감성과학
제24권 1호, 2021
<연구논문>

pISSN 1226-8593
eISSN 2383-613X
Sci. Emot. Sensib.,
Vol.24, No.1, pp.91-104, 2021
https://doi.org/10.14695/KJSOS.2021.24.1.91

Hi, KIA! 기계 학습을 이용한 기동어 기반 감성 분류

Hi, KIA! Classifying Emotional States from Wake-up Words Using Machine Learning

김태수[1] · 김영우[2] · 김근형[3] · 김철민[4] · 전형석[5] · 석현정[6†]
Taesu Kim[1] · Yeongwoo Kim[2] · Keunhyeong Kim[3] · Chul Min Kim[4] ·
Hyung Seok Jun[5] · Hyeon-Jeong Suk[6†]

Abstract

This study explored users' emotional states identified from the wake-up words —"Hi, KIA!"— using a machine learning algorithm considering the user interface of passenger cars' voice. We targeted four emotional states, namely, excited, angry, desperate, and neutral, and created a total of 12 emotional scenarios in the context of car driving. Nine college students participated and recorded sentences as guided in the visualized scenario. The wake-up words were extracted from whole sentences, resulting in two data sets. We used the soundgen package and svmRadial method of caret package in open source-based R code to collect acoustic features of the recorded voices and performed machine learning-based analysis to determine the predictability of the modeled algorithm. We compared the accuracy of wake-up words (60.19%: 22%~81%) with that of whole sentences (41.51%) for all nine participants in relation to the four emotional categories. Accuracy and sensitivity performance of individual differences were noticeable, while the selected features were relatively constant. This study provides empirical evidence regarding the potential application of the wake-up words in the practice of emotion-driven user experience in communication between users and the artificial intelligence system.

Key words: Voice-User Interface (VUI), Wake-Up Words, Machine-Learning, Acoustic Feature, svmRadial, Emotional User Scenario

요 약

본 연구에서는 승용차에서 사람들이 기기를 사용하기 위해 사용하는 기동어인 "Hi, KIA!"의 감성을 기계학습을 기반으로 분류가 가능한가에 대해 탐색하였다. 감성 분류를 위해 신남, 화남, 절망, 보통 총 4가지 감정별로 3가지 시나리오를 작성하여, 자동차 운전 상황에서 발생할 수 있는 12가지의 사용자 감정 시나리오를 제작하였다. 시각화 자료를 기반으로 총 9명의 대학생을 대상으로 녹음을 진행하였다. 수집된 녹음 파일의 전체 문장에서 기동어 부분만 별도로 추출하는 과정을 거쳐, 전체 문장 파일, 기동어 파일 총 두 개의 데이터 세트로 정리되었다. 음성 분석에서는 음향 특성을 추출하고 추출된 데이터를 svmRadial 방법을 이용하여 기계 학습 기반의 알고리즘을 제작해, 제작된 알고리즘의 감정 예측 정확성 및 가능성을 파악하였다. 9명의 참여자와 4개의 감정 카테고리를 통틀어 기동어의 정

※ 이 논문은 교육부의 재원으로 한국연구재단 4단계 BK21사업의 지원을 받아 수행됨(NO.4120200913638).
※ 이 논문은 2018년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행되었음(No.2018R1A1A3A04078934).
[1] 김태수: KAIST 산업디자인학과 박사과정
[2] 김영우: KAIST 산업디자인학과 학사과정
[3] 김근형: KAIST 문화과학기술대학원 박사과정
[4] 김철민: KAIST 원자력및양자공학과 박사과정
[5] 전형석: 현대자동차 기아디자인센터 기아내장디자인실
[6†] (교신저자) 석현정: KAIST 산업디자인학과 부교수 / E-mail: color@kaist.ac.kr / TEL: 042-350-4523

확성(60.19%: 22~81%)과 전체 문장의 정확성(41.51%)을 비교했다. 또한, 참여자 개별로 정확도와 민감도를 확인하였을 때, 성능을 보임을 확인하였으며, 각 사용자 별 기계 학습을 위해 선정된 피쳐들이 유사함을 확인하였다. 본 연구는 기동어만으로도 사용자의 감정 추출과 보이스 인터페이스 개발 시 기동어 감정 파악 기술이 잠재적으로 적용 가능한데 대한 실험적 증거를 제공할 수 있을 것으로 기대한다.

