다 그 종류가 훨씬 적다. 따라서 전역적 특징에 교차 검증 및 특징 선택 알고리즘을 적용하면 구간 특징에 적용할 때보다 훨씬 빠르게 수행된다. 그러나 연구자들은 전역적 특징이 흥분이 큰 감정, 예로 분노, 두려움 및 기쁨과 같은 감정과 대비하여 흥분이 낮은 감정, 즉 슬픔 등과 같은 감정을 구분할 때만 효율적이라고 주장하기도 한다. 그들은 전역적인 특징으로 비슷하게 흥분을 일으키는 감정, 즉 분노와 기쁨을 분류하는 것이 불가능하다고 주장한다. 전역적 특징의 또 다른 단점은 시간 정보 신호가 완전히 손실된다는 점이다. 더욱이, 훈련 벡터의 수가 모델 파라미터를 신뢰성 있게 추정하기에 충분하지 않을 수 있어서, 음성의 전역적 특징을 갖는 히든

마르코프 모델(HMM) 및 서포트 벡터 머신(SVM) 과 같은 복잡한 분류기를 사용하는 것은 신뢰할 수 없을 수 있다는 것도 단점으로 꼽는다.

다른 한편으로, 다수의 구간 특징 벡터를 이용하여 복잡한 분류기를 확실하게 훈련 시킬 수 있고, 따라서 그들의 파라미터가 정확하게 추정될 것이다. 이로 인해 전역적 특징을 사용하는 경우보다 분류 정확도가 높아질 수 있다.

2.3.3 음성 특징의 범주



<그림 2-4> 음성 특징의 범주

현재 음성 신호에서 많이 사용되고 있는 특징은 그림 <2-4>와 같이 Continuous feature, Qualitative feature, Spectral feature, TEO-Based feature 네 가지 범주로 구분한다[17]. 다양한 음성 신호 특징 중에서, 인간의 음성 신호에서 감정을 잘 표현할 수 있는 특징을 찾는 것이 중요하다. 자세한 음성 신호 특징들에 대한 소개는 이하 설명하도록 한다.

(1) Continuous feature

대다수 연구자는 피치나 에너지와 같은 운율적인 연속성이 있는 특징이 발화의 정서적 내용을 전달한다고 생각한다. Williams와 Stevens의 연구[18]에 따르면, 화자의 감정이 흥분

된(arousal) 상태의 높고 낮음은 주파수 영역의 전반적인 에너지 분포와 음성 신호의 정지(pause) 빈도와 지속 시간(duration)에 영향을 미친다. 음향적 특징은 아래와 같은 카테고리로 그룹화할 수 있다. 이러한 특징들은 통계적인 수치로도 계산해 많이 활용된다.

① 피치(pitch) 관련 특징
② 포먼트 주파수(formant) 관련 특징
③ 에너지(energy) 관련 특징
④ 시간(timing) 관련 특징
⑤ 조음(articulation) 관련 특징

기본 주파수 F_0(fundamental frequency), 에너지(energy)
: 평균, 중앙값, 표준편차, 최댓값, 최솟값, 범위(range), 선형회귀계수(linear regression coefficients), jitter, shimmer

지속시간(duration): 음성 발화 비율, 음성 구간과 음성 구간이 아닌 구간의 지속시간의 비, 가장 긴 음성 구간의 지속시간 길이

포먼트 주파수(formant): 첫 번째와 두 번째 포먼트, 대역폭(bandwidth)

(2) Qualitative feature

발화된 음성의 음질과 인지된 감정에 강한 상관관계가 있다는 실험 연구 결과가 있다[19]. 음성 신호와 상관있는 음성의 음질(voice quality)은 아래와 같은 그룹의 범주로 분류할 수 있다.

① 음성 수준(voice level): 신호의 크기(amplitude), 에너지(energy), 지속시간(duration)은 음성의 수준을 측정하는 방법이다.
② 음성 피치(voice pitch)
③ 어구(phrase), 음소(phoneme), 단어(word), 특징 경계(feature boundary)

④ 일시적 구조(temporal structure)

하지만 다른 특징들에 비해 음질의 이러한 특징들이 음성의 감성을 정확히 전달하지 못한다는 두 가지 이유가 있다. 첫 번째, tense, harsh, breathy와 같은 감정적 레이블(label) 용어는 연구자에 따라서 서로 다른 해석을 가질 수 있다. 예를 들어, 긴장된(tense) 목소리는 분노(anger), 기쁨(joy), 두려움(fear)과 관련이 있다고 제안하는 학자가 있는 반면에 호흡이 많은(breathy) 목소리가 분노(angry)와 기쁨(joy)이 관련 있다고 제안하고, 슬픔은 공명이 있는(resonant) 특징이 있다고 주장하는 학자도 있다. 이처럼 같은 용어에 대한 서로 다른 해석의 영향이 연구자들에게 미쳐 음질의 특징과 음성 신호 감성 간의 불일치를 만든다.

두 번째 이유는 음성 신호에서 바로 음질의 특징이라는 용어를 자동으로 결정할 수가 없다는 문제점이 있다.

(3) Spectral feature

피치 및 에너지와 같은 시간 의존적 음향 특성 외에도, 스펙트럼 특징은 종종 음성 신호 내에서 단시간에 대한 대푯값으로 선정된다. 감정에 따라 스펙트럼 에너지의 분포가 다른 주파수 대역대에서 나타난다.

그 예로, 기쁨(happiness)의 감정은 높은 주파수 영역에서 더 큰 에너지를 가지고, 슬픔(sadness)의 감정은 같은 영역에서 기쁨(happiness)보다 작은 에너지값을 가진다. spectral 특징은 전형적인 선형 예측 계수(LPC, Linear Predictive Coefficient)[16], One-Sided Autocorrelation Linear Predictor Coefficients(OSALPC)[17] 등의 방정식에서 추출할 수 있다.

이외에 가청주파수 영역 대의 스펙트럼 분포를 더 잘 활용하기 위해서 대역 통과 필터 뱅크(band-pass filter)를 사용한다. 추정 스펙트럼은 대역 통과 필터 뱅크를 통과한 후 추출된다. 인간이 pitch를 인지하는 과정은 선형 스케일을 따르지 않기 때문에, 필터 대역폭 또한 Bark scale[20] 또는 Mel-frequency scale[21]과 같은 비선형 주파수 스케일을 사용한다. 다양한 스펙트럼의 특징 중에서 단연 감성 인식에서 중요하게 다루어지고 있는 특징에는 MFCC(Mel Frequency Cepstrum Coefficients)가 있다.

MFCC는 음성 인식과 화자 인식 분야에서 널리 사용된다. MFCC는 사람의 청각기관이 저주파수 대역에서 민감하지만, 고주파수 대역에서 상대적으로 둔감한 특성을 갖는 것에

착안하여, 멜 스케일(mel scale)에 기반해 표현한 음성 특징이다. 멜 스케일은 Stevens 등에 의해 명명되었고, 물리적인 음의 높이와 청각 인지적 음높이의 관계를 표현한 것이다[22]. 식 (1)은 Hz 단위로 표현한 물리적 주파수 f를 mel 단위의 청각 인지적 음의 높이 m으로 변환하는 식이다.

$$m = 1127\log_e\left(1 + \frac{f}{700}\right) \quad (1)$$

MFCC 추출과정은 <그림 2-5>와 같다.

<그림 2-5> MFCC 추출과정

음성 신호에 윈도우 함수를 씌운 다음 매 프레임 단위로 이산 푸리에 변환(DFT, Discrete Fourier Transform) 과정을 통해 시간 영역에서 주파수 영역으로 변환시킨다. 하지만 실제 이산 푸리에 변환 과정은 연산의 효율성을 위해 빠른 푸리에 변환(FFT, Fast Fourier Transform)로 대체한다. 이 과정을 거친 후, 멜 스케일을 가지도록 식 (1)을 사용하여 주파수 축을 워핑한 후, 이 스케일과 동일한 대역폭을 갖는 삼각 필터뱅크를 통해 각 필터뱅크에 해당하는 에너지를 계산한다. 마지막으로 로그 함수를 취해서 <그림 2-5>의 다섯 번째 단계인 이산 코사인 변환(DCT, Discrete Cosine Transform)을 통해 최종적인 MFCC 값을 구한다[23].

(4) TEO-Based feature

Teager의 실험 연구에 의하면, 음성은 보컬 시스템의 비선형 공기 흐름에 의해 생성된다[24]. 음성에서 강세가 있는 상황에서, 발화자의 근육 긴장도가 공기 흐름에 영향을 주게 된다. 따라서 비선형 음성 특징은 감성 인식에 꼭 필요한 특징이 될 수 있다고 주장한다.

TEO란 Teager와 Kaiser 두 연구자가 사람의 청력이 에너지를 탐지하는 과정인 것을 고려하여 개발해 만든 Teager Energy Operator의 약자를 의미한다[25-26]. 이산신호 x[n]에 대해 TEO는 식 (2)와 같이 정의한다.

$$\Psi x[n] = x^2[n] - x[n-1]x[n+1] \quad (2)$$

앞서 서술한 바와 같이, TEO의 특징은 음성의 강세를 잘 찾아내기 때문에, Teager energy는 음성에 있어서 시끄러움(loud), 화남(angry), 명확함(clear) 그리고 중립(neutral)으로 그 효과들을 분류한다.

Cortes 등의 연구[27]에서 TEO 기반 특징(TEO-based feature)에는 TEO-decomposed FM variation(TEO-FM Var), normalized TEO, autocorrelation envelope area(TEO-Auto-Env), critical band-based TEO autocorrelation envelope area(TEO-CB-Auto-Env)가 중립적인 감정과 강세가 있는 발화구간을 찾는 특징들로 제안됐다.

2.4 감성음성 인식 기술 동향

2.4.1 국외 기술 동향

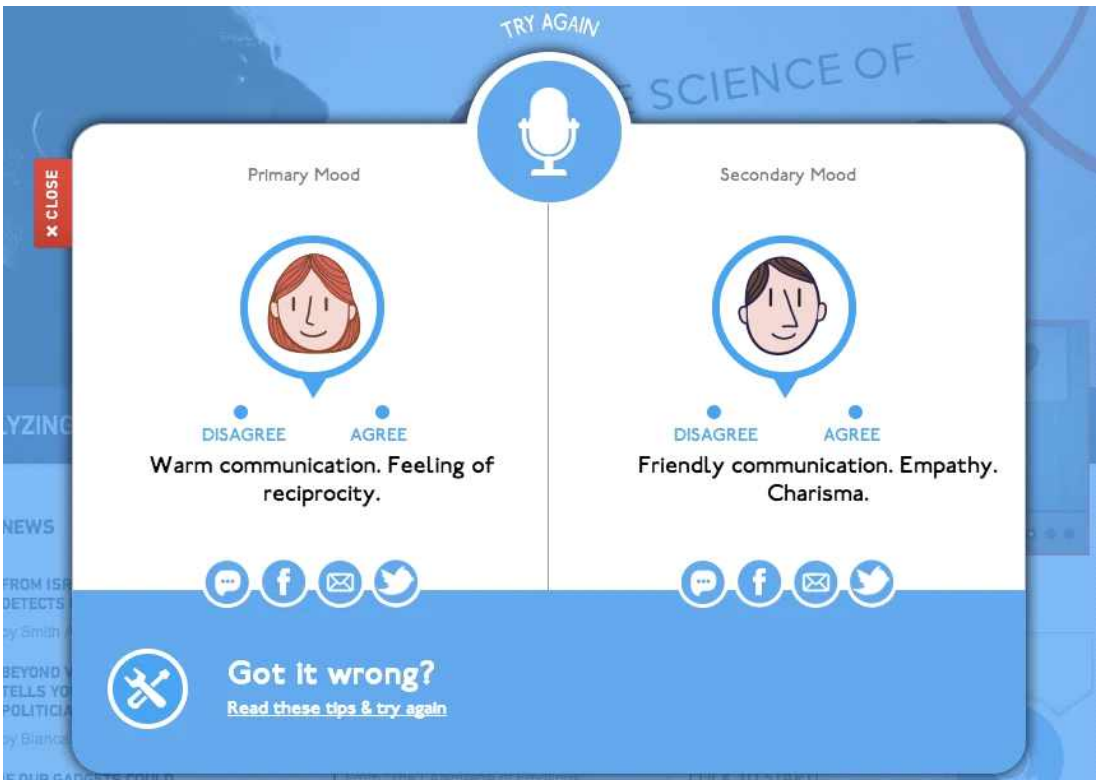
(1) Beyond Verbal(이스라엘)

2012년에 설립된 이 회사는 2014년에 Beyond Wellness API[28]를 출시했다. 2014년에는 스마트폰이나 마이크가 장착된 웨어러블 장치에서 실제 발화 내용이나 문맥을 고려하지 않고 발화자의 억양을 분석해 그의 감정을 추론하는 센서를 선보이며 시장을 확대해 나갔다. 이 회사에는 Moodies와 Empath라는 두 개의 무료 소비자용 앱과 Beyond Clinic이라는 임상의를 위한 앱이 있다. 특히 Empath는 사람의 음성으로부터 질병을 진단하는 연구에 적용할 수 있다.

