
저자 (Authors)	권순일, 손귀영, 김경인, 박능수, 이석필
출처 (Source)	전기의세계 68(10) , 2019.10, 22-27(6 pages) The Korean Institute of Electrical Engineers 68(10) , 2019.10, 22-27(6 pages)
발행처 (Publisher)	대한전기학회 The Korean Institute of Electrical Engineers
URL	http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09216314
APA Style	권순일, 손귀영, 김경인, 박능수, 이석필 (2019). AI 스피커를 위한 음성기반 감정인식 기술. 전기의세계, 68(10), 22-27
이용정보 (Accessed)	연세대학교 165.***.14.104 2019/11/16 23:33 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

AI 스피커를 위한 음성기반 감정인식 기술

권순일 교수
세종대학교

손귀영 선임연구원
세종대학교

김경인 연구원
세종대학교

박능수 교수
건국대학교

이석필 교수
상명대학교

1

서론

최근 4차 산업혁명의 시대로 접어들면서, 다양한 분야에서 인공지능 기반 융합적인 응용 기술개발이 빠르게 진행되고 있다. 특히, AI 스피커, 서비스 로봇 등 실생활에 밀접하게 연관되어 있는 다양한 스마트 디바이스의 사용이 빠르게 확산됨과 동시에, 사용자의 수요도 증가되고 있는 추세이다. (그림 1)

사용자의 얼굴이나 지문, 목소리를 인식하는 것에 한정되지 않고 더 나아가 사용자가 하는 말의 의미를 파악하고 이에 적절하게 대처할 수 있는 인공지능 시스템에 관한 관심이 고조되고 있다. 특히 사용자의 감정에 대해 목소리, 표정, 생체리듬 등을 활용하여 인지하는 기술이 앞으로 해결해야 할 과제로 부상하고 있다. 최근 IT 기술이 발달하면서 인간의 감정을 읽어내기 위하여 목소

리, 얼굴, 행동 등 다양한 생체신호 데이터 구축 및 인공지능 알고리즘까지 활용하여 그 접근방법도 다양해지고 있다.

음성 감정인식은 사용자의 음성을 활용하여 감정을 인지하는 기술이다. 과거의 음성 자체를 인식하여 텍스트로 변환하는 기술 단계에서 점차 발전되었고, 현재에는 사용자의 감정을 인식하고자 하는 시도가 늘고 있다. 감정인식기술을 기반으로 하는 미래의 서비스를 제공하기 위해 다양한 분야와의 융합을 통하여 끊임없이 노력하고 있다. 또한, 최근의 인공지능 기술에서의 음성분석기술을 도입하는 사례가 증대되고, 딥러닝 기술을 이용한 감정인식에서의 좋은 결과를 도출하기도 하였다. (그림 2)

사용자에 대한 감정을 읽어내고 이 정보에 기반한 개인화 서비스를 제공하는 기술은 미래의 기술과 산업의 새로운 돌파구를 마련하는데 기여할 수 있으며, 인공지능 및 음성인식 기술이 축적되어있는 국가들은 앞다투어 감정인식 분야에 관심을 가지고 지속적으로 인공지능 기술개발에 있어서 필수적인 연구주제로 인식

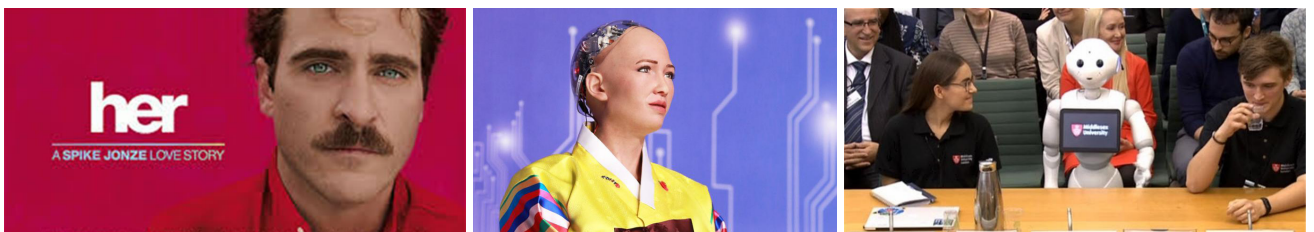


그림 1 인공지능 로봇 영화, 세계 최초 미국 시민권을 받은 AI 로봇 소피아, 영국의회에서 연설한 인간형 로봇 페퍼 (좌>우)



그림 2 IBM 딥러닝 기반 인공지능 시스템(미국)과 이용자의 감정 파악이 가능한 지능형 음성 인식 앱 'Moodie'(이스라엘) [1,2]

물적 자원을 투자하고 있다.