주제어: 보이스 인터페이스, 기동어, 기계 학습, 음성 피쳐, 서포트 벡터 머신, 감성 시나리오

1. 서론

사람의 음성을 이용해서 기기를 제어하는 방식인 보이스 인터페이스(Voice User Interface, 이하 VUI)에 대한 연구와 상용화 기술은 인공지능 스피커와 같은 음성인식 기기의 보급화와 함께 최근 들어 더욱 활발히 진행되고 있다. VUI는 사용자가 일상 용어를 활용하기 때문에 물리적 버튼(Physical User Interface)이나 스크린 GUI(Graphic User Interface)와 비교해서 더 쉽고 직관적으로 제어할 수 있다. VUI의 가장 큰 기술적 관심화자의 음성을 정확하고 신속히 이해하는 데 있는 만큼, VUI의 정확도 향상을 위한 음성인식 기술과 자연어 처리 기술이 적극적으로 접목되고 있는 것은 자연스러운 결과이다.

이와 동시에 기기에 입력된 음성 데이터는 문자로 변환되어(Speech-to-Text; 이하 STT) 단어 혹은 나아가 맥락을 파악하는 목적으로 활용될 수도 있지만 음성에 내포된 화자의 감성 상태 또한 파악이 가능하다. 인공지능 분석 기술을 접목한 음성기반 감정인식(Speech-Emotion Recognition; 이하 SER) 관련 연구가 다수 공개되어 왔고, 기술의 고도화를 위해 언어권별 특징을 반영하거나 분석 방법을 개선하는 등 발전을 거듭해 오고 있다(Swain et al., 2018). 이에 음성을 기반으로 작동하는 어시스턴트의 개발로 이어지고 있는데, 구글과 애플과 같은 전자기기 업체들은 휴대폰을 차량에 연결해 Android Auto, Apple’s Siri via CarPlay와 같이 기존의 어시스턴트를 차량에서 손쉽게 사용할 수 있도록 하였다. 또한, 이 움직임은 Toyota사의 Concept-I, Honda의 Houndify와 같이 차량 업계에서 출시하는 컨셉카에서 차량 요소 중 하나로 간주된다. 메르세데스 벤츠(Mercedes Benz)는 플래그십 모델인 S-Class 차량에서 차량 맞춤형 보이스 어시스턴트를 탑재하였다. 보이스 어시스턴트(voice assistant)는 브랜드의 상징성을 담고 있어 어시스턴트 실행을 위한 기동어(Wake-Up Word)를 브랜드 이름이나 상징과 연관지어 제작하는 경우가 빈번한데, 벤츠의 경우도 “Hi, Mercedes!” 라는 기동어를 사용하고 있다(Voicebot, 2020). 이렇게 차량 내 기기 제어를 위해 VUI가 주목을 받는 이유는 VUI는 운전 중 전방에 주의를 집중하면서도 차량 내 제어가 용이하여 운전자 안전과도 직결되기 때문이다. 실제로 NIO사의 NOMI, BMW의 IPA, Benz사의 MBUX와 같이 다수의 자동차 제조 업체가 차량 제어의 수월성과 운행 중 정보 검색의 편의성을 높이기 위해 차량 특화형 VUI의 개발과 그 사용성 증진을 위해 노력하고 있다.