비욘드 버벌은 메이요 클리닉(Mayo Clinic), 스크립스(Scripps)과 같은 의료기관과의 공동

연구를 통해 환자의 음성 신호가 심장질환 관련 신호를 포함하고 있다는 것을 발견하였다. 즉 음성 신호가 심장질환의 바이오 마커 역할 기능 수행한다는 것이다[29].

<그림 2-6>은 이 회사가 개발한 앱 무디스의 실행화면 일부분을 보여주는데 무디스를 사용하여 미국의 전 대통령 버락 오바마의 히로시마 평화공원 연설에 관한 분석을 한 사례가 있다. 이에 따르면 오바마의 연설 내용 자체는 전쟁의 처참함을 호소하는 것이었으나 세계 평화를 호소하는 고무적인 감정과 부진한 핵무기 폐기에 대한 좌절감이 복합적으로 나타난 연설이었다고 한다.



<그림 2-6> 비욘드 버벌의 앱, 무디스(Moodies)

Beyond Verbal이 가지고 있는 핵심 기술은 목소리를 분석하여 인간의 감정과 성격을 추론하는 기술로 음성으로 구동되는 장치 및 앱이 인간과 마찬가지로 감정적인 수준에서 사용자와 상호 작용할 수 있도록 하는 것이 이 회사의 궁극적인 목표이다. 21년 동안의 연구를 바탕으로 이 회사는 170개국에서 230만 개의 음성 샘플을 분류했으며 API에 대해 8개의 특허를 받은 저력으로 기업용 솔루션, 건강관리 도구 등 다양한 분야에 활용 영역을 확대하고 있다.

(2) audEERING(독일)

이 회사는 뮌헨공대에서 시작하여 2012년 설립된 기업(Gartner(2017))으로 기계학습의 알고리즘과 심리 모델을 기반으로 음성, 음악 및 기타 사운드 분석을 위한 솔루션 개발을 주력 사업으로 하는 업체이다. audEERING은 머신러닝 알고리즘과 정교한 심리 모델을 기반으로 음성, 음악과 같은 소리 분석을 위한 독점 솔루션을 개발하였다.

특히 이 회사가 개발한 sensAI 감정 제품은 특정 비즈니스 프로세스를 개선하기 위해 소비자의 다양한 감정적 상태를 식별하는 데 사용되고 있다. 이에 더하여 특허 출원 중인 VocEmoApI 기술을 사용하면 발화자의 사용 언어나 그의 국적과 관계없이 심리적으로 입증된 개념을 기반으로 50가지 이상의 미묘한 기분 상태를 추론하는 것이 가능하다[30].

또한, audEERING의 기술은 지역 오디오 설정 및 건강상태, 예를 들어 발화자가 감기에 걸렸는지와 같은 감정적인 행동에 영향을 미치는 상황을 감지할 수 있는데, 이 기술은 기계학습, 심화 학습 및 전송 학습 방법을 적용하여 짧은 양의 음성 데이터(1초)만 사용할 수 있더라도 실시간 분석에 적용 가능하다는 강점이 있다.

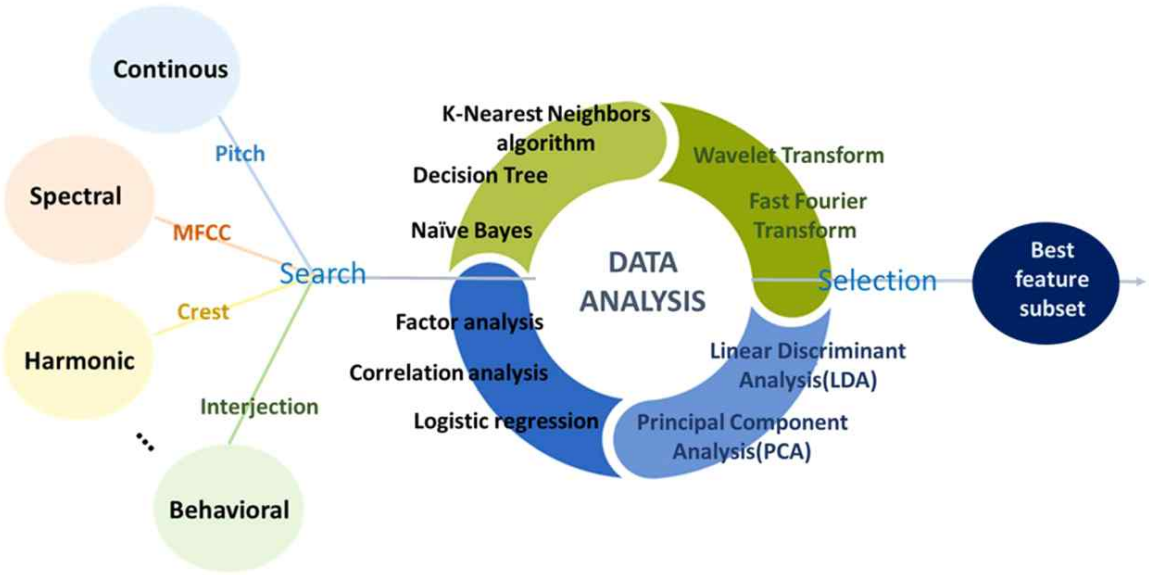
2.4.2 국내 기술 동향

(1) 세종대, 상명대, 건국대의 공동연구

세종대학교 연구진과 상명대학교, 건국대학교 연구진들은 최근 음성정보만을 이용한 감성 인식 기술을 개발하기 위해 국내에서 활용 가능한 감성음성 데이터베이스를 구축하고, 이를 기반으로 최적의 특징 추출, 분석, 인식 및 인덱싱 기술개발을 진행하고 있다[1]. 미디어 기반 감성 인식 기술개발에 있어서 필요한 핵심요소는 한국형 감성 데이터베이스를 구축하는 것이다. 이를 기반으로 하여 최적화된 음성기반 특징요소를 추출하고 알고리즘을 개발하는 것을 목표로 한다.

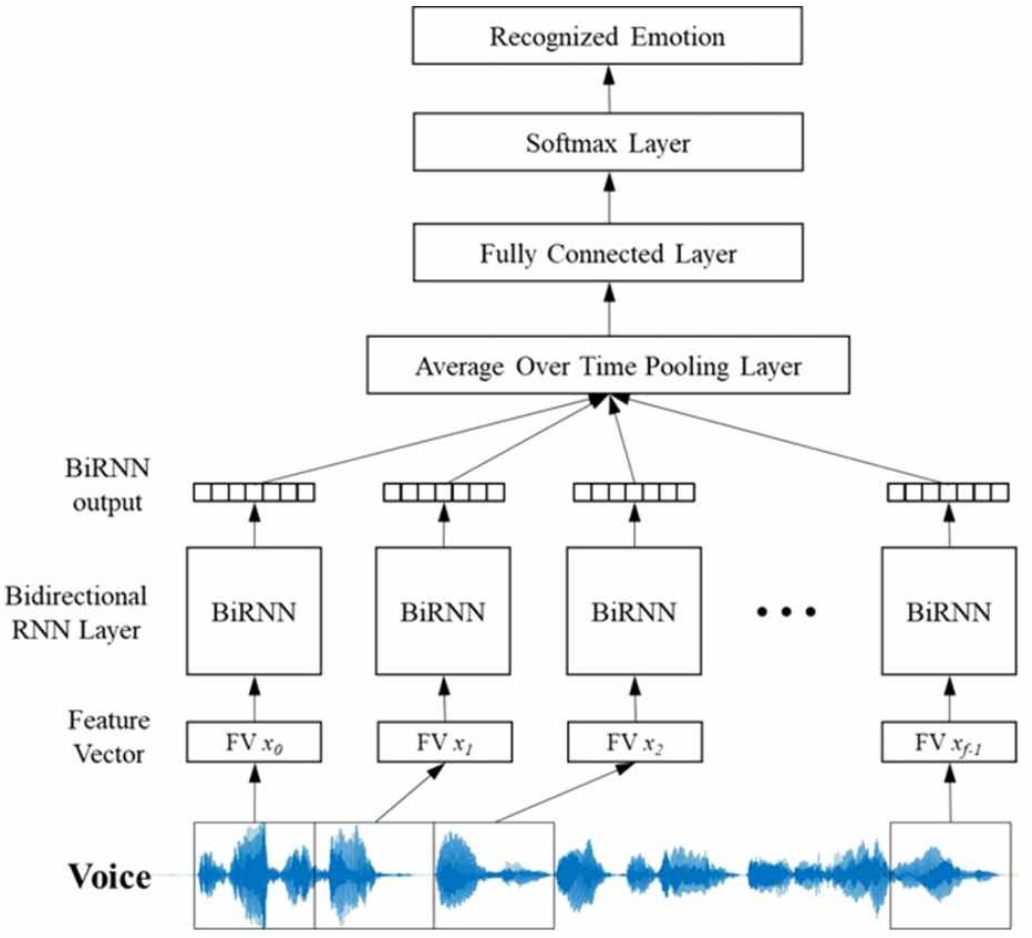
한국어 기반 표준화된 데이터베이스 구축을 위해 한국형 감성 데이터베이스는 기존의 훈련된 연기자를 통하여 감정유발 시나리오에 근거하여 음성을 녹음하고 이를 표준화하는 과정을 통하여 데이터베이스 구축이 진행되고 있다. 또한, 기존의 감성 인식에서 많이 사용되는 Pitch, MFCC 외의 언어학, 심리학 등의 여러 학제 간 융합의 측면에서 접근한 감정

표출의 비언어적 발화행태 등의 새로운 감성 인식 특징요소를 추출하고 최적화하려고 노력하고 있으며 이렇게 추출한 특징으로 감성 음성연구를 수행하면서 보다 효과적인 조합은 없는지 새로운 조합을 발굴하려고 노력하고 있다. <그림 2-7>은 이러한 연구의 방향을 그림으로 정리한 것이다.



<그림 2-7> 음성기반 최적의 특징요소 선정

딥러닝을 음성연구에 적용하는 부분은 다음과 같다. 인공지능 기술개발 중 음성에 최적화된 딥러닝 기법 중 감정 분류에 최적화된 알고리즘 개발을 진행하고 있다. 이 알고리즘은 발화의 특정 시점 이전과 이후 정보를 모두 사용하여 앞뒤 정보에 의존성을 갖는 발화의 특징을 더 잘 반영하고, 출력값에 AoT(Average over Time) 방식의 변형된 모델을 추가하여 음성 신호의 짧은 부분적인 구간뿐만 아니라 전체 구간에 대한 정보를 통해 감정 분류가 가능하도록 돕고 있다 <그림 2-8>은 이러한 AoT Bi-directional RNN Model을 개략적으로 정리한 것이다. 감정을 인식하고 이를 이용한 인터페이스 기술은 새로운 형태의 스마트 미디어 콘텐츠와의 소통 및 감정을 전달하는 방법이며, 감정 ICT 제품과 서비스 콘텐츠의 성장기반 구축 및 기술의 글로벌 확산과 신규 비즈니스 창출에 이바지할 수 있다.