음성기반 감정인식 기술 관련 국내외 개발 동향

독일, 미국 등 감정인식 연구를 지속적으로 수행한 국가들에게서는 실시간 음성기반 감정인식에서 가장 기초가 되는 감정유발 시나리오에 근거하여 감정기반 음성 데이터베이스 구축해 왔고, 전 세계적으로 동 분야를 연구하는 연구자들의 기술개발에 활용되고 있다 [3-6]. Berlin Emotional Speech DB는 독일 베를린 공과대학 연구팀에서 훈련된 연기자로부터 구축한 7가지의 감정에 대한 총 800개의 문장으로 구성된 독일어 데이터베이스이다. Danish emotional Database는 덴마크 커뮤니케이션 센터에서 구축한 감정 데이터베이스로, 비훈련자를 대상으로 구축되었다. 인도의 Vcl 공과대학 연구팀은 배우의 발화를 이용한 데이터베이스

스(훈련된 연기자)와 실제 상황에 가까운 인공적인 상황(콜센터, 환자-의사 대화)을 녹음한 데이터베이스로 8개 감정에 대하여 12,000개로 구성되어 있다. 미국의 USC 공과대학 연구팀은 10명의 훈련된 연기자를 대상으로 시나리오 기반, 상황 기반의 대화를 통하여 8개 감정으로 구성된 감정음성 데이터베이스를 구축하였다. (표 1)

수집된 데이터베이스를 활용하여 기계학습, 특히, 딥러닝 기반의 알고리즘을 활용하여 감정인식 연구가 지속적으로 이루어지고 있다. 독일 예를랑겐대 연구팀은 Berlin Emotional Speech DB를 이용하여, 운율, 묵음길이, 되풀이, 수정정보 분석을 통한 7가지 감정(Happiness, Sadness, Anger, Boredom, Disgust, Fear, Neutral)인식 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. 중국의 Geosciences 대학과 Advanced Control and Intelligent

표 1 감정음성 데이터베이스

데이터베이스	국가	언어	감정범주
Berlin Emotional Database(EMO-DB) [3]	독일	독일어	Anger, Happiness, Sadness, Fear, Disgust, Boredom, Neutral
Danish emotional Database [4]	덴마크	덴마크어	Anger, Joy, Sadness, surprise, neutral
IITKGP: SEHSC [5]	인도	힌두어	Anger, Happiness, Sadness, Fear, Disgust, Boredom, Neutral, Sarcastic, Surprise
IEMOCAP [6]	미국	영어	Happiness, Anger, Sadness, Neutral, Disgust, Fear, Excitement, Surprise

표 2 기계학습 기반 감정음성인식 최근 연구결과

Database	Literature	Acoustic features	Classifier	Accuacy(%)	Emotions
EMO DB	Quan et. al. (2017) [7]	Correlation, cepstral distance, MFCC, prosodic	SVM	80	Angry, Fear, Happy, Sad, Surprise, Neutral
	Palo et. al. (2018) [8]	LPCCVQC	MLP	83	Anger, Sad, Happy, Boredom
		MFCCVQC	RBFN	79	
		PLPVQC	PNN,DNN	76	
IEMOCAP	Lee et. al. (2015) [9]	LLDs	BLSTM-ELM	63.89	Angry, Sad, Happy, Neutral
	Mirsamadi et. al (2017) [10]	LLDs	BLSTM-WPA	58.8	Angry, Sad, Happy, Neutral
	Fayek et. al (2017) [11]	Spectrogram	LSTM	58.05	Angry, Sad, Happy, Neutral
	Tzinis et. al (2017) [12]	Statistical	LSTM	60.02	Angry, Sad, Happy, Neutral