자율주행 3단계 이하의 차량에서는 VUI가 운전자의 안전성 등을 최우선순위로 고려되어 왔다. 그래서 SER을 차량에 접목하는데 있어서도 감성적 만족감과 운전자의 주행 능력간 상호작용에 대해 많은 연구가 진행되었다. 예를 들어, Nass et al.(2005)의 연구에서는, 차량과 운전자의 감정이 일치할 경우, 운전자의 운전 성능과 태도 그리고 안전성에 영향을 준다는 사실을 확인한 바 있다. 그러나 점차 차량이 스스로 안전성을 확보하는 기술을 더 높은 수준으로 확보해 나가고 자율 주행 기술이 적극적으로 차량에 반영됨에 따라 SER 기술을 접목한 VUI 에 대한 탐색이 더욱 활발하게 진행되어 왔다. Jones & Jonsson(2005)와 Jones & Jonsson(2007)의 연구에서는 차량 시뮬레이터 내부에서 실험 참가자들의 음성 대화 데이터를 수집하여 운전자의 감정을 분석하였다. 음성에 대해 사람과 알고리즘이 각각 평가한 결과는 평균 65% 정도가 상호 일치하는 것을 보여준 바 있다. 해당 연구에서 감정 인식 시스템은 2초 간격으로 분할되었으며 한 번의 발화를 10개의 단위로 구분하여 학습 데이터셋으로 활용하였다. Schuller et al.(2006)에서는 보다 정확도를 높이기 위해 음성 샘플에서 음소 단위의 특징을 추출하여 감정을 인식하는 연구를 진행했다. 음성 신호를 10ms의 단위로 쪼개어 분석한 결과 평균 73.63%의 정확도로 운

전자의 감정을 분류해낼 수 있었다. 추가적으로 Support Vector Machine(이하 SVM) 모델에 Sequential Forward Floating Selection(이하 SFFS) 알고리즘과 feature space optimization까지 적용하여 평균 정확도 80.93%를 달성할 수 있었다. 이처럼 음절 전체의 음성적인 특징이 아닌 음소 단위의 특징을 감정 인식에 활용한 경우, 녹음자 개인의 음색이나 말하는 방식에 덜 의존적이면서도 정확도 높은 결과를 얻을 수 있었다. 이러한 기존 연구를 바탕으로, 기동어만으로 운전자와 탑승자의 감성을 파악할 수 있는 기술의 발전 가능성을 예측할 수 있다. 최근 Kim et al.(2020)의 연구에서 EmoDB 를 학습 데이터 셋으로 활용한 Bidirectional Long Short-Term Memory(이하 BLSTM) 구조의 딥러닝 분석 방식으로 기동어의 감성 분류를 시도한 바 있다. 학습 데이터 셋이 독일어였던 점, 7가지의 감성 분류로 도출되는 결과가 차량의 기동어 감성에 자연스럽게 매핑되기 어려웠던 점 등을 한계로 감성 분류의 정확도에 대한 구체적인 결과가 제시되지 않았다. Këpuska et al.(2009) 의 연구에서 짧은 길이의, 그러나 반복적 패턴을 가지는 측면을 기반으로 기동어 인식의 정확도가 확인된 바 있듯이 “Hi, OO!”로 규격화된 어휘에 한정하여 인공지능 기술이 효율적으로 접목될 수 있는 특수성이 있다. 기동어만으로도 이미 사용자의 감성을 파악할 수 있다면, 차량 시스템의 반응 방식에서도 사용자의 감성적 상태에 최적화된 방식의 콘텐츠나 맞춤형 소통 방식이 매칭 가능할 수 있다. 이에 기동어를 말하는 음성의 음향적 속성에 기초하여 기계 학습 방법을 적용시키며, 이를 통해 도출된 알고리즘을 활용하여, 기동어 기반 감성 예측의 가능성을 살펴보고자 한다.

2. 연구 방법

본 연구에서는 국내 자동차 대표 브랜드 중 하나인 기아자동차를 대상으로 “Hi, KIA!”를 기동어로 활용하였다. 차량 시스템 사용자가 주어진 감성 상태에 따라 기동어를 어떻게 표현하는가에 대한 기계학습을 진행한 후, 도출된 모델의 성능 평가를 계획하였다. 모델의 성능을 평가하기 위한 기준으로는 기동어가 포함된 전체 문장에서 파악될 수 있는 감성을 대조군으로 간주하였다.