<그림 2-8> AoT Bi-directional RNN Model

최근 다국적 대기업은 정체된 ICT 제품구매 촉진을 위해 감성전달 기술에 역량을 강화함으로써 실감·감성기술 중심의 시장을 주도하려는 추세를 보인다. 소비자 참여형 미디어 생산 및 공유 기술을 통해 사업자 주도의 서비스에서 사용자의 감성 키워드 검색기반 개인화 서비스로의 진화된 감성 미디어 서비스를 통해 사용자 경험 창출 및 신 ICT 산업이 조성될 수 있을 것이다.

감성 인식 기술은 개인화 서비스 제공을 통한 사용자의 긍정적 심리 형성 및 정서 불안 해결 등 개인적 삶의 질 향상, 감성 기반 원격 의료에의 연동으로 사용자의 심리상태를 진단하는 의료서비스 및 심리적 문제 해결에 이바지할 수 있는 감성 기반 심리치료 서비스를 제공할 수 있다. 또한, 한국인의 정서를 반영한 한국어로 수집된 데이터 통한 한국인 사용자 맞춤형 감성 미디어 빅데이터 확보가 가능하며, 다양한 스마트 기기를 활용한 사용

자 감성 최적화 콘텐츠 검색 및 추천, 감성 프로파일링 등 기술개발의 활성화에 큰 도움을 줄 수 있다. 이러한 감성 기반 연구를 통해 미디어 콘텐츠의 소비 증가에 따른 새로운 패러다임 제시와 신시장 개척, 스마트 가전, 문화기술 등의 타 산업과의 융합을 통한 동반 성장이 기대된다.

(2) 감성 인공지능 조나단

딥러닝 기술을 전문으로 하는 ㈜아크릴[31]은 인공지능 플랫폼 '조나단'을 통해 사람의 말과 글, 표정을 읽는 인공지능 기술을 선보이고 있다. 조나단은 조나단 브레인, 조나단 프레임, 조나단 툴로 구성된 인공지능 플랫폼인데 이 중에서 감성 인식을 담당하는 것은 조나단 브레인이다. 이는 사람의 말과 글, 표정, 감정을 이해하는 지능형 프레임워크이다. 챗봇, 휴머노이드와 같은 대화형 인터페이스에 적용되며, 금융, 의료, 쇼핑 등 다양한 분야에 활용될 수 있다. 조나단 브레인의 핵심으로 여겨지는 감성 지능도 여러 기술로 구성되어 있는데 텍스트 감성 인식, 음성 감성 인식, 이미지 감성 인식, 멀티모달 감성 인식, 총 네 가지 기술이다. 이 중 음성 감성 인식은 주어진 음성을 분석해 감정을 인식하는 딥러닝 모듈인데 7종의 감정을 파악할 수 있다.

멀티모달 감성 인식은 텍스트, 음성, 이미지를 모두 한 번에 분석해 감정을 인식하는 기술인데 7여 종의 감정을 읽는다.

<그림 2-9> 조나단의 멀티 모달 감성 인식

감성 지능의 가장 큰 특징은 감정을 세밀하게 들여다본다는 것이다. ㈜아크릴에 따르면 “보고, 듣고, 읽고 이해하는 AI는 많지만 34가지 감성 인식이라는 높은 해상도로 사람의 감성을 판단하는 AI는 세계 어디에도 없다”라고 한다.

마지막으로 대화 지능은 음성과 텍스트로 사람과 대화할 수 있는 지능이다. 공감 질의-응답 매칭, 생성형 대화 봇, 두 가지로 구성되어 있다. 질의-응답 매칭은 이용자가 작성한 글에 가장 적합한 답변을 매칭해 주는 기술이다. 작성 글과 그에 걸맞은 답변을 머신러닝으로 학습시킨 뒤 이 데이터를 근간으로 매칭한다. 생성형 대화 봇은 자연어 질의를 분석한 뒤 이에 부합하는 답글을 자연어로 내보내는 봇이다.

제3장 감성 음성 합성 기술

제3장에서는 감성 음성 합성의 기술적 측면에 대해 논의하고, 더 높은 수준의 프레임워크를 기반으로 한 실제 응용을 보여주며, 균일하지 않은 단위 선택 기반 합성 및 음성 변환 기술로 감성 음성 합성을 구현하는 것과 관련된 딥러닝 기반 감성음성합성 최신 기술을 소개하도록 하겠다.

3.1 감성 음성 합성의 개요

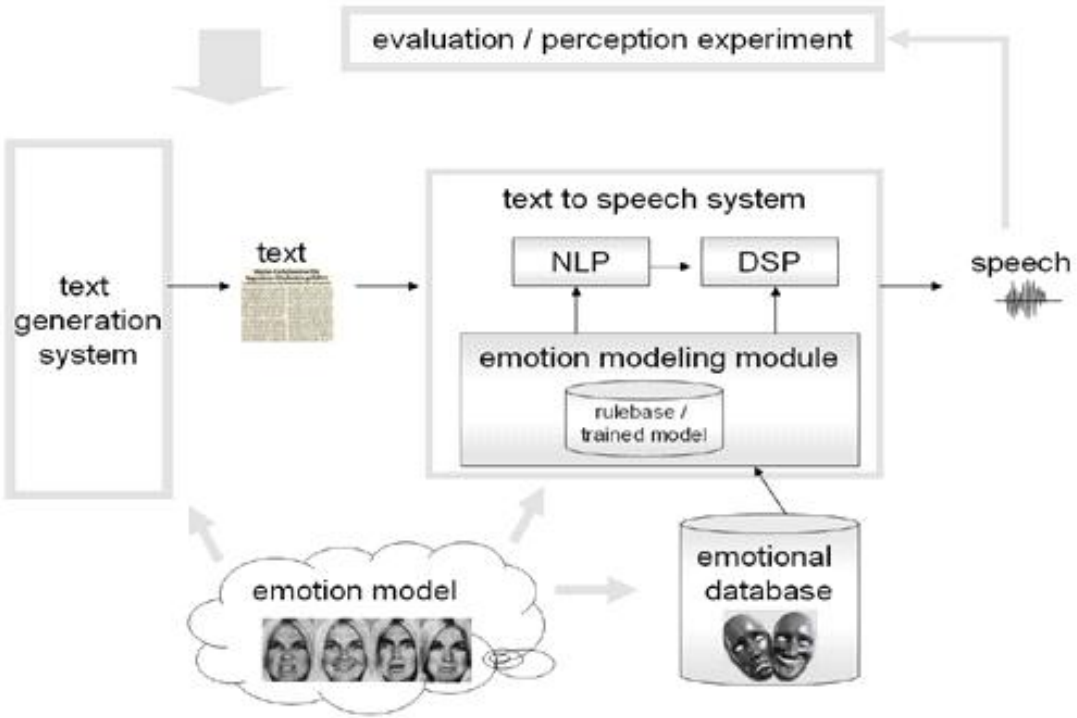
3.1.1 감성 음성 합성의 개념

감성 음성 합성은 인간과 기계의 상호작용이라는 머나먼 여정에서 해결해야 할 중요한 과제이다. 발화할 때 강세의 변화와 감정 없이 대화하는 사람은 없다. 감성은 명백히 음성에 존재해야 하는 필수요소임에도 불구하고 감성음성 합성은 현재 음성 합성에서 누구나 손쉽게 구현할 수 있는 주제는 아니다. 그 이유는 사람 음성 표현의 복잡성에 있다. 현재의 최신 음성 합성 기술로도 이해하기 쉬우면서도 자연스럽게 실제 사람이 말하는 것과 같은 감정 삽입이 자유자재로 되는 음성 합성은 실제 구현이 쉽지 않은 것이 현실이다. 본장은 사용 사례 및 감정 모델에서 기술적 접근에 이르기까지 감성 음성 합성의 다양한 측면과 관련하여 최신 기술에 대한 이해를 돕기 위해 작성하였다.

심리학자들조차 “심리학자를 제외한 모든 사람이 감정이 무엇인지 알고 있다.”라고 말하는 일화는 유명하다. 이처럼 감정을 다루는 주제는 설명하기 가장 부정확하고 어려운 문제이다. 다행인 것은 이 문제는 감정이 담긴 음성 합성보다 감성 인식과 훨씬 밀접한 관련이 있는데, 감성음성을 인식하는 문제는 명확히 정의 내리기 어렵고 그 종류도 많은 "실제 세계"의 정확한 감정 모델이 필요하지만, 감성 음성 합성의 경우에는 몇 가지 기본 감정으로 구현하는 것이 가능하기 때문이다[32]. 일반적으로 말하면, 음성 합성은 다음의 3가지 종류로 나눌 수 있고 이 세 개의 복합 형식으로 음성 합성을 하는 것도 가능하다.

① 지하철 안내 시스템에서 사용되는 것과 같은 음성 응답 시스템
② 재합성 또는 복사 합성(Re- or copy-synthesis)은 음성 관련 기능에서 음성 신호를 변경하는 데에 사용된다. 특별한 경우는 음성 변환, 예를 들어 음성 변환의 경우 원시 화자에서 목표 화자로 음성 신호를 변경하는 것이다. 이러한 합성 기술은 음성 변환 및 감정표현을 생성하는데 사용할 수 있다.
③ 임의 음성 합성기(Arbitrary speech synthesizer)는 ①, ②와는 대조적으로 목표 언어의 한계를 참작해도 모든 종류의 입력을 처리할 수 있다. 다만, 임의 음성 합성기라 할지라도 다소 사용 범위가 제한된다는 점에 유의해야 한다.

텍스트와 함께 제공되는 정보에 따라 텍스트 음성 변환(text-to-speech system)과 개념 음성 변환 시스템(concept-to-speech)으로 나눌 수 있다. 당면한 주제와 관련하여 개념 음성 변환시스템은 대상 감정으로 텍스트에 자동으로 레이블을 지정할 수 있다.



<그림 3-1> 감성 음성 합성 시스템의 일반적인 구조

감성 음성 합성 시스템의 전체적인 구조가 <그림 3-1>에 나와 있다. 합성될 텍스트는 입

력으로 제공되거나 텍스트 생성 시스템에 의해 생성된다. 텍스트 생성 시스템은 본 장에서 직접 다루지는 않지만, 텍스트에 감정표현을 위해 주석을 기재하는 방법은 3.3절에서 기술한다.

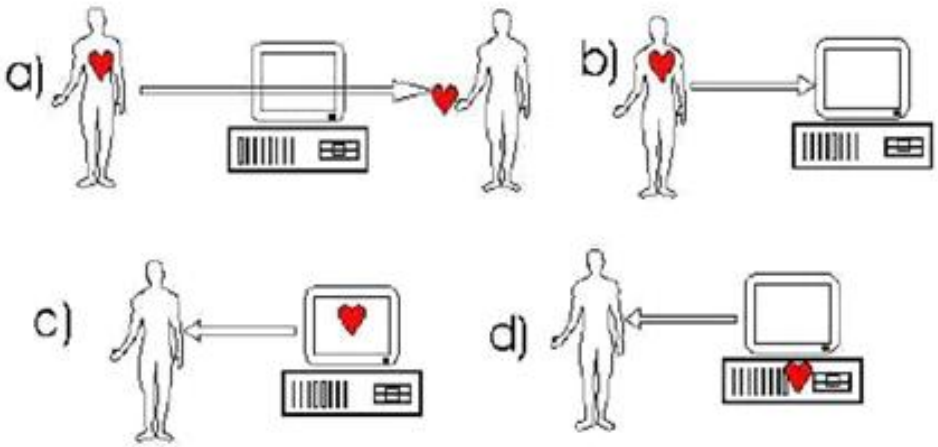
TTS(text-to-speech)라고도 불리는 음성 합성은 자연어처리(NLP, Natural Language Processing)로 텍스트를 먼저 분석하고 운율적 특징(prosodic feature)으로 정렬된 발음 표현으로 변환하여 텍스트를 음성으로 변환한다. 이러한 하위 모듈은 모두 감정 모델링 구성 요소의 영향을 받을 수 있는데 음성 신호를 생성하기 위해 디지털 신호처리(DSP, Digital Speech Processing) 컴포넌트로 전달되는 구조로 감정적 표현을 생성하는 접근법은 3.5절에서 더 자세히 서술할 것이다.

