

음성 인터페이스를 활용한 감성 인식과 감성 음성 합성에 대한 기술 동향

목차

- ▷ 제1장 서론
- ▷ 제2장 감성 음성 인식 기술
- ▷ 제3장 감성 음성 합성 기술
- ▷ 제4장 감성 음성기술과 로봇공학
- ▷ 제5장 맺음말





제1장 서론



연구의 배경 및 서술 형식

- 현대인들의 정서적 • 심리적인 문제 + 음성 인터페이스의 편리성

ex) 소셜 네트워크상의 사용 언어, 음성의 미묘한 변화에 대한 감성파악을 통해 우울증을 진단하고 자살을 사전에 방지

- 1.3 음성 공학 용어 정리: <표 1-1> 용어

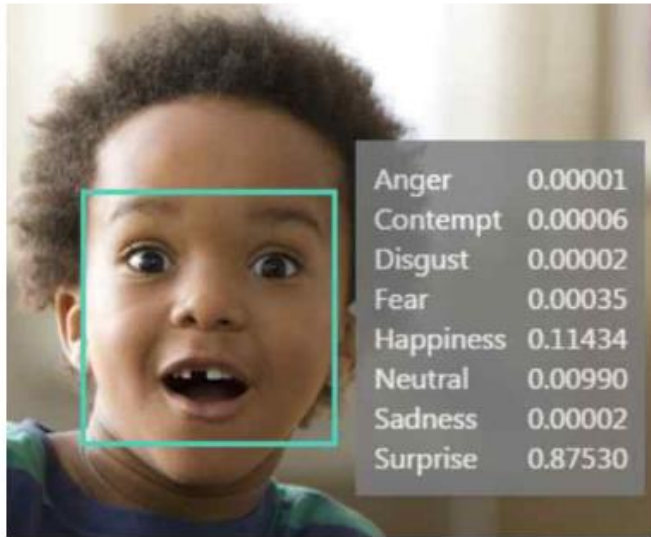
한글 용어	영문 용어	의미
음성	speech	사람의 발음 기관을 통해 내는 구체적이고 물리적인 소리. 발화자와 발화시에 따라 다르게 나는 소리로서 자음과 모음으로 나뉘는 성질이 있다. (유의어) 말소리
음성 인식	speech recognition	사람이 발성한 음성의 의미 내용을 컴퓨터 따위를 사용하여 자동으로 인식하는 것 사람이 말하는 음성 언어를 컴퓨터가 해석해 그 내용을 문자 데이터로 전환하는 처리를 말한다. STT(=Speech To Text)라고도 한다. 키보드 대신 문자를 입력하는 방식으로 주목을 받고 있다. 로봇, 텔레매틱스 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.



제2장 감성 음성 인식 기술



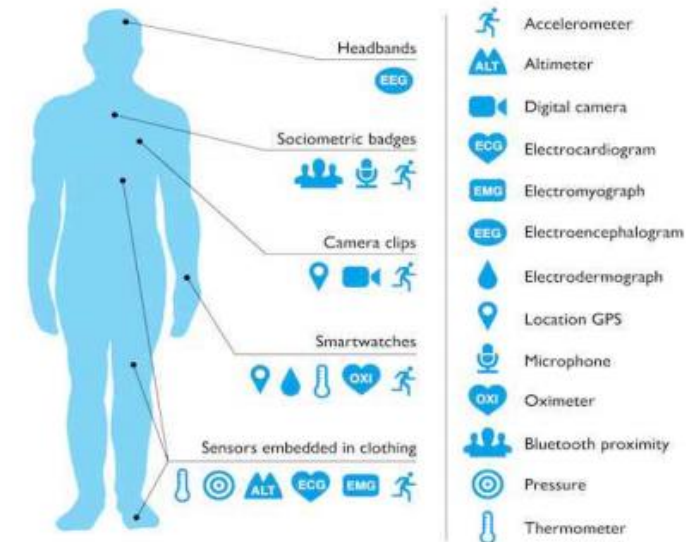
감성 음성 인식의 개요



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



국외 감성 음성 DB

<표 2-1> 감성음성 데이터베이스

데이터베이스	국가	언어	감정범주
Berlin Emoitional 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