Automation for Complex Systems 연구소 그리고 Central South 대학에서는 Fisher를 통해 Feature를 선택하고 ELM(Extreme Learning Algorithm) 의사 결정 트리 기반의 기계학습 방법을 통해 6가지 감정(Happiness, Sadness, Surprise, Anger, Fear, Neutral) 인식 기술을 연구하였으며 평균 89.6%의 인식 성공률을 보였다. 이 밖에도, 최근 국제적으로 발표된 논문들에서는 음성 기반의 감정인식 성능을 향상시키기 위한 다양한 노력을 보여주고 있다. 하지만 아직까지 AI 기반 상품에 적용할 만한 수준의 정확도에는 미치지 못하고 있다[7-12]. (표 2)

음성기반 감정인식 알고리즘 개발을 통하여 실생활에 활용되는 사례도 증대되고 있다. 대표적으로 뉴럴네트워크 기반 사용자 음성톤을 분석으로 감정을 인지하는 IBM의 딥러닝기반 감성시스템 톤-애널리저가 있으며, MIT의 컴퓨터과학 및 인공지능연구소(CSAIL)는 사람의 말소리 패턴과 생체 상태를 기반으로 대화의 감정강도(행복, 슬픔, 중립)를 예측하는 인공지능 웨어러블 시스템을 개발하였다. 마지막으로, 스마트폰 카메라를 활용하여 실시간으로 사용자의 감정을 인식하는 카네기 멜론대 로봇연구소의 InterFace는 음성과 표정까지 추가하여 우울증을 파악하는 임상실험도 진행하고 있다.



감성 미디어 서비스를 위한 음성기반 감정인식 기술 개발과 활용

감성 미디어 서비스를 위한 음성기반 감정인식 기술은 사용자의 음성을 통해 감정을 인식한 후, 이를 바탕으로 사용자에게 감정에 따른 최적의 미디어 서비스를 제공하는 기술이다. 최근 방송 스마트미디어 분야의 서비스 기술이 빠르게 발전됨에 따라 사용자 맞춤형 콘텐츠를 제공하기 위한 사용자의 감정인식을 통하여 사용자 수요에 부합되는 맞춤형 미디어를 제공할 필요성이 증대되면서 기술 개발이 진행되고 있다. (그림 3)

기계학습을 이용한 감정인식 연구가 많이 진행되고 있으나 영상정보만을 이용하여 감정인식을 하는 연구가 대다수이다. 하지만 영상의 경우 휴대용 기기를 이용한 모바일 환경에서 취득하거나 정보를 처리하는 데 어려움이 있지만, 이를 극복하는 데 있어서 음성만을 이용한 감정인식의 필요성이 증대되고 있다.

세종대학교 연구진과 상명대학교, 건국대학교 연구진들은 최근 음성정보만을 이용한 감정인식 기술을 개발하기 위해 국내에서 활용 가능한 감성음성 데이터베이스를 구축하고, 이를 기반으로 최적의 특징추출/분석/인식 및 인덱싱 기술 개발을 진행하고 있다. 미디어 기반 감정인식 기술 개발에 있어서 필요한 핵심요소는 한국형 감정 데이터베이스를 구축, 이를 기반으로 하여 최적화