음성 신호의 특징은 스펙트럼(spectral, 음성의 '소리'), 운율적 특징(prosodic, 음성의 '멜로디'), 발음적 특징(phonetic, 음소의 감소와 정교한 조작), 개인 특유의 언어(idiolect, 단어의 선택) 및 의미적 특징(semantic features)이다. 이 모든 것들이 감정표현에 영향을 받으므로, 음성 합성 시스템은 이러한 특징 중 일부에 치우쳐서는 안 되며 전체적으로 고려해야 한다. 규칙이 데이터에서 파생되는지 또는 통계 알고리즘이 학습되는지에 관계없이 정서적 표현을 생성하는 구성 요소는 감성음성 데이터베이스에서 학습해야 한다[9-12]. 이것들로부터, 규칙 기반(rule base) 또는 모델이 생성되어 음성 합성기를 제어하는 데에 사용할 수 있다. 감정 처리 시스템의 모든 구성 요소는 기본적으로 동일한 감정 모델을 가져야 한다.

감정을 지정하고 설명하기 위한 다양한 접근 방식은 3.2절에서 자세히 설명한다. 감성음성 합성 시스템이 작동 중이면 청취자만이 시스템의 정상 구동 여부를 판단할 수 있다. 감성음성 합성기의 성능은 물론 그 사용 목적에 달려 있다. 만화 인물의 음성을 만드는 음성 합성기는 언어 장애인의 목소리를 더 자연스럽게 만드는 시스템과는 그 요구 사항이 다르다. 감성 음성 합성의 응용 분야는 다음 3.1.2절에서 기술한다.

3.1.2 감성음성합성 적용 분야

<그림 3-2>에서는 인간의 기계 상호작용에서 감정 처리의 가능성을 도시한다. <그림 3-2>와 관련된 보다 자세한 내용은 참고문헌 [33]에서 확인할 수 있다. 이처럼 감정은 정보처리 시스템 여러 곳에서 다뤄질 수 있다.



<그림 3-2> 사람-컴퓨터 시스템 간 감정 처리의 방법

(1) 방송(Broadcast): 감정표현은 인간 커뮤니케이션에서 중요한 정보 채널입니다. 전기 통신에서는 감정적인 의사소통을 위한 특별한 채널을 제공하는 것이 바람직할 수 있다. 전자메일 통신에 사용되는 소위 이모티콘이 대표적인 예이다.

(2) 인식(Recognition): 인간의 감정표현은 다양한 방식으로 분석될 수 있으며 이 분석을 통해 얻은 지식은 시스템 반응을 변경하는 데에 사용된다.

(3) 시뮬레이션(Simulation): 자연스러운 인터페이스를 향상하거나 동요된 음성 스타일로 긴급 메시지를 발언하는 등의 추가 커뮤니케이션 채널에 액세스하기 위해 시스템에서 감정표현을 흉내 낼 수 있다.

(4) 모델링(Modeling): 감정표현의 내부 모델을 사용하여 사용자 또는 시스템 상태를 나타내거나 의사 결정에 영향을 미치는 인공지능 모델로 사용할 수 있다.

(1), (3) 및 (4)의 경우, 감정표현을 사용하여 감정 상태를 표현하거나 전달할 수 있다. 감성 음성 합성은 아래와 같은 곳에 적용될 수 있다.

- 감정을 삽입한 인사
- 보형물
- 감정을 삽입한 채팅용 아바타(chat avatars)
- 게임, 그럴 듯한 캐릭터

- 적응형 대화 설계(dialog design)
- 적응형 모습 설계(persona design)
- 대상 그룹에 특화된 광고(target-group specific advertising)
- 믿을 수 있는 대리인(believable agents)
- 인조인간(artificial humans)

실제로 이러한 응용 프로그램은 이미 시장에 나와 있다[34]. 이러한 응용은 인공지능 개발과 밀접한 관련이 있다. 감정과 지능이 밀접하게 섞여 있으므로[35], 대화 능력에 대한 사용자의 기대에 부응하기 위해 지능이 없는 컴퓨터 시스템이 감정적으로 반응하는 것처럼 보일 때 각별한 주의가 필요하다. 감성 음성 합성을 위해서는 많은 미묘한 말하기 스타일의 모델링과 자연스러운 언어학적 요소를 합성된 음성에 추가해야 한다. 과장된 기본 감정을 주로 합성한 과거의 연구 결과를 고려할 때 아직 감성 음성 합성은 갈 길이 멀다.

3.2 감정 모델링

정서적 표현은 일반적으로 분노(anger)나 권태(boredem)와 같은 특정 감정 범주 세트를 구별하거나 흥분(arousal), 합의(valence) 또는 지배(dominance)와 같은 정서적 차원을 사용하여 범주형 시스템으로 모델링 된다. 범주형 모델은 인간의 일상적인 의사소통에서 직관적으로 이해하고 잘 정립할 수 있는 매력이 있다.
차원 모델을 사용하면 감정은 흥분 또는 활성(이완에서 긴장에 이르기까지), 유쾌함 또는 합의(감정의 주관적 긍정을 구분) 및 지배(사람이 느끼는 감정의 강도)로 나눌 수 있다. 수많은 추가적인 차원 모델이 제안되었지만, 이 세 가지가 전통적으로 가장 일반적인 차원 모델이다.

음성 합성을 염두에 두고, 흥분(arousal)과 활성(activation) 차원에 대한 적절한 음향 수정 규칙을 쉽게 도출할 수 있다. 둘 다 근육 긴장과 직접적인 관련이 있지만 다른 차원에서는 이를 도출하는 것이 매우 어렵기 때문이다. 따라서 감정 상태는 일반적으로 범주형으로 모델링 된다. 차원 시스템은 더 실용적이고 소위 '완전한' 감정이 거의 발생하지 않는 '실제 세계'의 부정확성을 모델링 하는 데에 더 적합하다.

감정 이론 또는 OCC(Ortony, Clore, Collins) 모델과 같이 심리학과 인공지능과 더 가까운 다른 모델은 오늘날 감정적 음성 합성과 관련하여 이전에는 큰 역할을 하지 않았지만, 점차 딥러닝을 적용한 감성 음성 합성으로 바뀌고 있기는 하다. 이와 관련하여서는 3.6절에서 한국의 음성 공학 기술 신생 기업인 네오사피엔스의 감성 음성 합성 최근 연구를 살펴보겠다.

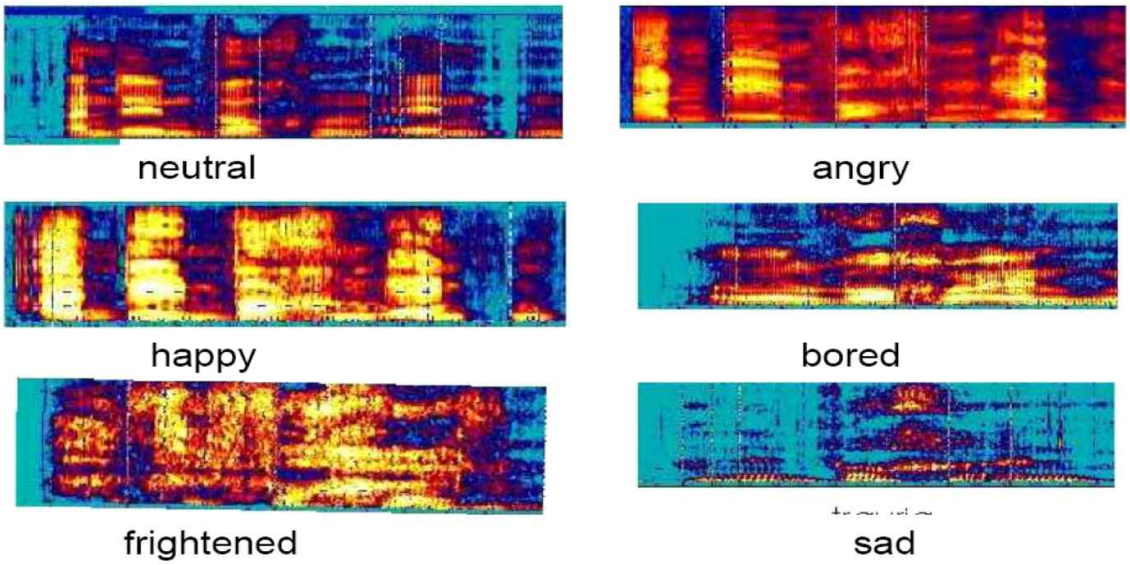
3.3 주석 형식

Loquendo와 같은 현재의 상용 음성 합성기는 IBM 연구팀의 연구 결과[36]에 설명된 것처럼 감정적 준 언어학적인 요소를 감성 음성 합성의 요소 단위로 추가했다. 텍스트 음성 변환시스템을 조정하는 도구인 Loquendo TTS 디렉터를 사용하면 계층적 드롭다운 메뉴에서 감정 단위(expressive units)를 선택할 수 있다. 정서적 마크업 언어를 위한 W3C 인큐베이터 그룹이 SSML(Speech Synthesis Markup Language)을 개발하였는데 이 그룹은 감정표현을 위한 마크업 언어의 개발을 목표로 한다. 인간과 기계의 상호작용에서 이를 이용하여 감성 인식, 합성 및 모델링에 사용할 수 있다. SSML에 대한 상세한 설명은 부록을 참고하기 바란다.

3.4 음성 신호 지각 단서

감성음성의 지각적 단서를 조사하기 위해, EmoDB[9]를 사용하여 감정별로 스펙트로그램을 획득한 결과는 <그림 3-3>과 같다. EmoDB은 독일어로 된 감성음성 데이터베이스이며 <그림 3-3>은 독일어 문장 “In sieben Stunden wird es soweit sein” (뜻: 그것은 7시간 안에 준비될 것입니다)를 각각 다른 감정으로 낭독했을 때의 스펙트로그램을 보여주고 있다. 주파수 대역의 다른 진폭에서 볼 수 있듯이, 표시된 감정은 운율, 음성 구현 및 음성 품질과 관련하여 다양하게 나타난다. 의미 및 개인화 언어와 같은 특징의 수정은 텍스트 생성 시스템의 범위 내에 있지만, 음향 특징은 음성 합성기 자체의 제어로 가능하다. 3.5절에서 볼 수 있듯이, 다양한 감성 음성 합성 접근법은 현재 음성의 자연스러움과 음성 조작에 대한

유연성 사이에서 상충 관계에 있다.



<그림 3-3> 같은 문장을 다양한 감정으로 녹음한 음성의 스펙트로그램

3.5 음성 합성 방법

이 절에서는 현재 연구 환경에서 가장 중요한 합성 방법에 대하여 설명한다. 감성 음성 합성 방법의 주요 차이점은 부자연스럽지만 유연한 파라미터 합성(parametric synthesis)인지 도메인 외부 발화와 관련하여 유연성은 떨어지지만, 자연적인 소리를 만드는 데에 중점을 두는 오디오 샘플 연결방식으로 합성인지 여부이다.

3.5.1 규칙 기반 합성

순수한 데이터 기반 통계적 접근 방식과 달리 이 합성 모델은 인간 음성 생성 메커니즘의 측면을 다양한 각도에서 모델링 하여 음성을 합성하는 방식이다. 이는 '신호 모델링'과는 달리 '시스템 모델링' 접근법이라는 용어로 요약할 수 있다.