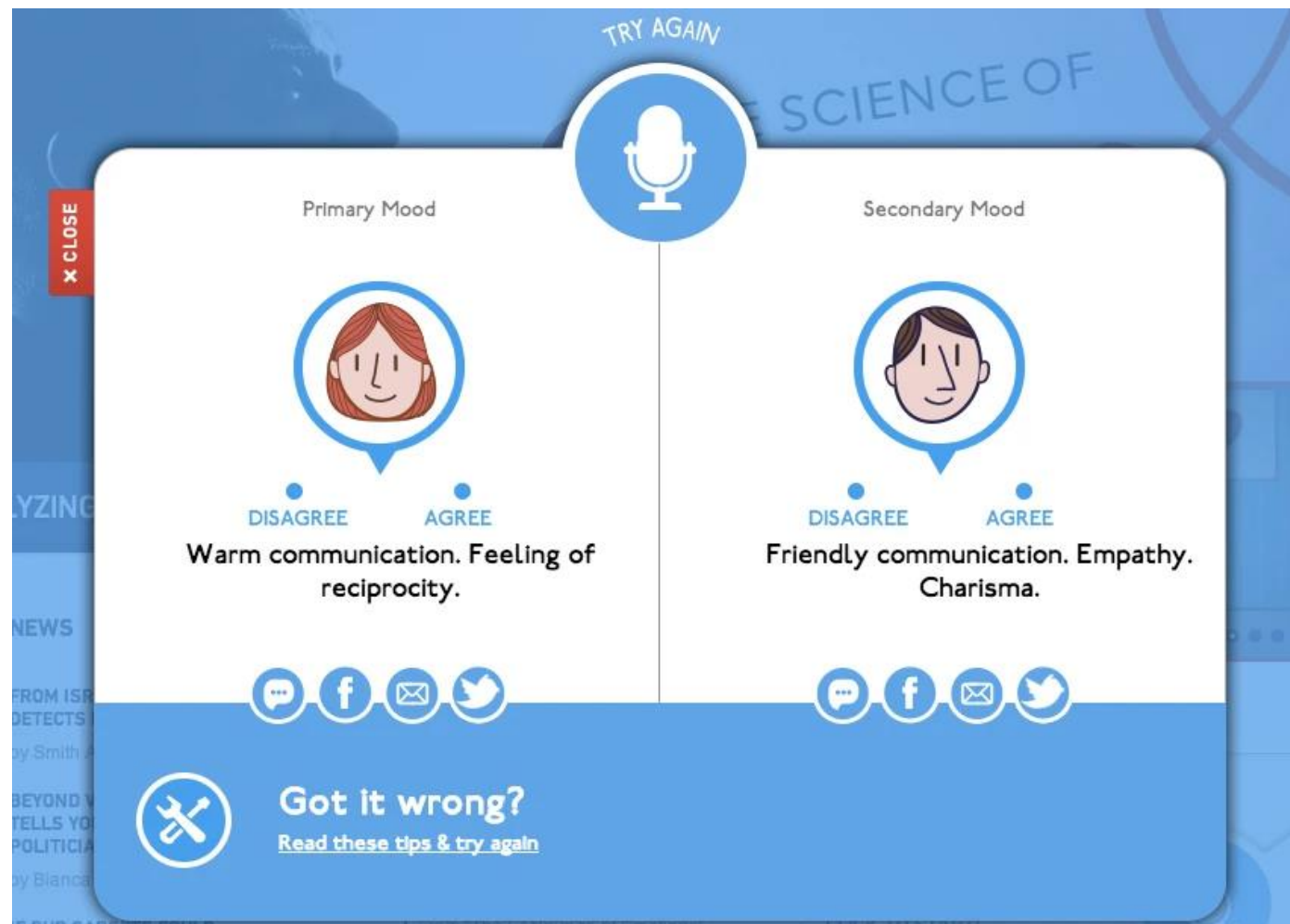


국내 감성 음성 DB

필요성	<ul style="list-style-type: none">- 스마트폰/로봇 등을 기반으로 한 대화형 비서 서비스, 안내 서비스 등 음성 합성에 대한 필요성은 시간이 지날수록 높아지고 있음- 사람과 비슷한 음성을 내기 위해서는 상황과 대화 흐름에 따라, 해당 감정에 알맞는 음성을 합성하는 기술이 필요하나, 동일 인물에 대한 다감정 음성 데이터셋은 공개된 바가 없음
구축내용	<ul style="list-style-type: none">- 30대 여성 성우 1인, 7가지 감정에 대해서 각각 3,000개 발화에 대한 음성 녹음을 수행하였음. 총 21,000개 음성 파일 구축
구조	<ul style="list-style-type: none">- raw 폴더 아래에 acriil_(감정)_(문장번호).raw 파일 존재- 해당 파일은 16bit, mono, 16KHz, PCM format의 음성 파일임- txt 폴더 아래에 acriil_(감정)_(문장번호).txt 파일이 해당 pcm 파일의 텍스트- 실제 발화 내용(발음)에 따라 텍스트가 수정되었으므로 txt 파일은 감정에 따라 상이할 수 있음
활용 예	
관련정보	<ul style="list-style-type: none">- home.iacryl.com:40022- id/pw : flagship/flagship1234- /emoTTS/ directory