2.1. 다양한 감성이 내포된 “Hi, KIA!” 시나리오 구성

‘하이 기아’ 라는 음절이 ‘ㅋ’, ‘ㅊ’, ‘ㅆ’ 등과 같은 센 파열음이나 파찰음을 내포하지는 않기 때문에 화난 감정을 담아서 문장을 말하기에 효과적이지 않을 수 있다(Jang & Kim, 2005). 그리고 기동어가 발화되는 시간은 1초 미만인 경우가 일반적이므로 SER 연구에서 사용되는 음성 데이터와 비교하여 길이가 매우 짧아 기계 학습이 용이하지 않을 수 있다. 반면, 동일한 음절을 반복 학습하는 조건이므로 학습의 효과를 기대할 수 있는 잠재력 또한 특징이다. 이에 여러 가지 상황에서 발화되는 “Hi, KIA!”를 자연스러운 상황에서 수집하기 위하여 차량에 탑승하는 다양한 상황과 상황 별 감성적 특징을 설정해 보았다.

감성적 상황들을 나열하기 위하여, 연구자들 5명이 모여 1차적으로 총 53가지의 상황을 제안하였다. 상황은 현재 차량에 탑승하여 이용하는 상황 외에도, Wiegand의 연구에서 제안하는 바와 같이, 미래 자율주행 기술이 상용화 되었을 때, 스마트 홈 연동 혹은 가상 비서로서 자동차의 역할과 같이 미래 자동차의 역할과 관련한 상황도 포함되도록 폭 넓게 제안하도록 하였다(Wiegand et al., 2019). 이 중 긍정(Pleasure)과 각성(Arousal) 차원의 감성적 특징이 명확한 경우로 한정하고 일반인이 빈번하게 경험하가를 기준으로 12가지의 지시용 문장으로 Table 1과 같이 수렴 구성하였다. 예를 들어, E1번 상황의 경우 재미있는 일이 생겨서 신나는 기분으로 차량에 탑승한 상황에 해당하는데, 운전자가 차량에 탑승하면서 신나고 기대된다는 느낌으로 차량 시스템을 호출하려는 장면으로 구체화될 수 있다. 상황과 장면의 감성 상태에 몰입이 된 것을 전제로 녹음 문구를 제시하여 “Hi, KIA, 철수에게 전화해 줘~”를 녹음하는 것을 계획하였다. 또한 각 상황 별 감성 분류에 있어 긍정과 각성 차원의 특징이 복합적으로 반영된 신남(Excited), 화남(Angry), 절망(Desperate), 그리고 보통(Neutral)의 구분을 적용하였다. 이러한 구분은 기본적으로는 긍정과 각성의 독립적인 축으로 구성된 2차원 감성 평면(Russell, 1980)을 기준으로 각 사분면을 대표하는 감성과 중심점을 설명하고자 함이다. 긍정과 차분함으로 설명될 수 있는 4사분면의 경우, 차량에 탑승하는 운전자가 경험하는 상황에서 수집되지

Table 1. 14 Emotional user scenarios include situations, recorded sentences, and emotions. Korean sentences that were used for recording were notated together