본질적으로, 조음 모델은 조음기관의 목표 위치 사이에서 보간 하지만, 속도 및 자유도에 대한 운동학적 제한이 고려된다. 이 모델은 공기 역학적 음향 모델에 기초하여 음성 신호가 생성된다.

감성음성합성에 있어서 조음 합성은 매우 매력적이다. 근육과 조직에 대한 감정적 각성의 영향과 언어 생성 과정과 관련된 신체 부위의 음향 특성을 직접 모델링 할 수 있기 때문이다. 그러나 조음 합성은 여전히 데이터 부족으로 어려우며 음성과 조음기관 움직임의 상관관계는 효율적인 측정 방법의 부족과 후두와 조음기관의 위치와 음파 사이의 연결은 매우 복잡하다. 기존의 모델은 사람의 성대를 대략 단순화한 것에 불과하다.

3.5.2 단위 선택 합성

단위 선택 합성(unit selection synthesis)은 음성 합성에 대한 상업적으로 가장 성공적인 접근 방식이다. 균일하지 않은 단위를 선택함으로써 대형 데이터베이스에서 가장 적합한 음성이 연결되어 큰 위험 없이 음성을 합성할 수 있다. 신호 조작이 가능한 한 많이 줄어들어서, 합성하는 발화가 데이터베이스의 원래 영역에 가까우면 그 결과로 나오는 음성은 음성 데이터를 녹음한 발화자와 유사하며 가장 자연스럽다. 감정 음성을 합성하기 위해서 세 가지 접근 방식이 가능하다.

① 각 감정 스타일에 대한 데이터베이스를 복제하는 방법으로 이는 브루트 포스(brute force) 방법으로 볼 수 있다. 방법이 매우 단순해 매우 제한된 수의 스타일의 음성 합성만 가능하다.
② 단위 선택 과정에서 감정적 목표 기능을 통합하는 방법으로 비언어적인 소리를 표시할 가능성을 명시적으로 포함하는 매우 우아한 방법이지만 매우 큰 데이터베이스에서 수작업으로 기재한 주석이 필요하므로 구현이 쉽지 않다.
③ 음성 신호에 신호 조작 방법을 적용한다. 이를 위해서는 단위를 (반) 파라 메트릭 방식으로 코딩해야 하며 음성 변환 기술을 사용하여 음성 스타일을 변경할 수 있다.

3.5.3 통계 기반 합성

통계적 접근법은 파라메트릭 합성의 유연성과 큰 데이터베이스의 음성 데이터가 가지는 자연스러움을 결합한다. 현재 HMM (Hidden Markov Model) 기반 모델[37]이 가장 성공적이다. 음성은 소스 필터 모델로 모델링 되며 여기 신호와 성대와 관련된 MFCC(Mel

Frequency Cepstrum Coefficients) 또는 포먼트 위치와 관련된 LSP(Line Spectrum Pairs)로 매개 변수화된다. 감성 음성 합성은 일반적으로 목표 감성과 원시 음성 신호의 파라미터를 이동(shift)시킴으로써 이루어진다.

파라메트릭[1] 합성에서 가장 큰 과제는 프로세스의 보코딩 단계에서 비롯된 자연스러움의 부족이다. 음성의 재합성은 품질 손실 없이 가능하더라도 소스 또는 필터 매개 변수가 변경되고 가청 왜곡이 나타난다. 소스 필터 상호작용 관련하여 현재의 파라메트릭 기술에 내재된 부자연스러운 윙윙거리는(buzziness) 잡음 등이 섞이는 것은 모델이 충분하지 않기 때문일 수 있다.

3.5.4 합성 방법의 비교

<표 3-1> 음성 합성 방법 비교

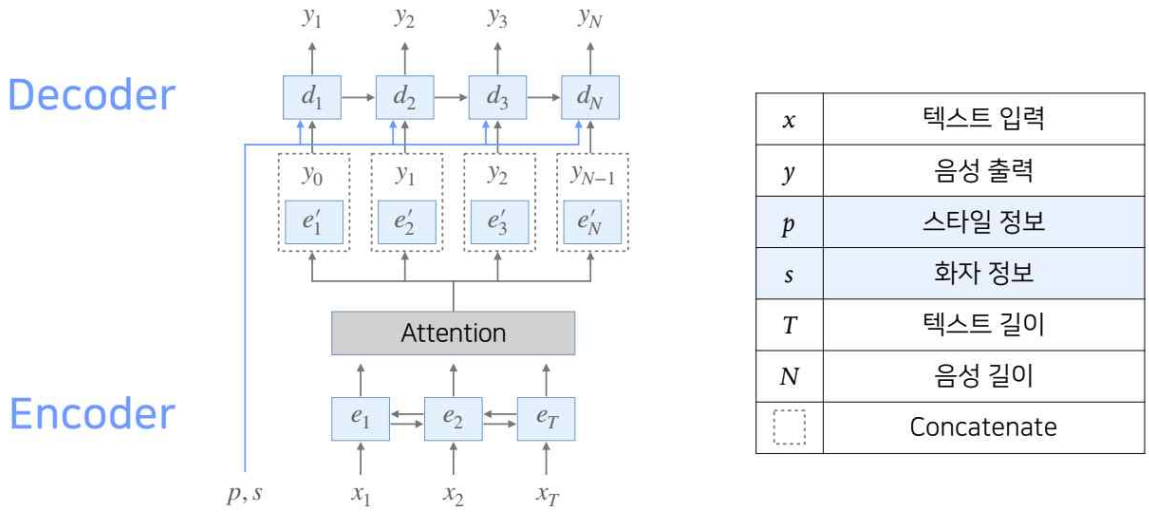
	natural sounding	flexible modification	prosody manipulation	voice quality manipulation
formant synthesis	inherent buzziness due to inadequate source-filter modeling	very flexible due to full parametric control	rule based flexibility	rule based flexibility
articulatory synthesis	only copy synthesis possible	theoretically very flexible due to direct physical modeling of muscle tension	without limitation	yes
diphone-based synthesis	limited naturalness due to many concatenation points	only with respect to prosodic variation	yes, but high jumps introduce audible distortions	only by using several databases
non-uniform unit selection	quite natural within the domain of the database	no flexibility outside the scope of the database	only if emotional models are part of the concatenation function	only if emotional models are part of the concatenation function
HMM-based synthesis	introduces slight buzziness	yes, even linear interpolation between styles possible	yes, if emotional data is part of the database	yes, due to underlying source-filter model

1) 여러 개의 독립적 변수를 사용한 공식에 의하여 정의되는 직선이나 곡선 또는 표면 등의 그래픽 데이터를 처리하는 것으로서 컴퓨터 지원 설계(CAD) 시스템에 쓰이는 기법의 하나.

<표 3-1>은 감정 모델링과 관련하여 다양한 음성 합성 접근법의 특성 일부를 요약한 것이다[38].

3.6 최근 감성 음성 합성 기법

DEVIEW 2019[2] 에서 음성 공학 기술로 새로운 바람을 일으키고 있는 스타트업 네오사피엔스가 “감정연기와 외국어가 가능한 인공지능 성우”라는 제목으로 감성 음성 합성의 최신 연구 성과[39]를 발표하였다. 이 회사에서 감성 음성 합성을 위해 사용한 방법은 음성 합성에서 사용하는 입력값으로 음성을 출력하고자 하는 대본, 즉 문장(text) 이외에 스타일을 받아 출력 음성에 감성을 부여하는 것이다.



<그림 3-4> 네오사피엔스에서 공개한 감성 음성 합성 방법

그러나 이 기술의 문제점은 1) 스타일을 한 문장 단위에서 조정할 수 있었다는 점, 2) 특정한 음소, 단어 단위의 스타일 조정은 여전히 불가능하다는 점이다. 그러나 감성 음성 합성에 있어서 Tacotron을 시발점으로 최근 비약적으로 적용하고 있는 End-To-End 기술을 적용했다는 점에서 이 연구는 큰 의미가 있으며 발전 가능성이 크다.

2) https://deview.kr/2019/schedule

제4장 감성 음성기술과 로봇공학

제4장에서는 비약적으로 발전하고 있는 로봇공학에 사람과 같은 감정이 들어간 음성을 활용하는 것이 가능한지, 로봇이 할 수 있는 감성적 문맥 인식은 무엇인지를 살펴보고 감성음성 기술을 로봇에 구현한 특허 내용을 분석해 본다.

4.1 감성적 문맥 인식

감성적 문맥 인식(emotional context perception)[40]이란 센서들을 통하여 감성 로봇의 감성을 변화시킬 수 있는 상황을 인식하는 것이다. 기존에 주로 연구된 감성적 문맥으로는 시각 센서와 청각 센서를 통해 사람의 감성 상태(human affective state) 또는 인식된 사용자와 관계라고 할 수 있다.

사용자의 관계는 로봇이 사용자를 인식하여 자신에게 적대적인 사람인지 우호적인 사람인지 판단하고 그에 따라 변화된 감성을 행동으로 표현함으로써 사용자에게 지속적인 관심을 가지도록 연구하고 있다.

그러나 사람의 감성 상태는 주로 로봇의 감성을 변화시키는 요인으로 사용되기보다는 로봇이 사람의 감성 상태에 따라 서비스를 제공할 수 있도록 연구되었다. 시각 센서와 청각 센서뿐만 아니라 최근 촉각센서를 통해 사람과의 상호작용에서 인식될 수 있는 감성적 문맥에 관한 연구가 진행되고 있다. 지금까지 엔터테인먼트 로봇 혹은 애완용 로봇에서 일반적으로 촉각센서는 강화학습(reinforcement learning)에서 보상(reward)과 처벌(penalty)을 주기 위한 채널로 이용되었다.

이외에도 로봇의 감성을 변화시킬 수 있는 상황을 감지하기 위해 조도 센서, 온도 센서, 가속도 센서, 후각 센서, 배터리 잔량 센서가 보편적으로 로봇에 사용되고, 타이머(timer)를 이용하여 로봇의 24시간 주기 리듬(circadian rhythm)을 계산하여 감성 변화에 반영시키고 있다.

로봇이 사용자와 의사소통을 하면서 사용자의 감성 상태를 인식할 수 있는 또 다른 중요한 채널은 청각 센서를 이용한 음성이다. 얼굴의 표정에 따른 사용자 감성 상태 인식과

는 다르게, 음성을 통한 사람의 감성 상태를 인식할 때에는 감성 상태에 따른 구조화된 발화(structured utterance)를 찾기가 어려우므로 일반적으로 화자(話者)들에게 주어진 문장들을 감성 상태에 따라 말하도록 하여 화자마다 감성 상태에 따른 발화를 생성한다.

그러나 이렇게 생성된 구조화된 발화는 화자가 의식해서 감성 상태에 따라 발화하기 때문에 실생활에 나타나는 감성에 따른 발화와는 차이가 있을 뿐만 아니라 화자 간의 차이도 있다. 이러한 문제들 때문에 Martínez 등의 연구[41]에서는 영화에서 화자의 감성이 나타나는 대사들을 인위적으로 추출하여 실험에 사용하였고, Kazunori Komatani 등의 연구[42]는 불특정 다수의 감성 상태를 인식할 수 있도록 발화의 기본 주파수 F_0, 크기, 길이, 그리고 발화 간의 시간을 기반으로 계산된 29개의 특징을 제안하였다.

로봇이 같이 대화하는 상대방의 감성 상태에 따라 대화를 한다면, 사람에게 더욱 친근한 느낌을 줄 수 있는 사회성 있는 로봇이 될 것이다. 따라서 이러한 목적을 가지고 많은 연구가 진행되고 있다. 대화 중에 화자에게 나타날 수 있는 감성은 재미있고 즐거운 감정, 당황함과 같이 일시적으로 나타나는 감성과 호감, 긴장감과 같이 대화 중에 변하지 않은 감성으로 구분된다.