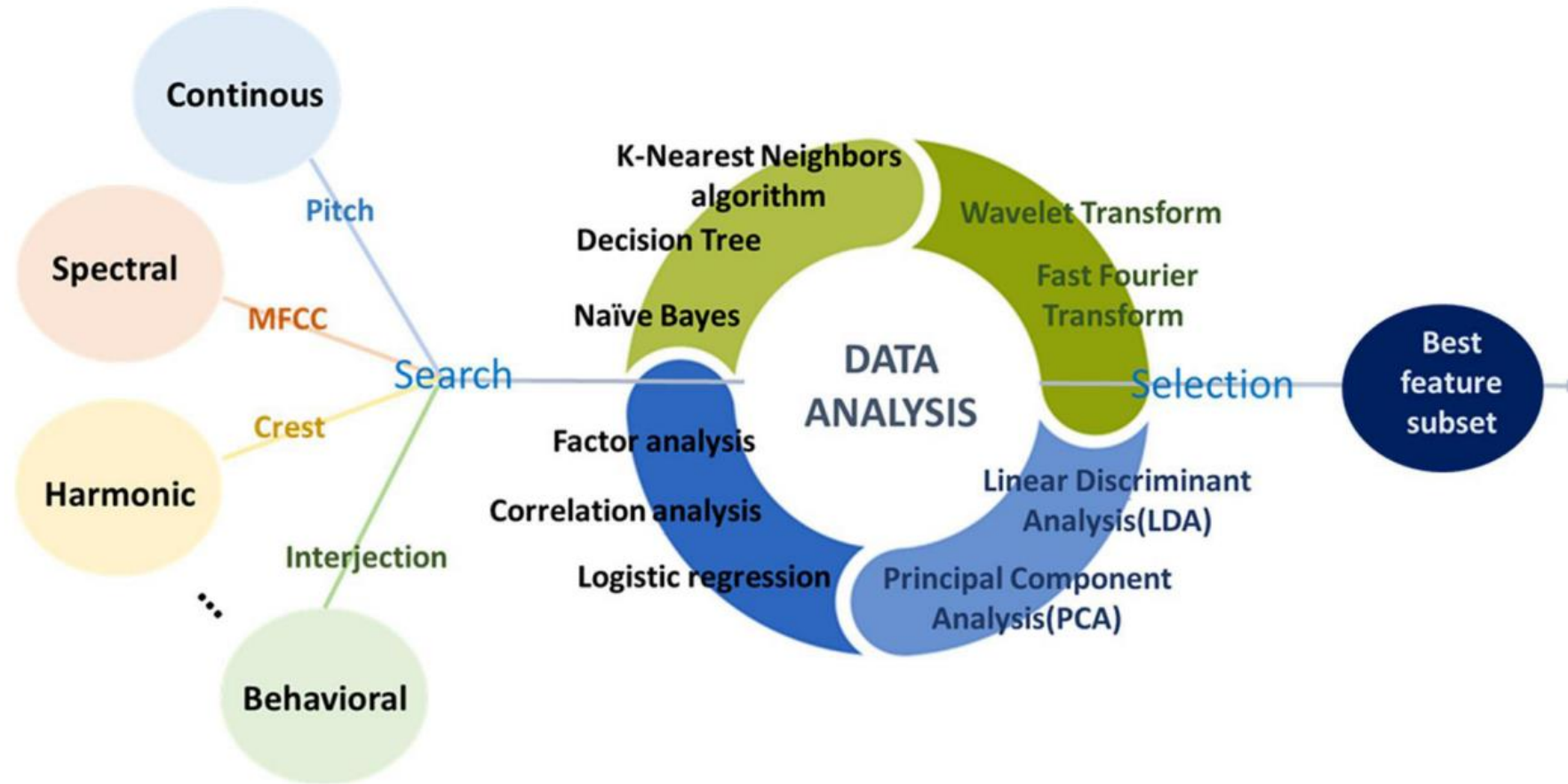
국외 감성음성 인식 기술



<그림 2-6> 비온드 버벌의 앱, 무디스(Moodies)



국내 감성음성 인식 기술



<그림 2-7> 음성기반 최적의 특징요소 선정



국내 감성음성 인식 기술

작동 방식



<그림 2-9> 조나단의 멀티 모달 감성 인식



제3장 감성 음성 합성 기술



학술 논문 번역 – 참고문헌 [38]

EMOTIONAL SPEECH SYNTHESIS: APPLICATIONS, HISTORY AND POSSIBLE FUTURE

Felix Burkhardt, Joachim Stegmann

*Deutsche Telekom Laboratories
Berlin, Germany*

Abstract: Emotional speech synthesis is an important part of the puzzle on the long way to human-like artificial human-machine interaction. During the way, lots of stations like emotional audio messages or believable characters in gaming will be reached. This paper discusses technical aspects of emotional speech synthesis, shows practical applications based on a higher level framework and highlights new developments concerning the realization of affective speech with non-uniform unit selection based synthesis and voice transformation techniques.

1 Introduction

No one ever speaks without emotion. Despite this fact, emotional simulation is not yet a self evident feature in current speech synthesizers. One reason for this lies certainly in the complexity of human vocal expression: current state-of-the-art synthesizers still struggle with the challenge to generate understandable and natural sounding speech, although the latter demand already indicates the importance of affective expression.