그림 3 감성미디어 서비스를 위한 음성기반 감정인식 기술 개발의 개요

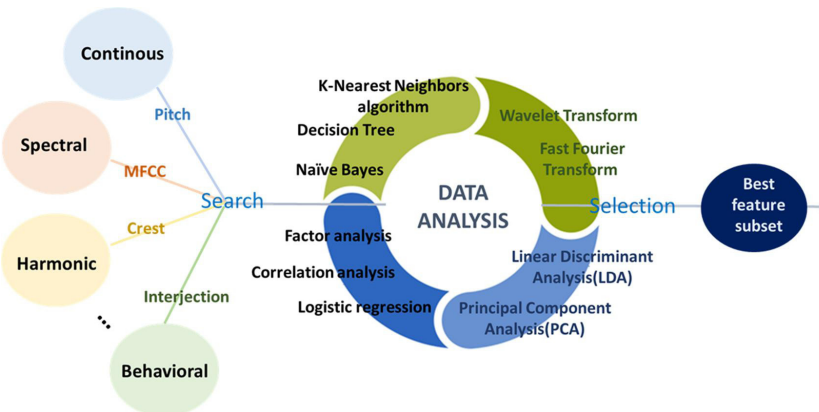


그림 4 음성기반 최적의 특징요소 선정

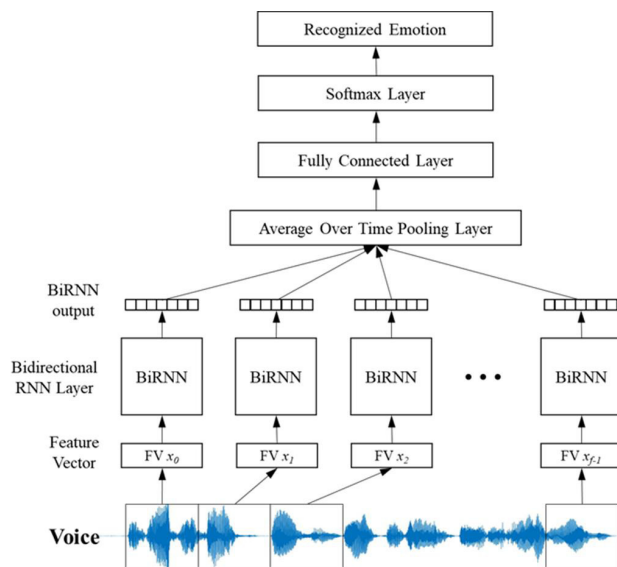


그림 5 AoT Bi-directional RNN Model

정 데이터베이스는 기존의 훈련된 연기자들
통하여 감성유발 시나리오에 근거하여 음성
을 녹음하고 이를 표준화하는 과정을 통하여
데이터베이스 구축이 진행되고 있다. 또한,
기존의 감정인식에서 많이 사용되는 Pitch,
MFCC 외의 언어학, 심리학 등의 다학제적
융합적 측면에서 접근을 통한 감성표출의 비
언어적 발화형태 등의 새로운 감정인식 특징
요소를 추출 및 최적화 특징요소 조합을 발굴
하려고 하고 있다(그림 4). 마지막으로, 인공
지능 기술개발 중 음성에 최적화된 딥러닝 기
반 감성분류에 최적화된 알고리즘 개발을 진
행하고 있다. 이 알고리즘은 발화의 특정 시
점 이전과 이후 정보를 모두 사용하여 앞뒤
정보에 의존성을 갖는 발화의 특징을 더 잘
반영하고, 출력값에 AoT(Average over Time)
방식의 변형된 모델 추가하여 음성신호의 짧
은 부분적 구간뿐만 아니라 전체 구간에 대
한 정보를 통해 감성분류가 가능하다(그림 5).

감정을 인식하고 이를 이용한 인터페이스
기술은 새로운 형태의 스마트 미디어 콘텐츠와의 소통 및 감정을
전달하는 방법이며, 감성 ICT 제품과 서비스 콘텐츠의 성장기반
구축 및 기술의 글로벌 확산과 신규 비즈니스 창출에 기여할 수
있다. 최근 글로벌 기업은 정체된 ICT 제품구매 촉진을 위해 감
성전달 기술에 역량을 강화함으로써 실감·감성기술 중심의 시장
을 주도하려는 추세를 보이고 있다. 소비자 참여형 미디어 생산
및 공유 기술을 통해 사업자 주도의 서비스에서 사용자의 감성 키
워드 검색기반 개인화 서비스로의 진화된 감성 미디어 서비스를
통해 사용자 경험 창출 및 신 ICT 산업이 조성될 수 있을 것이다.