ID	Situation	Specific Situation	Recorded sentences (“ ”)	Emotion
E1	Excited situation happened today, riding a car with excited emotion.	Call my friend for sharing~! (excite and anticipation)	Hi KIA, Call to CheolSu~! (철수에게 전화해줘)	Excited
E2	Enjoying nice and sunny weather.	The weather is lovely~! (bright and cheerful)	Hi KIA, Open the sunroof~! (썬루프 열어줘~!)	Excited
E3	Going on a trip with friends.	Done for loading? Here we go~! (amusement and excitement)	Hi KIA, Go to Haeundae~! (해운대로 가자~!)	Excited
A1	Voice assistant of the car does not understand my request.	Why doesn’t it understand me? (explicit and angry)	Hi KIA, Call. To. Doctor. Kim. (김.철.수 에게 전.화.해.)	Angry
A2	Voice assistant of the car does not respond.	Is it broken? What happened? (clear and annoying)	Hi KIA!! Hi KIA~ Hi KIA? (하이 기아~ 하이 기아?)	Angry
A3	Right after leaving the house, reminded of forgetting to turn off the boiler.	I do not think I turned off the boiler! (panic and depressed)	Hi KIA!! Check the boiler status!! (보일러 확인좀!)	Angry
D1	Ride the car in bad weather with rain pouring and thunder lightning.	It is raining too much... (panic and fear)	Hi KIA... When will rain stop... (비가 언제 그칠까...)	Desperate
D2	Ride the car to take a rest after finishing hard work.	Ahhh... I am so tired... (tired and relaxed)	Hi KIA. Massage mode. (마사지 모드.)	Desperate
D3	Ride the car on a gloomy day.	I could have done a little better... (melancholy and sorrowful)	Hi KIA... I... failed again... (나.. 또 떨어졌어...)	Desperate
N1	Briefing today’s weather from AI in the car.	What’s the weather like today~ (daily and mundane)	Hi KIA, How’s the weather today? (오늘 날씨 어때?)	Neutral
N2	Brand new updated voice assistant of K5 is greeting you.	I am gonna say hi, too~ (calm and normal)	Hi KIA, Hello? (안녕?)	Neutral
N3	Ride the car to go home with tired but excited emotion.	It is finally time to go home~ (happy and normal)	Hi KIA. Go to the home. (집에 가자)	Neutral

않아 제외되었다.

감성적 표현을 효과적으로 유도하는 과정에서 해당 상황을 현실감 있게 묘사할수록 몰입도가 높아 감성 표현 수집에 용이하다고 밝혀진 바 있다(Alcamo, 2008; Ogilvy, 2011). 본 실험의 경우, 운전 중의 상황에 어시스턴트를 호출하는 상황보다, 주행 목적지나 차량 탑승 직전까지 있었던 일에 몰입한 후 감정을 담아 발화를 해야 하기에, 과거와 미래를 한번에 표현할 수 있도록 이미지 보드를 제작하여 녹음을 진행하였다. 또한, 올해 확산된 코로나 질병의 확산으로 비대면으로 감정 유도 및 녹음이 가능할 수 있는 방법을 탐색하던 중, 시나리오 이미지를 이용해 감정 유도와 음성 수집이 가능하다면, 실험을 보다 원활히 진행할 수 있다고 판단해, 본 연구에서 기동어를 호출하는 상황을 묘사하는 이미지를 사용하였다. 예를 들어, Fig. 1에 제시된 이미지는 신나는 상황에서 친구에게 전화를 걸기 위한 상황에 참여자가 몰입을 쉽게 할 수 있도록 제작되었음을 확인할 수 있다.

Fig. 1. Twelve scenarios were sequentially displayed to the survey participants in random order. The above scenario in the center corresponds to the E1

2.2. 실험 방법

실험은 Phonic.ai에서 제공하는 설문 플랫폼을 이용하여 온라인으로 진행하였다. 플랫폼에 기본적인 인적사항을 입력하고 나면, 화면에 제시된 각 장면마다 녹음해야 할 문구가 제공되었고, 각 참여자는 각 장면을 보고 감정을 유도한 뒤 충분한 연습을 한 후 녹음 버튼을 눌러 음성 녹음을 진행하도록 유도되었다. 1차 인터뷰와 평가를 통해서 일반인들 중 음성 에이전트를 사용해본 경험이 있으며, 하루에 한번 이상은 에이전트를 호출하는 사람을 1차적으로 선정하였다. 이후, 상대적으로 감성적 음성 표현이 정확하고 풍부하다고 연구자가 판단한 참여자들을 구인하였으며, 이에 대한 정보를 Table 2에 제시하였다. 남학생 5명과 여학생 4명으로 구성된 총 9명의 대학생이 참여하였고, 참여자들의 평균 나이는 24.88(표준 편차 = 2.13)세였다. 모든 참여자에게는 사례비가 제공되었다.