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

👉 3.1 ~ 3.5 내용

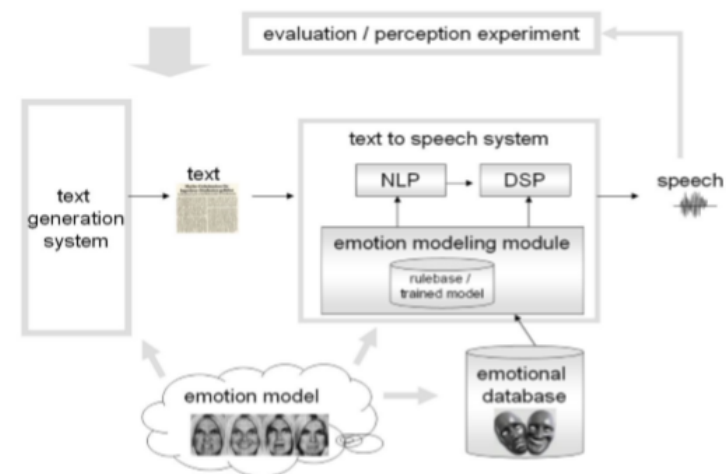
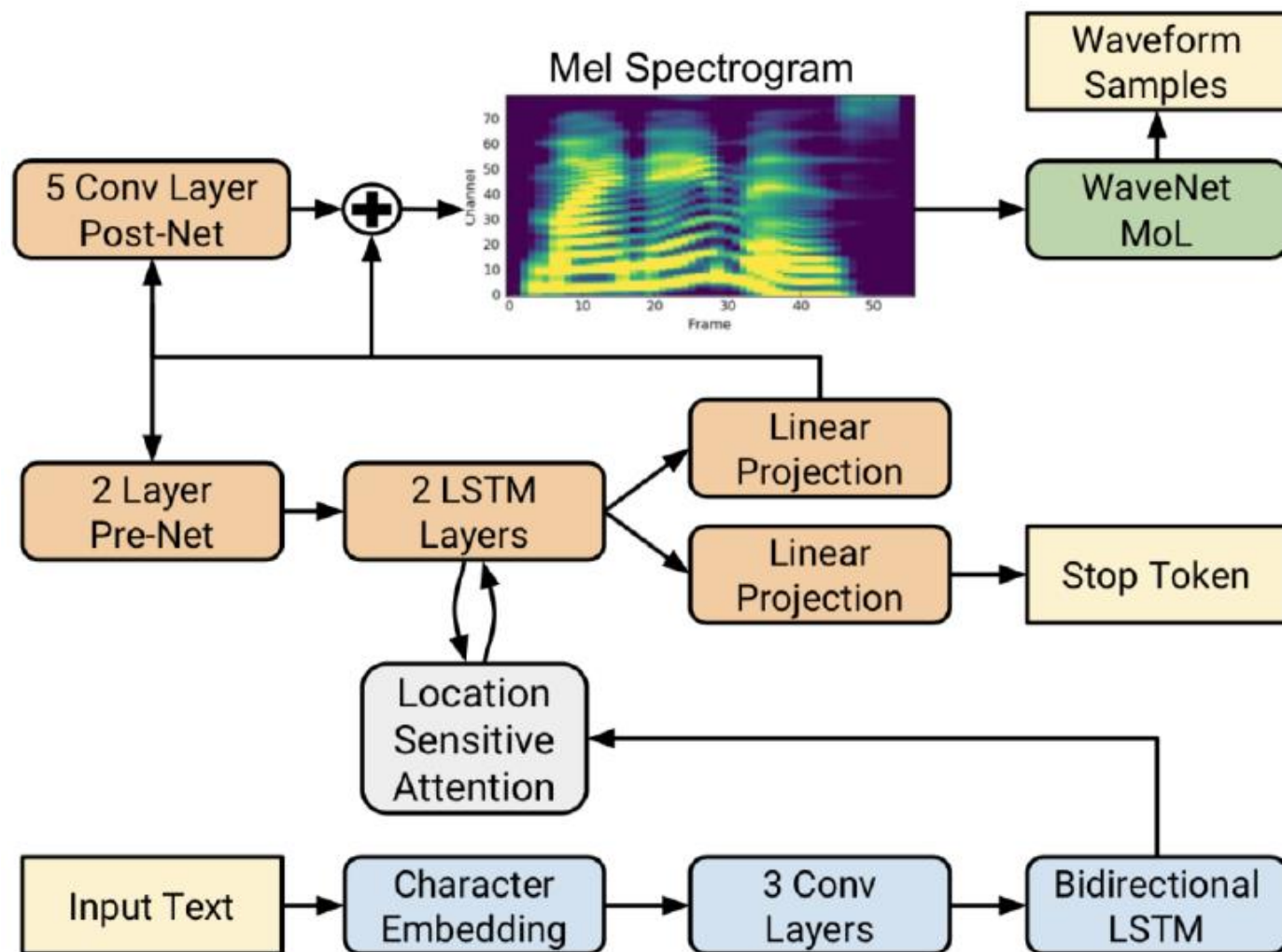


Figure 1 - General architecture of an emotional Text to Speech system.



감성음성합성 모델 - Multi-Speaker Tacotron2



- Multi-Speaker Tacotron2를 감성음성합성에 적용
- 7개의 극성을 가지는 감성음성 데이터를 각각 하나의 화자로 하여 Multi-Speaker Tacotron2로 훈련
- 결과 음성 출력 시에 설정한 Speaker ID는 다음과 같음

0: Angry 1: Disgust 2: Fear
3: Happy 4: Neutrality
5: Sad 6: Surprise



합성문장: 지금 이 순간 한없이 행복하네!