Kazunori Komatani는 전자를 일시적 감성(temporary emotion)으로 정의하고 후자를 영구적 감성(persistent emotion)으로 정의하였다. Takayuki Kanda[43]는 일반적으로 사람들은 긴장한 상태로 로봇과 말을 하게 하므로 대화 중에 즐거움과 같은 자율 신경의 작용이 사람에게 나타나지 않게 되어 기존 방법의 성능이 낮아진다고 분석하였다. 따라서 그는 화자의 발화와 표정을 이용하여 화자가 긴장하고 있는지를 판단하고 화자의 긴장을 풀 수 있도록 도와주는 행동을 선택하는 긴장 완화 메커니즘(tension-moderating mechanism)을 제안하였다.

점차 엔터테인먼트, 청소, 방범, 개호(介護) 로봇[3] 등과 같은 다양한 로봇이 우리 생활의 일부가 되어 그 임무를 수행함에 따라, 인간-로봇 간의 상호작용을 통한 감성적 교감이 가능한 형태의 연구가 더욱 필요하며, 사용자에게 지속적인 흥미를 주기 위한 감성 표현에 관한 연구도 병행되어야 할 것이다.

3) 곁에서 돌보아주는 로봇

4.2 로봇에 감성음성 기술을 적용한 특허

최근 센서기술의 발달로 사람의 감성(기쁨, 슬픔, 화남, 놀람, 공포, 혐오 등)을 인식하는 감성 인식 기술이 부상하고 있다. 이러한 감성 인식 기술이 적용된 모바일 기기는 사용자의 마음을 스스로 판단하여 사용자가 우울하다고 여겨지면 기분 전환용 음악을 전송할 수 있다.

특허청은 2014년 보도자료[44]에서 음성, 표정, 생체 데이터를 통하여 사람의 감성을 인식하는 모바일 기기의 출원이 매우 증가하는 추세에 있다고 밝혔다. 특허청에 따르면, 2008년까지 총 43건에 불과하던 특허 출원은 2009년 이후 2014년 10월까지 총 105건이 출원되어 지속적인 증가세를 이어가고 있다. 감성을 인식하기 위한 센서종류별로는 복수의 센서를 사용하는 경우가 가장 많았고, 오디오 센서(마이크로폰), 이미지센서(카메라)가 그 뒤를 이었다. 미국의 정보 기술 자문기관인 가트너에 의하면 현재 감성 컴퓨팅 기술은 태동기를 지나 5년에서 10년 이내에 기술 성숙기에 도달하리라고 예상한다.

4.2.1 특허 분석대상 및 전체 특허 동향

인간-로봇 상호작용(HRI)기술은 다양한 의사소통 채널을 통해 인간과 로봇이라는 두 개체 간의 상호작용 및 의사소통 연결 고리를 형성하는 기술로서, 음성 인식, 동작 인식, 촉각, 힘 인식 및 감성 인식 등 로봇이 인간의 의사표시를 인식하기 위한 인식 기술과, 먼 곳에 있는 로봇의 동작을 제어하기 위한 원격조작을 위한 인터페이스 기술 및 인간과 로봇 간의 암묵적, 쌍방향적 의사소통을 위한 인지 및 감정 상호작용기술로 크게 구분할 수 있다[45]. 따라서 이들 세 가지 기술을 분석 대상인 특허의 대분류로 정하고 각 대분류에 대한 관련 하위기술을 분류하여 본 과제의 기술분류체계를 작성하였다.

<표 4-1>에서 확인할 수 있듯이 음성 인식 기술이 가장 많은 출원 건수를 보인다. 전체적으로 살펴볼 때, 인간-로봇 상호 작용기술 관련 특허는 현재 미국과 일본이 기술개발을 선도하고 있으며 1990년대 말부터 한국 및 유럽의 여러 국가도 관련 기술개발에 박차를 가하고 있다.

주요 기술적 결과물로는 우선 1999년 Sony사가 개발, 시장에 출시한 애완용 로봇 AIBO를 들 수 있다. 이후, 2000년 Honda사가 인간형 로봇 ASIMO를 선보이는 등, 전체적으로

일본의 관련 출원의 현격히 증가하였고, 한국에서도 2002년에 KAIST에서 KHR-1 및 KHR-2를 개발한 것을 그 시발점으로 하여, 2004년에는 음성 인식 및 합성 기능 등을 가진 인간형 로봇 HUBO를 개발하여 선보임으로써 로봇 관련 기술개발에 박차를 가하고 있으며, 그 결과가 2000년대에 들어서 국내에서의 관련 출원 증가로 나타나고 있다.

<표 4-1> 휴먼-로봇 인터페이스의 기술분류체계 및 분석 건수

대분류	중분류	소 계
인식기술	음성 인식	159
	얼굴 인식	57
	제스처 인식	42
	촉감/힘 인식	52
	감정 인식	40
	소 계	350
원격조작을 위한 인터페이스	매개 인터페이스	271
	힘 반향 원격조종 장치	110
	힘 반향 제어 및 통신	210
	정보표현 및 공유	233
	소 계	824
인지 및 감정 상호작용기술	사용자 의도인식 및 대응기술	231
	감정생성 및 표현기술	181
	소 계	412
합 계		1,586

80년대 중반의 기술의 태동기, 90년대 중반까지 기술의 발전기를 지나 90년대 후반 이후부터 출원인 수와 출원 건수가 동시에 급격하게 증가하는 기술성장기를 맞고 있으며, 성장 추세는 앞으로도 당분간 계속될 것으로 추정하고 있다.

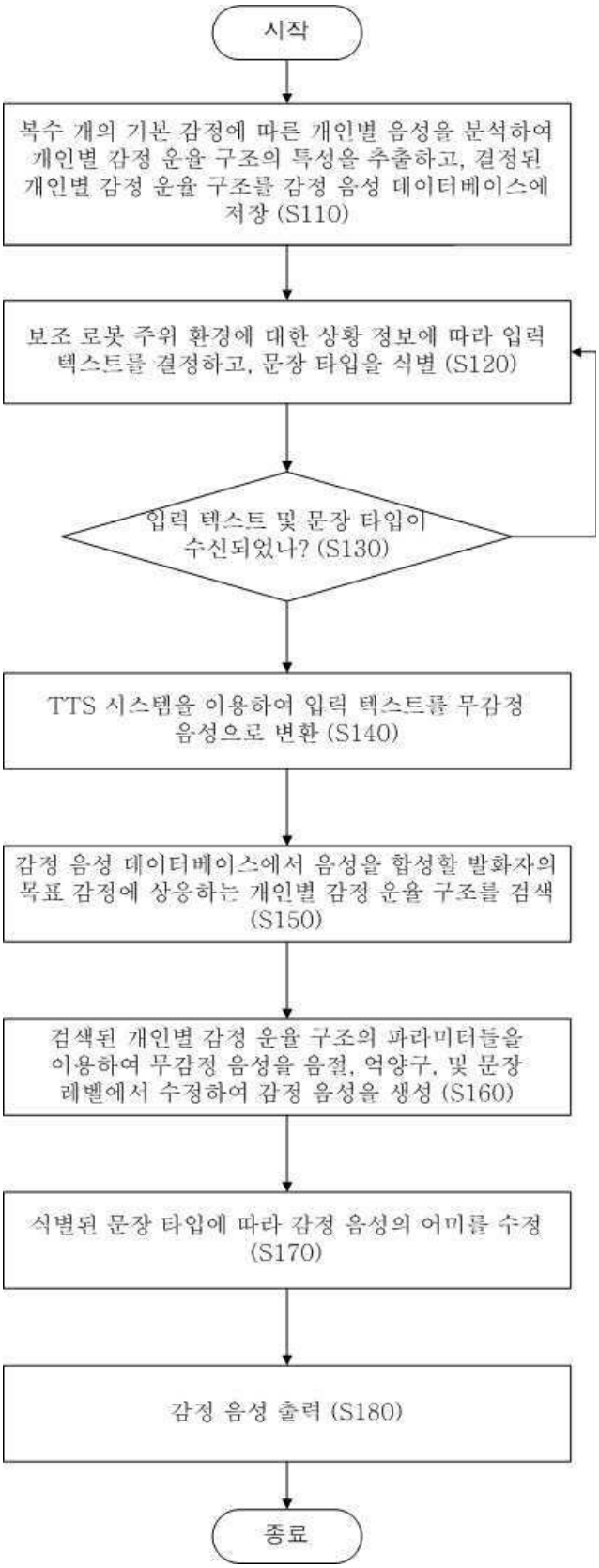
4.2.2 감성음성 기술 기반 특허권

감성음성 기술 관련한 많은 국내·외 특허가 있지만, 한국과학기술원(KAIST)에서 2011년에 출원하여 2013년 등록된 특허 10-1221188, “감정 음성 합성 기능을 가지는 보조 로봇

및 보조 로봇용 감정 음성 합성 방법 및 기록 매체"[46]에 대해 설명하고자 한다.

이 특허권은 현재 등록료 불납으로 국내에서는 그 권리가 소멸하였으나 감성음성 관련하여 그 특징을 상세히 기술하고 있어 본 보고서에서 소개하기로 한다. 이 특허의 명세서에는 인간과 상호작용하는 보조 로봇에서 출력할 감정 음성을 개인 운율 모델에 기반하여 합성하는 방법 및 이러한 보조 로봇이 개시된다. 본 발명에 따른 방법은, 개인별 음성을 분석하여 개인별 감정 운율 구조(personal emotional prosody structure)의 특성을 추출하는 감정 운율 구조 특성 추출하고, 추출된 개인별 감정 운율 구조를 감정 음성 데이터베이스에 저장하는 단계, 보조 로봇의 주위 환경에 대한 상황 정보를 수신하고, 수신된 상황 정보에 따라 입력 텍스트를 결정하며, 결정된 입력 텍스트의 문장 타입을 식별하는 단계, 입력 텍스트 및 목표 감정을 수신하는 수신 단계, 감정 음성 데이터베이스로부터 음성을 합성할 발화자(speaker)에 상응하는 개인별 감정 운율 구조를 검색하는 단계, 및 입력 텍스트를 무감정 음성(emotionless speech)으로 변환하고, 변환된 감정이 없는 음성을 목표 감정에 상응하는 개인별 감정 운율 구조에 기반하여 수정함으로써 발화자에 상응하는 감정 음성을 생성하며, 생성된 감정 음성의 문장 어미(sentence final syllable)를 식별된 문장 타입에 따라서 수정하여 출력하는 감정 음성 합성 단계를 포함한다.

본 발명의 목적은 상황 정보에 따라서 감정 음성을 합성하고, 합성된 감정 음성을 문장 타입에 따라서 수정함으로써 인간과 보조 로봇 간의 상호작용을 향상할 수 있는 보조 로봇을 제공하는 것으로 인간과 상호작용하는 보조 로봇(assistive robot)에서 출력할 감정 음성을 개인 운율 모델에 기반하여 합성하려는 방법에 관한 것이다. 개인별 음성을 분석하여 개인별 감정 운율 구조(personal emotional prosody structure)의 특성을 추출하는 감정 운율 구조의 특성을 추출하고, 추출된 개인별 감정 운율 구조를 감정 음성 데이터베이스에 저장하는 단계, 보조 로봇의 주위 환경에 대한 상황 정보를 수신하고, 수신된 상황 정보에 따라 입력 텍스트를 결정하며, 결정된 입력 텍스트의 문장 타입을 식별하는 단계, 입력 텍스트 및 목표 감정을 수신하는 수신 단계, 감정 음성 데이터베이스로부터 음성을 합성할 발화자(speaker)에 상응하는 개인별 감정 운율 구조를 검색하는 단계, 및 입력 텍스트를 무감정 음성(emotionless speech)으로 변환하고, 변환된 무감정 음성을 목표 감정에 상응하는 개인별 감정 운율 구조에 기반하여 수정함으로써 발화자에 상응하는 감정 음성을 생성하며, 생성된 감정 음성의 문장 어미(sentence final syllable)를 식별된 문장 타입에 따라서 수정하여 출력하는 감정 음성 합성 단계를 포함한다.