구인된 참여자들은 개인 노트북을 이용해 플랫폼에 접속하면, 12가지의 상황에 대해 녹음을 진행하게 된다. 첫 번째 녹음이 종료되면, 연구자들이 파일을 듣고 녹음 품질에 대한 피드백 후 추후 녹음을 진행하도록 하였다. 녹음은 첫 번째 녹음 후 2일 내 본인이 편안한 시간에 남은 녹음을 진행하도록 했다. 그 결과 동일인이 총 12가지 상황을 녹음하는 세션을 4~6회 반복 진행하였고, 총 653개의 음성 파일을 9명의 참여자로부터 수집하였다.

각 시나리오에 대해서 동일인이 여러 차례 반복 녹음을 하였으므로, Fig. 2에 나타나듯이 동일한 문장의 경우 매우 유사한 패턴을 관찰할 수 있다. 앞서 Table 1에 소개된 내용과 같이 각 시나리오 마다 문장의 길이와 구성이 크게 차이가 나지만, 동일 시나리오에 대한 반복 녹음은 전체적으로 유사한 패턴이 나타나는 것이다. 따라서 3.2 장에 진행된 문장 단위의 트레이닝과 테스트 데이터 셋 구분 시 시나리오 간 구분이 적용되도록 하였다. 이를 통해 동일 문장의 패턴을 학습한 데 기인한 예측의 정확도 상승을 배제하고자 하였다.

Table 2. Participant demographics and usage of voice assistants they used

ID	Age	Gender	Voice agent he/she have experienced	Voice disabilities
F01	25	Female	Siri, Clova Friends	X
F02	27	Female	Bixby, Google Home	X
F03	26	Female	Siri, Google Home	X
F04	27	Female	Bixby	X
M01	24	Male	Siri, Google Home	X
M02	28	Male	Siri	X
M03	23	Male	Bixby	X
M04	22	Male	Siri, Clova	X
M05	22	Male	Siri, Clova	X

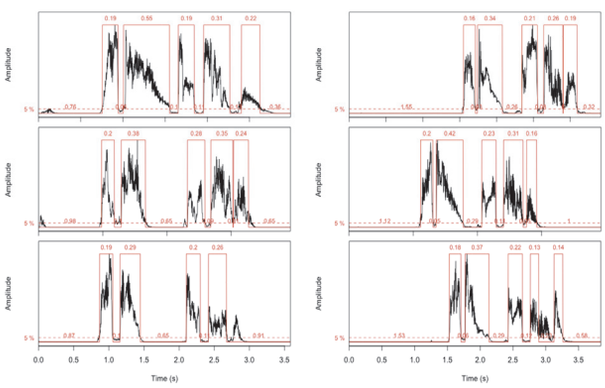

Fig. 2. The voice amplitudes of the participant, F03, recorded for the E2, “(brightly and cheerfully) Hi, KIA, open the sunroof~!”, repeating six times

2.3. 기동어 음향 속성 기반 학습 데이터 구성 수집

총 653개의 녹음된 음성 파일은 기동어 부분과 기동어를 포함한 전체 문장 두 가지로 구분하였다. 참여자별 표현 방식의 차이가 있고 시간 편차를 감안하여, 오픈 소스 기반 언어 R의 seewave 패키지를 이용하여 각 음성 파일마다 녹음 시그널의 패턴을 파악한 후 문장 앞부분에서 발화된 기동어 부분을 추출하였다.