Neutrality



Angry



Disgust



Fear



Happy



Sad

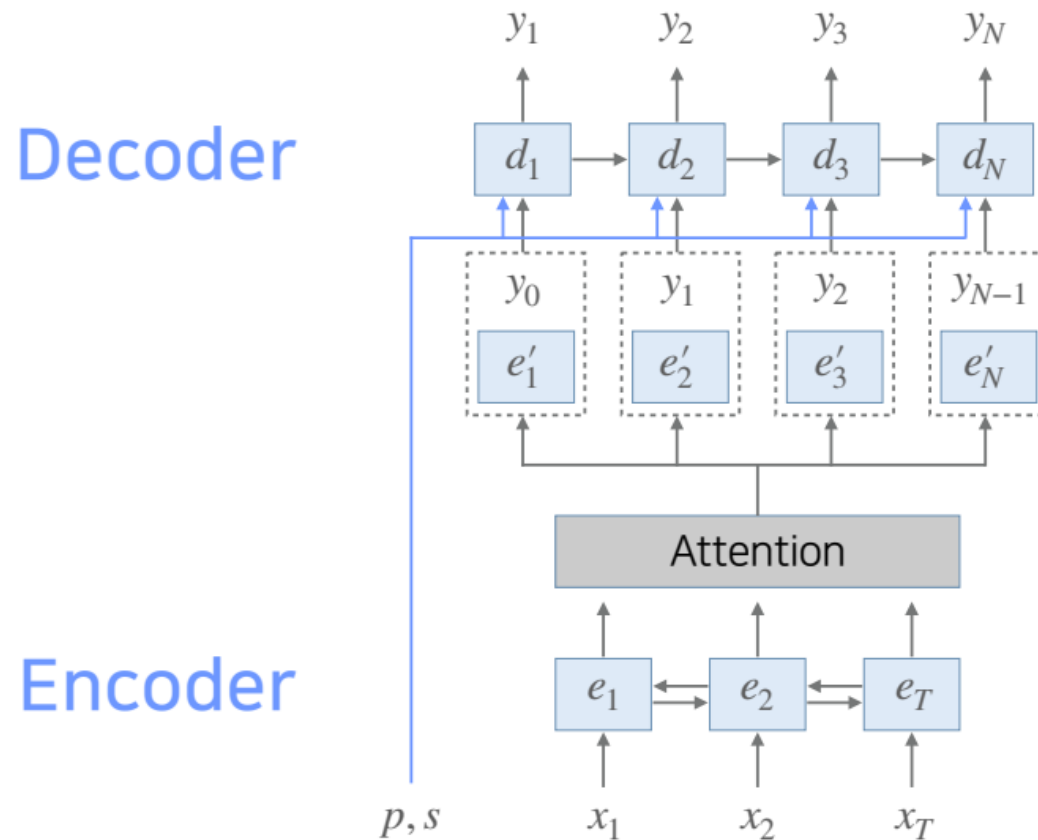


Surprise



3.6 최근 감성 음성 합성 기법

<https://deview.kr/2019/schedule>



x	텍스트 입력
y	음성 출력
p	스타일 정보
s	화자 정보
T	텍스트 길이
N	음성 길이
 	Concatenate

<그림 3-4> 네오사피엔스에서 공개한 감성 음성 합성 방법



부록 - SSML

회사 별 SSML 구문정리 사이트

Google: <https://cloud.google.com/text-to-speech/docs/ssml?hl=ko>

Microsoft: <https://docs.microsoft.com/ko-kr/azure/cognitive-services/speech-service/speech-synthesis-markup>

Amazon: https://docs.aws.amazon.com/ko_kr/polly/latest/dg/ssml.html

Kakao: https://developers.kakao.com/assets/guide/kakao_ssml_guide.pdf

IBM: <https://cloud.ibm.com/docs/services/text-to-speech?topic=text-to-speech-ssml>