<그림 4-1> 보조 로봇용 감정 음성 합성 방법 흐름도

<그림 4-1>은 본 특허 명세서의 도면 1로 기재된 보조 로봇용 감정 음성 합성 방법 흐름도이다. 특히, 감정 운율 구조 특성 추출 단계는, 발화자의 기본 감정(basic emotion)에 따른 음성정보를 포함하는 데이터로부터 일반 감정 운율 구조를 추출하는 일반 감정 운율 구조 추출 단계, 및 각 발화자의 개인별 감정 운율 구조를 일반 감정 운율 구조와 비교하여 개인별 감정 운율 구조의 상대적 차분치를 파라미터화하는 개인별 감정 운율 구조 추출 단계를 포함한다. 또한, 개인별 감정 운율 구조 추출 단계는, 발화자의 각각의 감정에 따른 음성의 전체 피치(overall pitch), 강도(intensity), 및 발화 속도(speech rate)를 파라미터로서 문장 레벨에서 분석하는 문장 레벨 감정 운율 구조 분석 단계, 발화자의 각각의 감정에 따른 음성에 포함되는 억양구(IP, Intonation Phrase)들 간의 휴지 길이(pause length)를 파라미터로서 억양구 레벨에서 분석하는 억양구 레벨 감정 운율 구조 분석 단계, 및 발화자의 각각의 감정에 따른 음성에 포함되는 억양구들 각각의 억양구 경계 패턴(IP boundary pattern)을 파라미터로서 음절 레벨에서 분석하는 음절 레벨 감정 운율 구조 분석 단계를 포함한다.

더 나아가, 문장 레벨 감정 운율 구조 분석 단계는, 감정별 음성의 피치 값의 사분위 간 평균(IQM, InterQuartile Mean)을 연산하는 단계, 감정별 음성의 강도 중 소정의 값 이상의 강도를 선택하는 단계, 감정별 음성의 전체 발화 길이로부터 발화 속도를 연산하는 단계, 피치 값, 강도 및 발화 속도를 정규화하여 문장별 불일치(disparity) 및 발화자별 불일치를 제거하는 단계, 및 정규화된 결과를 이용하여, 중립의 감정 상태에 상응하는 개인별 감정 운율 구조를 표준으로 하여 감정별 개인별 감정 운율의 파라미터들의 차이를 연산하고, 연산 결과를 이용하여 개인별 감정 운율 구조를 구성하는 단계를 포함한다.

특히, 억양구 레벨 감정 운율 구조 분석 단계는, 감정별 음성의 억양구 간 휴지 영역(pause region)을 검출하는 단계, 및 휴지 영역들의 전체 길이를 합산하여 전체 휴지 길이를 연산하는 단계를 포함하고, 음절 레벨 감정 운율 구조 분석 단계는, 음성의 억양구 경계 패턴을 L%, H%, LH%, HL%, LHL%, HLH%, HLHL%, LHLH% 및 LHLHL% 중 하나에 상응하는 피치 컨투어(pitch contour)로서 분석하는 단계를 포함한다. 또한, 감정 음성 합성 단계는 TTS(Text-to-Speech) 시스템을 이용하여 입력 텍스트를 무감정 음성으로 변환하는 단계, 감정 음성 데이터베이스로부터 발화자의 목표 감정에 상응하는 감정 운율 구조를 검색하는 단계, 검색된 감정 운율 구조의 파라미터들을 이용하여 무감정 음성을 수정함으로써 감정 음성을 생성하는 음성 수정 단계, 및 생성된 감정 음성의 어미를 문장 타입에 따라서

수정하는 어미 수정 단계를 포함하고, 음성 수정 단계는 개인별 감정 운율 구조로부터 목표 감정에 상응하는 피치 컨투어를 파라미터로서 추출하는 단계, 및 추출된 피치 컨투어를 이용하여 무감정 음성의 피치 컨투어를 수정하는 음절 레벨 수정 단계를 포함한다.

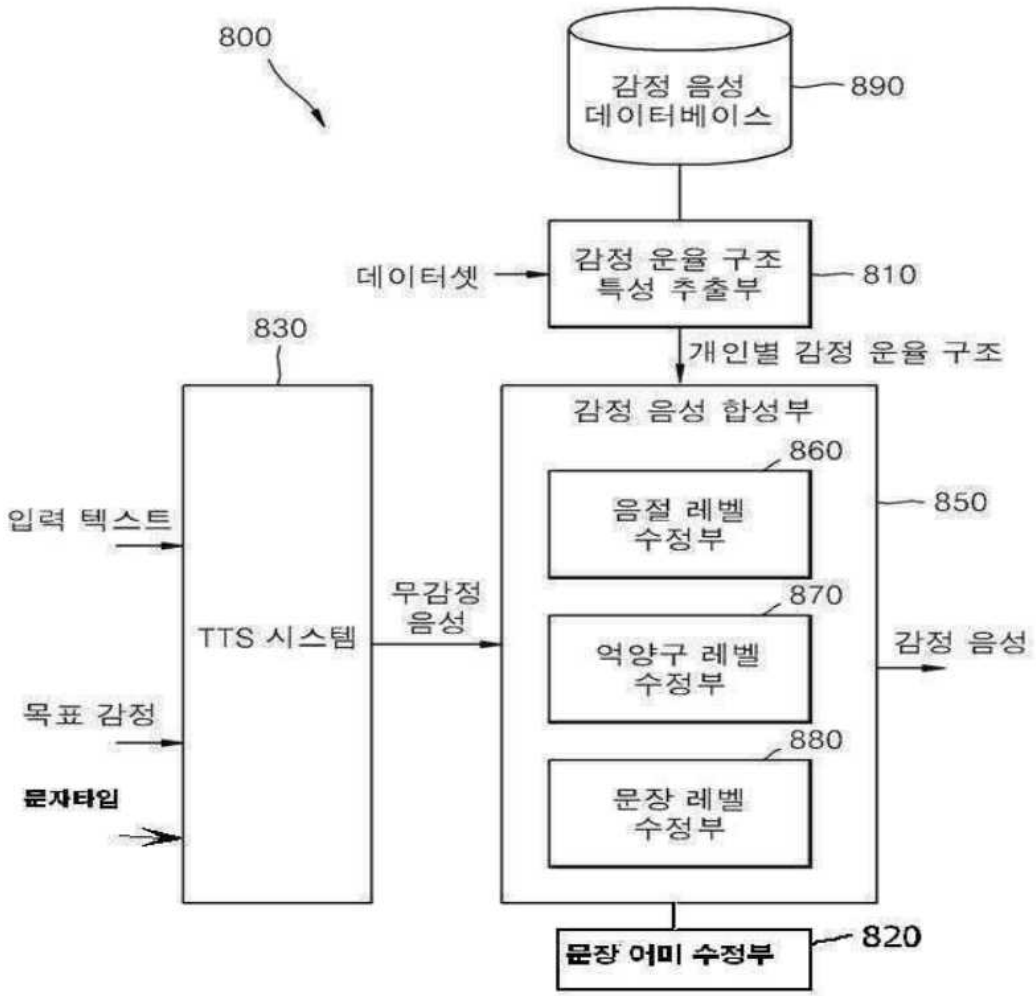
더 나아가, 음성 수정 단계는 개인별 감정 운율 구조로부터 목표 감정에 상응하는 휴지 길이를 파라미터로서 추출하는 단계, 및 추출된 휴지 길이를 이용하여 무감정 음성의 휴지 길이를 수정하는 억양구 레벨 수정 단계를 포함한다.

또한, 음성 수정 단계는 개인별 감정 운율 구조로부터 목표 감정에 상응하는 전체 피치, 전체 강도, 및 발화 속도를 파라미터로서 추출하는 단계, 및 추출된 전체 피치, 전체 강도 및 발화 속도를 이용하여 무감정 음성의 전체 피치, 전체 강도, 및 발화 속도를 수정하는 문장 레벨 수정 단계를 포함하고, 어미 수정 단계는 감정 음성의 억양이 L%, LH%, LHL%, LHLH%, LHLHL%, H%, HL%, HLH% 및 HLHL% 중 하나가 되도록 수정하고, 수정된 억양에 따라서 감정 음성의 어미를 수정하는 단계를 포함한다.

상기와 같은 목적들을 달성하기 위한 본 발명의 다른 면은, 본 발명의 일면에 따른 방법을 구현하기 위한 컴퓨터에 의하여 실행될 수 있는 명령어들을 포함하는 컴퓨터 프로그램이 기록된 컴퓨터에 의하여 독출될 수 있는 기록 매체에 관한 것이다. 상기와 같은 목적들을 달성하기 위한 본 발명의 또 다른 면은, 인간과 상호작용하며, 상황 정보에 따른 감정 음성을 개인 운율 모델에 기반하여 감정 음성을 합성하기 위한 보조 로봇에 관한 것이다.

<그림 4-2>는 감정 음성 합성 기능을 가지는 보조 로봇의 블록도이다. 본 발명에 따른 보조 로봇은 개인별 음성을 분석하여 개인별 감정 운율 구조의 특성을 추출하는 감정 운율 구조 특성 추출부, 추출된 개인별 감정 운율 구조를 저장하는 감정 음성 데이터베이스, 보조 로봇의 주위 환경에 대한 상황 정보를 수신하기 위한 상황 정보 수신부, 수신된 상황 정보에 따라 입력 텍스트를 결정하며, 결정된 입력 텍스트의 문장 타입을 식별하는 문장 타입 식별부, 및 입력 텍스트 및 목표 감정이 수신되면 입력 텍스트를 무감정 음성으로 변환하고, 감정 음성 데이터베이스로부터 음성을 합성할 발화자에 상응하는 개인별 감정 운율 구조를 검색하며, 변환된 무감정 음성을 목표 감정에 상응하는 개인별 감정 운율 구조에 기반하여 수정함으로써 발화자에 상응하는 감정 음성을 생성하고, 생성된 감정 음성의 문장 어미를 식별된 문장 타입에 따라서 수정하여 출력하는 감정 음성 합성부를 포함한다. 특히, 감정 운율 구조 특성 추출부는 발화자의 기본 감정에 따른 음성정보를 포함하는 데이터셋으로부터 일반 감정 운율 구조를 추출하는 동작, 및 각 발화자의 개인별 감정 운율

구조를 일반 감정 운율 구조와 비교하여 개인별 감정 운율 구조의 상대적 차분치를 파라미터화하는 동작을 수행하도록 적응되고, 감정 운율 구조 특성 추출부는, 개인별 감정 운율 구조를 추출하기 위하여, 발화자의 각각의 감정에 따른 음성의 전체 피치, 강도, 및 발화 속도를 파라미터로서 문장 레벨에서 분석하는 문장 레벨 감정 운율 구조 분석 동작, 발화자의 각각의 감정에 따른 음성에 포함되는 억양구(IP)들 간의 휴지 길이를 파라미터로서 억양구 레벨에서 분석하는 억양구 레벨 감정 운율 구조 분석 동작 및 발화자의 각각의 감정에 따른 음성에 포함되는 억양구들 각각의 억양구 경계 패턴을 파라미터로서 음절 레벨에서 분석하는 음절 레벨 감정 운율 구조 분석 동작을 수행하도록 적응된다.



<그림 4-2> 감성 음성 합성 기능을 가지는 보조 로봇의 블록도

특히, 감정 운율 구조 특성 추출부는, 문장 레벨 감정 운율 구조를 분석하기 위하여, 감

정별 음성의 피치값의 사분위간 평균(IQM)을 연산하는 동작, 감정별 음성의 강도 중 소정값 이상의 강도를 선택하는 동작, 감정별 음성의 전체 발화 길이로부터 발화 속도를 연산하는 동작, 피치값, 강도 및 발화 속도를 정규화하여 문장별 불일치 및 발화자별 불일치를 제거하는 동작, 및 정규화된 결과를 이용하여, 중립의 감정 상태에 상응하는 개인별 감정 운율 구조를 표준으로 하여 감정별 개인별 감정 운율의 파라미터들의 차이를 연산하고, 연산 결과를 이용하여 개인별 감정 운율 구조를 구성하는 동작을 수행하도록 적응된다. 더 나아가, 감정 운율 구조 특성 추출부는, 억양구 레벨 감정 운율 구조를 분석하기 위하여, 감정별 음성의 억양 구간 휴지 영역을 검출하는 동작, 및 휴지 영역들의 전체 길이를 합산하여 전체 휴지 길이를 연산하는 동작을 수행하도록 적응되고, 감정 운율 구조 특성 추출부는, 음절 레벨 감정 운율 구조를 분석하기 위하여, 음성의 억양구 경계 패턴을 L%, H%, LH%, HL%, LHL%, HLH%, HLHL%, LHLH% 및 LHLHL% 중 하나에 상응하는 피치 컨투어로서 분석하는 동작을 수행하도록 적응된다.