감정인식 기술은 개인화 서비스 제공을 통한 사용자의 긍정적
마인드 형성 및 정서 불안 해결 등 개인적 삶의 질 향상, 감성 기
반 원격 의료연동으로 사용자의 심리상태를 진단하는 의료서비
스 및 심리적 문제 해결에 기여할 수 있는 감성 기반 심리치료 서
비스를 제공할 수 있다. 또한, 한국인의 정서를 반영한 한국어로
수집된 데이터 통한 한국인 사용자 맞춤형 감성 미디어 빅데이터

된 음성기반 특징요소를 추출 및 알고리즘 개발을 목표로 한다.

한국어 기반 표준화된 데이터베이스 구축을 위해 한국형 감


확보가 가능하며, 다양한 스마트 기기를 활용한 사용자 감성 최적화 콘텐츠 검색 및 추천, 감성 프로파일링 등의 기술 개발의 활성화에 큰 도움을 줄 수 있다. 마지막으로 미디어 콘텐츠의 소비 증가에 따른 새로운 패러다임 제시와 신시장 개척, 스마트 가전, 문화기술 등의 타 산업과의 융합을 통한 동반 성장이 기대된다.

4

결론

인간과 컴퓨터(혹은 로봇) 간의 상호작용 및 인공지능 기술 개발에 있어서 주목을 받기 시작하고 있는 음성기반 감정인식 기술을 소개하였다. 인간이 감정을 인식하는 메커니즘을 생각해 볼 때 다양한 종류의 정보를 통하여 정확한 감정인식을 수행할 수 있을 것이고 이를 위해서는 멀티모달 형태의 특징요소들을 기반으로 감정인식 기술 개발도 중요하지만, 이에 앞서 음성정보에 기반한 싱글모달 형태의 특징요소 기반 감정인식의 기반을 구축하는 것도 요소기술 개발 차원에 있어서 중요한 부분이라고 볼 수 있다. 하지만, 이러한 감정인식을 위한 기본적인 감정인식 데이터베이스는 부족하고, 또한 사회·문화적 요소를 반영한 데이터베이스는 거의 없는 실정이다. 무엇보다도, 사용자를 이해하고, 사용자를 대신할 수 있는 보다 나은 기술 개발을 위하여 사용자들이 원하는 바, 추구하는 바가 무엇인가를 보다 면밀한 관찰을 통하여 기초부터 차근차근 기술개발에 접근할 필요가 있다.

인공지능을 가진 시스템이나 디바이스와 사용자 상호 간의 인터페이스에 있어서 단순한 명령이나 정보전달 형태의 의사소통을 뛰어넘어 사용자의 의도와 감성을 소통할 수 있으려면 반드시 감정인식 기술이 필요하며 이를 상용화하기 위해서는 감정을 이해하는 연구부터 응용연구에 이르기까지 다양한 연구가 지속적으로 진행될 필요가 있으므로 국가 차원에서 중장기적인 투자가 선행되어야 할 것이다. 마지막으로 사용자의 감정인식을 위한 연구 및 서비스가 진행될 경우 정보의 속성상 민감한 개인정보를 다루고 있다고 볼 수 있고, 이러한 개인정보 유출과 해킹에 대한 우려가 클 것으로 판단되며, 나아가 프라이버시와 보안 문제

의 법적, 기술적 해결이 해당 기술 상용화에 도약할 수 있는 발판이 될 것이다. 

참고문헌

- 1 <https://www.ibm.com/watson/services/tone-analyzer/>
- 2 <https://thenextweb.com/apps/2014/01/23/beyond-verbal-releases-moodies-standalone-ios-app/>
- 3 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.
- 4 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.
- 5 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.
- 6 Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335.
- 7 Quan, C., Zhang, B., Sun, X., & Ren, F. (2017). A combined cepstral distance method for emotional speech recognition. International Journal of Advanced Robotic Systems, 14(4), 1729881417719836.
- 8 Palo, H., & Mohanty, M. N. (2018). Comparative Analysis of Neural Networks for Speech Emotion Recognition. Int. J. Eng. Technol, 7, 112-116.
- 9 Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In Sixteenth Annual Conference of the International Speech Communication Association.
- 10 Tzinis, E., & Potamianos, A. (2017). Segment-based speech

- emotion recognition using recurrent neural networks. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 190-195). IEEE.
- 11 Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92, 60-68.
 - 12 Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231). IEEE.