29

회사 별 SSML 구문의 특성

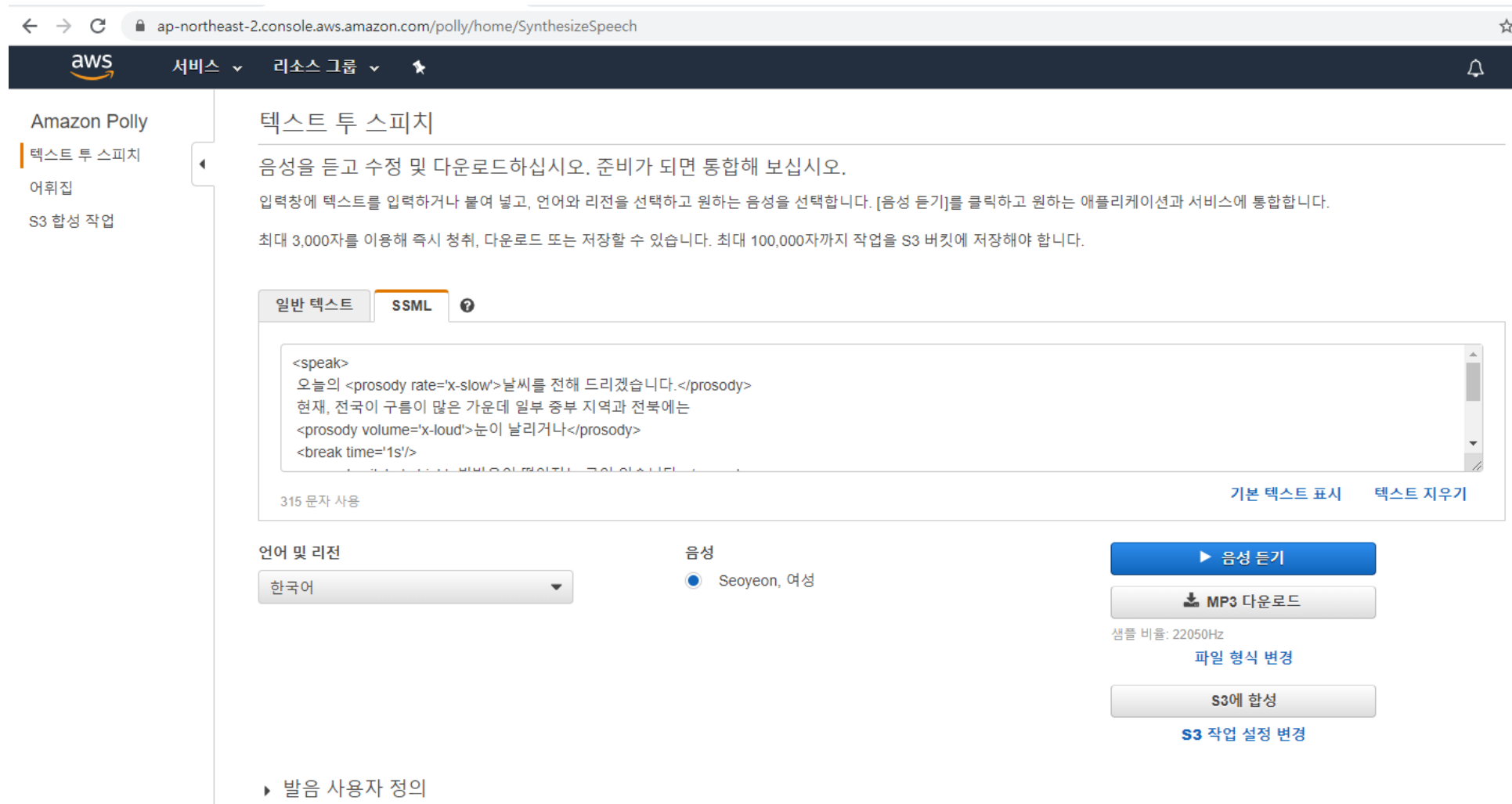
Google	Kakao
<p>1. <par></p> <ul style="list-style-type: none">* 여러 미디어 요소를 한 번에 재생할 수 있게 해주는 병렬 미디어 컨테이너로 허용되는 콘텐츠는 하나 이상의 <par>, <seq>, <media> 요소로 구성된 세트* Google Cloud TTS API 추가 기능으로 제공	<p>1. voice</p> <p>WOMAN_READ_CALM : 여성 차분한 낭독체(default) MAN_READ_CALM : 남성 차분한 낭독체 WOMAN_DIALOG_BRIGHT : 여성 밝은 대화체 MAN_DIALOG_BRIGHT : 남성 밝은 대화체</p>
<p>2. <seq></p> <ul style="list-style-type: none">* 미디어 요소를 하나씩 재생할 수 있게 해주는 순차적 미디어 컨테이너* 콘텐츠: 하나 이상의 <seq>, <par>, <media> 요소로 구성된 세트이고 미디어 요소의 순서는 렌더링 되는 순서와 같음* 하위 요소의 begin, end 속성은 오프셋 값으로 설정* Google Cloud TTS API 추가 기능으로 제공	<p>2. SSML 중 <prosody>, <break> 세부 속성이 SSML 표준과 다르나 속성값의 범위는 표준의 범위를 만족함</p> <ul style="list-style-type: none">- prosody rate: slow, medium, fast- prosody volume: soft, medium, loud- break time: 단위는 ms만 사용하며 150(0.15초)~1500(1.5초)까지 지원
<p>3. <media></p> <ul style="list-style-type: none">* <par> 또는 <seq> 요소 내 미디어 레이어를 나타냄* 콘텐츠: SSML <speak> 또는 <audio> 요소* Actions on Google 플랫폼의 추가 기능으로 제공	<p>3. kakao:effct</p> <ul style="list-style-type: none">* 카카오에서 제공하는 custom tag* tone 속성 사용으로 존댓말→반말 변경 가능 (반대로는 불가능)- default: 기본 말투, 변경 없음- friendly: 친구 같은 말투

30

SSML의 작동 원리와 회사 별 SSML 구문의 특징을 비교, 분석하여 부록으로 정리



Amazon Polly



The screenshot shows the Amazon Polly console in the AWS Management Console. The browser address bar displays `ap-northeast-2.console.aws.amazon.com/polly/home/SynthesizeSpeech`. The left sidebar shows the navigation menu with "Amazon Polly" selected, and "Text to Speech" is the active sub-menu. The main content area is titled "텍스트 투 스피치" (Text to Speech). It contains instructions in Korean: "음성을 듣고 수정 및 다운로드하십시오. 준비가 되면 통합해 보십시오." and "입력창에 텍스트를 입력하거나 붙여 넣고, 언어와 리전을 선택하고 원하는 음성을 선택합니다. [음성 듣기]를 클릭하고 원하는 애플리케이션과 서비스에 통합합니다. 최대 3,000자를 이용해 즉시 청취, 다운로드 또는 저장할 수 있습니다. 최대 100,000자까지 작업을 S3 버킷에 저장해야 합니다." Below the instructions are two tabs: "일반 텍스트" (General Text) and "SSML" (Selected). The SSML tab is active, showing a text input area with the following SSML code:

```
< speak>
오늘의 < prosody rate='x-slow'>날씨를 전해 드리겠습니다.</ prosody>
현재, 전국이 구름이 많은 가운데 일부 중부 지역과 전북에는
< prosody volume='x-loud'>눈이 날리거나</ prosody>
< break time='1s'>
```

 Below the text input, it says "315 문자 사용". To the right of the text input are two links: "기본 텍스트 표시" (Show Default Text) and "텍스트 지우기" (Clear Text). Below the text input area, there are two dropdown menus: "언어 및 리전" (Language and Region) set to "한국어" (Korean) and "음성" (Voice) set to "Seoyeon, 여성" (Seoyeon, Female). To the right of these dropdowns are three buttons: "▶ 음성 듣기" (▶ Listen to Audio), "MP3 다운로드" (Download MP3), and "S3에 합성" (Synthesize to S3). Below the "MP3 다운로드" button, it says "샘플 비율: 22050Hz" and "파일 형식 변경" (Change File Format). Below the "S3에 합성" button, it says "S3 작업 설정 변경" (Change S3 Job Settings). At the bottom left, there is a link "▶ 발음 사용자 정의" (▶ Custom Pronunciation).