특히, 감정 음성 합성부는 TTS 시스템을 이용하여 입력 텍스트를 무감정 음성으로 변환하는 동작, 감정 음성 데이터베이스로부터 발화자의 목표 감정에 상응하는 감정 운율 구조를 검색하는 동작, 및 검색된 감정 운율 구조의 파라미터들을 이용하여 무감정 음성을 수정함으로써 감정 음성을 생성하는 음성 수정 동작, 및 생성된 감정 음성의 어미를 문장 타입에 따라서 수정하는 어미 수정 동작을 수행하도록 적응된다.

또한, 감정 음성 합성부는, 음성 수정 동작을 수행하기 위하여, 개인별 감정 운율 구조로부터 목표 감정에 상응하는 피치 컨투어를 파라미터로서 추출하는 동작, 및 추출된 피치 컨투어를 이용하여 무감정 음성의 피치 컨투어를 수정하는 음절 레벨 수정 동작을 수행하도록 적응되고, 음성 수정 동작을 수행하기 위하여, 개인별 감정 운율 구조로부터 목표 감정에 상응하는 휴지 길이를 파라미터로서 추출하는 동작, 및 추출된 휴지 길이를 이용하여 무감정 음성의 휴지 길이를 수정하는 억양구 레벨 수정 동작을 수행하도록 적응된다. 더 나아가, 감정 음성 합성부는, 음성 수정 동작을 수행하기 위하여, 개인별 감정 운율 구조로부터 목표 감정에 상응하는 전체 피치, 전체 강도 및 발화 속도를 파라미터로서 추출하는 동작, 및 추출된 전체 피치, 전체 강도, 및 발화 속도를 이용하여 무감정 음성의 전체 피치, 전체 강도, 및 발화 속도를 수정하는 문장 레벨 수정 동작을 수행하도록 적응되고, 어미 수정 동작을 수행하기 위하여, 감정 음성의 억양이 L%, LH%, LHL%, LHLH%, LHLHL%, H%, HL%, HLH% 및 HLHL% 중 하나가 되도록 수정하고, 수정된 억양에 따라

서 감정 음성의 어미를 수정하는 동작을 수행하도록 적응된다.

제5장 맺음말

감성 ICT는 궁극적으로 개인의 감성 정보를 기반으로 한 제품 및 서비스를 제공하는 것을 목표로 한다. 그 결과는 인간 사이의 커뮤니케이션 증진, 개인의 이해, 가족관계 증진, 여론 및 정책형성, 사회통합 증진 등 인공적 생산물, 제품과의 커뮤니케이션에 의한 애착형성 및 선택, 구매 행동의 증진 등으로 나타날 것이다. 즉, 감성 ICT는 일상의 전 산업에 융합 적용될 것이며 미래 우리 삶의 질을 책임질 기술이다.

또한, 개인의 정보를 다루는 일이라 민감한 부분이기도 하므로 개인의 사생활 보호 문제 해결을 위한 기술적 부분과 법제도 정비가 필요한 부분에 대해서는 세심한 검토와 접근이 요구된다. 더 나아가, 현재의 플랫폼에 감성 정보 처리기능을 개발, 적용하기 위해서는 감성 정보의 규격 및 나아가 표준화가 국가 주도로 이루어져야 하며 국제 표준화에도 앞장서야 할 것이다.

무엇보다도 감성 인식 기술은 진정한 인간-컴퓨터 또는 인간-로봇 간 상호작용을 위한 핵심 요소기술로 정확한 감성 인식을 위해서는 감성 인식의 재료가 되는 사람 얼굴의 표정(영상), 목소리(음성), EEG나 맥박과 같은 생체 신호, 이 모든 데이터가 함께 이용되는 것이 바람직하다. 그러기 위해서는 소위 멀티모달 센싱 및 딥러닝 기술이 꾸준히 발전되어야 한다.

아울러 감성 관련 데이터베이스의 구축도 매우 중요하다. 일반적인 객체 인식 데이터베이스에 비해 감성 인식 데이터베이스는 턱없이 부족한 것이 현실이다. 기술발전을 위해서는 이에 대한 투자와 연구가 절실하다. 마지막으로 감성 인식 기술은 정상인뿐만 아니라 자폐나 우울증 같은 정신적 질환이 있는 분들에게 진찰은 물론 치료 목적으로 매우 유용하게 사용될 것으로 보인다.

참고문헌

[1] 권순일, 손귀영, 김경인, 박능수, 이석필. (2019). AI 스피커를 위한 음성기반 감성 인식 기술. 전기의세계, 68(10), pp.22-27

[2] 송병철, 김대하, 최동윤, 이민규. (2018.10.17). 감정 인식 기술 동향. 주간기술동향 1868호[온라인].
주소: https://www.itfind.or.kr/publication/regular/weeklytrend/pastList/read.do?selectedId=1049&selectedCategory=B ITA 01&pageSize=10&pageIndex=0

[3] 방재훈, 이승룡. (2014). 감성기반 서비스를 위한 통화 음성 감성 인식 기법, 정보과학회논문지, 41(3), pp.208-213.

[4] A. B. Kandali, A. Routray, T. K. Basu, Emotion recognition from Assamese speeches using MFCC features and GMM classifier, IEEE Region 10 Conference(TENCON), Nov, 2008, pp.1-5, 19-21.

[5] Z. Xiao, Dellandrea, L. Chen, W. Dou, Recognition of emotions in speech by a hierarchical approach, ACII 2009, 2009, pp.401-408.

[6] Y. Cho, K. S. Park, A Study on The Improvement of Emotion Recognition by Gender Discrimination, Journal of IEEK, vol.45, 2008, pp.401-408.

[7] 이지원 외, 다중 작업 기반의 합성곱 신경망을 이용한 음성 감성 인식, 2017년 한국통신학회 하계종합학술대회, 2017년 6월.

[8] 강소연, 최욱, Random forest를 이용한 음성신호 기반 감성 인식, 2017년 한국통신학회 동계 종합학술발표회, 2017년.

[9] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In Ninth European Conference on Speech Communication and Technology.

[10] Engberg, I. S., Hansen, A. V., Andersen, O., & Dalsgaard, P. (1997). Design, recording and verification of a Danish emotional speech database. In Fifth European Conference on Speech Communication and Technology

[11] Koolagudi, S. G., Reddy, R., Yadav, J., & Rao, K. S. (2011). IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In 2011 International conference on devices and communications (ICDeCom) (pp. 1-5). IEEE.

[12] Busso, Carlos & Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. pp.335-359. 10.1007/s10579-008-9076-6.

[13] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1), pp.235-238.

[14] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), pp.90-99.

[15] D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, July 2005, pp.1500−1503.

[16] H. Hu, M.-X. Xu, W. Wu, Fusion of global statistical and segmental spectral features for speech emotion recognition, in: International Speech Communication Association—8th Annual Conference of the International Speech Communication Association, Interspeech 2007, vol. 2, 2007, pp.1013−1016.

[17] Ayadi, Moataz & Kamel, Mohamed S. & Karray, Fakhri. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 44. pp.572-587. 10.1016/j.patcog.2010.09.020.

[18] C. Williams, K. Stevens. (1981). Vocal correlates of emotional states, Speech Evaluation in Psychiatry, Grune and Stratton, pp.189–220.

[19] Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. Speech communication, 40(1-2), pp.189-212.

[20] L. Rabiner, R. Schafer. (1978). Digital Processing of Speech Signals, first edition, Pearson Education.

[21] Deller Jr, J. R., Proakis, J. G., & Hansen, J. H. (1993). Discrete time processing of speech signals. Prentice Hall PTR.

[22] Motion Capture (IEMOCAP) Database
주소: https://sail.usc.edu/iemocap/

[23] Choi, Y. H., Ban, S. M., Kim, K. W., & Kim, H. S. (2015). Evaluation of Frequency Warping Based Features and Spectro-Temporal Features for Speaker Recognition. Phonetics and Speech Sciences, 7(1), pp.3-10.

[24] Teagar, H. M., & Teagar, S. M. (1990). Evidence for Nonlinear Production Mechanisms in the Vocal Tract. In Speech Production and Modelling. Kluwer.

[25] Teager, H. (1980). Some observations on oral air flow during phonation. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(5), pp.599-601.

[26] Kaiser, J. F. (1990, April). On a simple algorithm to calculate the'energy'of a signal. In International conference on acoustics, speech, and signal processing (pp. 381-384). IEEE.

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), pp.273-297.

[28] Heather Mack. Emotion-detecting voice analytics company Beyond Verbal raises $3.3M. mobihealthnews. September 01, 2016.
주소:https://www.mobihealthnews.com/content/emotion-detecting-voice-analytics-company-beyond-verbal-raises-33m

[29] 김문구, 박종현. (2018). AI 기반 감성증강 10대 유망 서비스 탐색. 대전: 한국전자통신연구원 미래전략연구소 기술경제연구본부

[30] audEERING 공식 홈페이지
주소: https://www.audeering.com/

[31] 감성을 읽는 인공지능, 조나단 AI
주소: https://blog.naver.com/skaibril/221397337775

[32] Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2), pp.1097-1108.

[33] Batliner, A., Burkhardt, F., Van Ballegooy, M., & Nöth, E. (2006). A taxonomy of applications that utilize emotional awareness. Proceedings of IS-LTC, pp. 246-250.

[34] examples of emotion-aware applications
주소: http://emotionalapplications.syntheticspeech.de/

[35] Marg, E. (1995). DESCARTES'ERROR: emotion, reason, and the human brain. Optometry and Vision Science, 72(11), pp. 847-848.

[36] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., & Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In Fifth ISCA Workshop on Speech Synthesis.

[37] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Sixth European Conference on Speech Communication and Technology.

[38] Burkhardt, F., & Stegmann, J. (2009). Emotional speech synthesis: Applications, history and possible future. Proc. ESSV.

[39] Lee, Y., & Kim, T. (2019, May). Robust and fine-grained prosody control of end-to-end speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5911-5915). IEEE.

[40] 박천수, 류정우, 손주찬. (2007년 4월). 로봇 감성 기술. 전자통신동향분석 제22권 제2호

[41] Martínez, C. Á., & Cruz, A. B. (2005, August). Emotion recognition in non-structured utterances for human-robot interaction. In ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005. (pp. 19-23). IEEE.

[42] Komatani, K., Ito, R., Kawahara, T., & Okuno, H. G. (2004, May). Recognition of emotional states in spoken dialogue with a robot. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 413-423). Springer, Berlin, Heidelberg.

[43] Kanda, T., Iwase, K., Shiomi, M., & Ishiguro, H. (2005, August). A tension-moderating mechanism for promoting speech-based human-robot interaction. In 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 511-516). IEEE.

[44] 김민철. 음성 표정 등 감성 인식 모바일 기술 특허 증가. 세종경제신문 2014년 11월 24일자.
주소: http://www.sejongeconomy.kr/news/articleView.html?idxno=5064

[45] [IT응용] 인간로봇 상호작용(HRI)기술
주소: http://weekly.tta.or.kr/weekly/files/20065806025819 admin.pdf

[46] 박종철, et al. (2011). 감정 음성 합성 기능을 가지는 보조 로봇 및 보조 로봇용 감정 음성 합성 방법 및 기록 매체.

부록 (SSML)