SSML EXAMPLES-ENG_01

< speak version="1.0" xml:lang="en-GB" >

< prosody rate = "x-fast" >

This sentence is spoken fast

</prosody>

< prosody pitch = "x-low" >

This sentence is spoken low pitch

</prosody>

< emphasis level = "strong" >

This sentence is spoken with stress

</emphasis>

< prosody duration = "10s" >

This sentence will be spoken out in ten seconds </prosody>

</speak>



SSML EXAMPLES-ENG_02

< speak >

< prosody volume="x-loud" pitch="x-high" rate="x-fast" > **Today we preview the latest music from Example.** < /prosody >

Hear what the < emphasis level="strong" > **Software Reviews** < /emphasis >
said about Example's newest hit.

Today is < say-as interpret-as="date" format="md" > **12/18** < /say-as >

He sings about issues that touch us all. < lang xml:lang="es-ES" > **Mucho gusto.** < /lang >

Would you like to < break time="300ms" / > **buy it?**

< /speak >



SSML EXAMPLES-KOR_01

< speak >

안녕하세요, 저는 은주입니다.< break time="1s"/> 오늘 날짜는

2019년 12월 18일 이구요,

저는 이렇게 얘기하고 싶습니다, "SSML은 신기해요!"

< amazon:effect name="whispered" > "SSML은 신기해요!" </amazon:effect >

</ speak >



SSML EXAMPLES-KOR_02

```
< speak>  
오늘의 < prosody rate='x-slow'>날씨를 전해 드리겠습니다.</prosody>  
현재, 전국이 구름이 많은 가운데 일부 중부 지역과 전북에는  
< prosody volume='x-loud'>눈이 날리거나</prosody>  
< break time='1s' />  
< prosody pitch='x-high'>빗방울이 떨어지는 곳이 있습니다.</prosody>  
서울의 경우 북부 지역을 중심으로  
< prosody rate='10%'>눈이 날리고 있으나,</prosody>  
공식적인 첫눈으로 기록되지는 않습니다.  
</ speak>
```



제4장 감성 음성기술과 로봇공학



감성기술 관련 특허 동향

<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



감성음성기술 관련 특허 분석

박종철, et al. (2011).

감정 음성 합성 기능을 가지는 보조 로봇 및
보조 로봇용 감정 음성 합성 방법 및 기록 매체



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년01월10일
(11) 등록번호 10-1221188
(24) 등록일자 2013년01월04일

(51) 국제특허분류(Int. Cl.)
B25J 13/08 (2006.01) G06F 19/00 (2011.01)
G06F 9/44 (2006.01)

(21) 출원번호 10-2011-0039199

(22) 출원일자 2011년04월26일

심사청구일자 2011년04월26일

(65) 공개번호 10-2012-0121298

(43) 공개일자 2012년11월05일

(56) 선행기술조사문헌

KR100814569 B1

JP2005283647 A

KR100644814 B1

JP2008243043 A

전체 청구항 수 : 총 23 항

(73) 특허권자

한국과학기술원

대전 유성구 구성동 373-1

(72) 발명자

박종철

대전광역시 유성구 대학로 291, 한국과학기술원
전산학과 (구성동)

이호준

대전광역시 유성구 대학로 291, 한국과학기술원
전산학과 (구성동)

(74) 대리인

김강욱

심사관 : 명대근

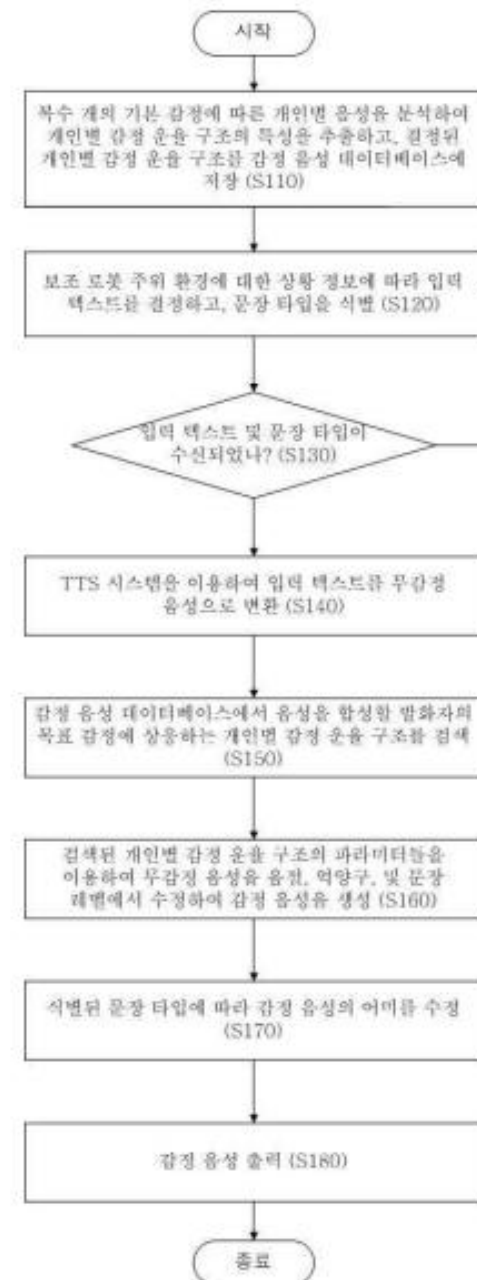
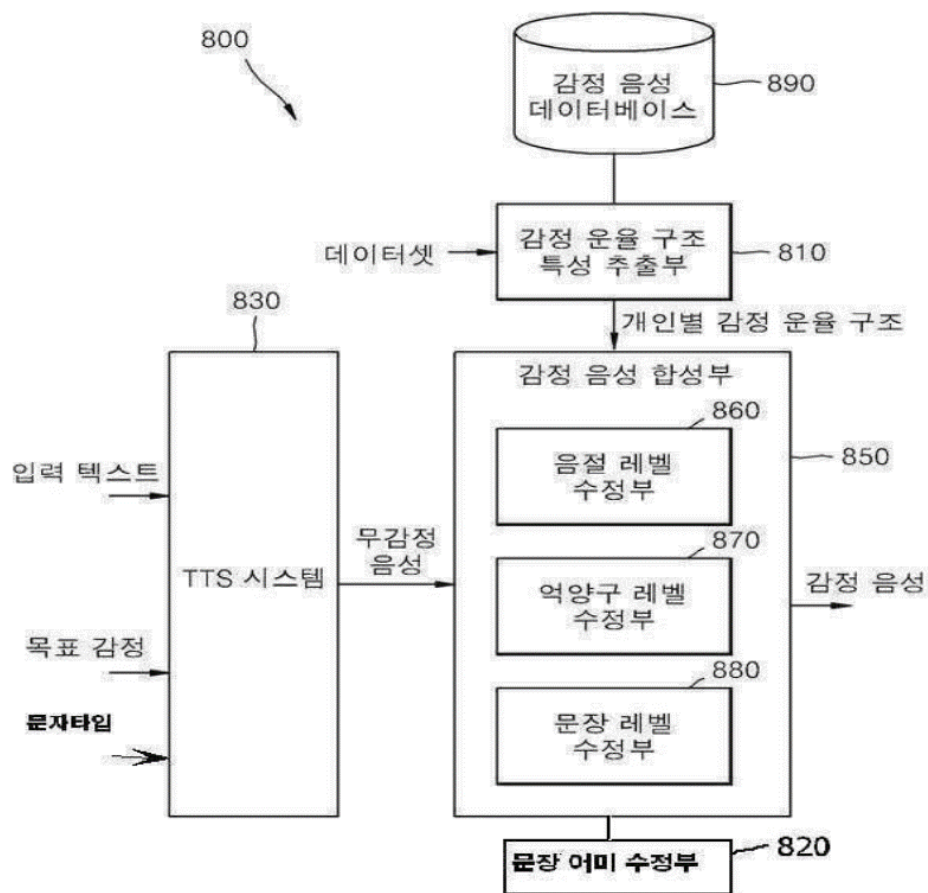
(54) 발명의 명칭 감정 음성 합성 기능을 가지는 보조 로봇 및 보조 로봇용 감정 음성 합성 방법 및 기록 매체

(57) 요약

인간과 상호작용하는 보조 로봇에서 출력할 감정 음성을 개인 운율 모델에 기반하여 합성하기 위한 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 대응하는 개인별 감정 운율 구조를 검색하는 단계, 및 입력 텍스트를 무감정 음성(emotionless speech)으로 변환하고, 변환된 무감정 음성을 목표 감정에 대응하는 개인별 감정 운율 구조에 기반하여 수정함으로써 발화자에 대응하는 감정 음성을 생성하며, 생성된 감정 음성의 문장 어미(sentence final syllable)를 식별된 문장 타입에 따라서 수정하여 출력하는 감정 음성 합성 단계를 포함한다. 본 발명에 의하여 사용자와 보조 로봇 사이의 상호 작용의 품질을 향상시킬 수 있다.



감성음성기술 관련 특허 분석



<그림 4-2> 감성 음성 합성 기능을 가지는 보조 로봇의 블록도





제5장 맺음말



맺음말

- 사람 사이의 커뮤니케이션 증진, 개인의 이해, 가족관계 증진, 여론 및 정책형성 등의 효과
- 감성 ICT는 일상의 전 산업에 융합 적용될 것이며 미래 우리 삶의 질을 책임질 기술
- 사생활 보호 문제 해결을 위한 기술적 부분과 법제도 정비가 필요한 부분
- 멀티모달 센싱 및 딥러닝 기술이 꾸준히 발전되어야 함
- 감성 관련 데이터베이스의 구축도 매우 중요
- 정신적 질환(자폐, 우울증 등)이 있는 분들에게 진찰은 물론 치료 목적으로 매우 유용하게 사용될 수 있는 기술





감사합니다.