추출된 기동어 녹음 길이는 평균 595.29 ms(표준 편차 = 133.40 ms)로, 각 감성별 구분에 따른 평균 시간은 Table 3에 제시된 바와 같이 화가났을 때가 평균 544.29 ms 로 가장 짧고 절망적일 때가 687.09 ms 로 가장 긴 것으로 관찰되었다. 전반적으로 상당히 짧은 구간임에도 불구하고 화가 났을 때 말의 속도가 빨라지고 우울할 때 느려지는 말하기 속도에 대한 Davitz(1964)의 원론적 연구와 같은 경향성을 확인할 수 있다. 목소리와 음악의 단순한 음향적 속성을 토대로 감성적 특징을 파악하고자 한 Nordström과 Laukka 의 연구에서는 기동어보다도 훨씬 더 100 ms 이하의 짧은 자극에 대해서도 기쁨과 슬픔을 구분하는 기본적인 감성 분류가 정확하게 판별될 수 있음을 보여주었다(Nordström &

Laukka, 2019). 또한 동일한 음소, 음절, 단어 개수라는 제한된 조건 내에서 미묘한 차이에도 상대적인 비교 분석이 가능할 잠재성 또한 예상 가능하다.

Table 3. The average duration of the Wake-up Words, “Hi, KIA” across the emotion categories. Standard deviations are in parentheses. Unit: millisecond (ms)

	Excited	Angry	Desperate	Normal
Average duration (SD)	606.74 (102.93)	544.29 (104.28)	687.09 (180.16)	576.54 (115.47)

2.4. 분석 데이터 처리

기동어는 일상 언어와는 달리 동일한 음소, 음절, 단어 개수를 가지고 있어 인공지능 모델이 정확성 높게 인지하기에 용이한 장점이 있다. 문장에 비하여 길이가 매우 짧고 상대방을 호출한다는 한정된 조건으로 인하여 다양한 감성을 담기에는 한계가 있는 측면도 갖고 있다. 본 연구에서는 문장 전체에서 기동어 부분을 발췌한 데이터 셋을 별도 구성하여 기동어를 학습한 경우와 문장 전체, 나아가 기존의 알려진 알고리즘을 학습한 경우와 비교하여 그 분류 결과에 대하여 문장 전체에 대한 정확도와 비교하고자 하였다.

녹음된 음성의 다양한 음향 속성을 파악하고자 R 의 soundgen 패키지를 활용하여 각 녹음 파일에 대해 매 25ms 간격으로 총 37가지 음향 속성을 추출하였다. 이 중 사람의 목소리로 판별된 구역(‘Voiced’ 피쳐값이 ‘TRUE’에 해당)에 대해서만 분석값을 산출하는 피처들이 중복 산출되어 있어, 최종적으로는 총 24개의 피처가 연산에 활용되었다. Soundgen 패키지로 추출한 음성 피처의 종류는 크게는 ‘시간적(Time)’, ‘크기적(Loudness)’, ‘높낮이(Pitch)’로 구분되어 있다. 데이터 셋을 구성하는 전체 관측치는 총 12,219개로, 결과적으로는 “Hi, KIA!” 한 개의 기동어 녹음에 대해 평균 18.71회(12,219 관측치 / 653 녹음건수)에 대한 추출값을 분석에 사용하였다(부록 A 참조).

2.5. 기계 학습 기반 데이터 분석

학습-테스트 데이터 구분을 위해 한 번의 기동어 음성 녹음을 테스트 데이터로, 그리고 해당 음성 녹음을 제외한 전체 데이터를 학습 데이터로 간주하는 1 out of N 방식을 적용하였다. 분류를 위한 기계 학습을 위한 방법으로 k-NN이나 recursive partitioning, Random Forest (RF) 계열과 같은 트리 구조의 분류를 적용하거나, SVM 계열과 같이 신경망을 모사한 방법 등 다양한 알고리즘을 적용해 볼 수 있다. 본 연구에서는 Kuhn (2008)이 고안한 Caret 패키지를 활용하여, 다양한 기계 학습 알고리즘을 피험자 1에 대해서 테스트한 후, 연산속도가 빠르면서도 정확도와 결과값의 안정된 정도가 높은 편으로 획득되는 svmRadial을 선택하여 분석에 적용하였다. 데이터 학습 과정에서는 RF 패키지의 Boruta 함수를 이용하여 영향력이 높은 피쳐를 선택

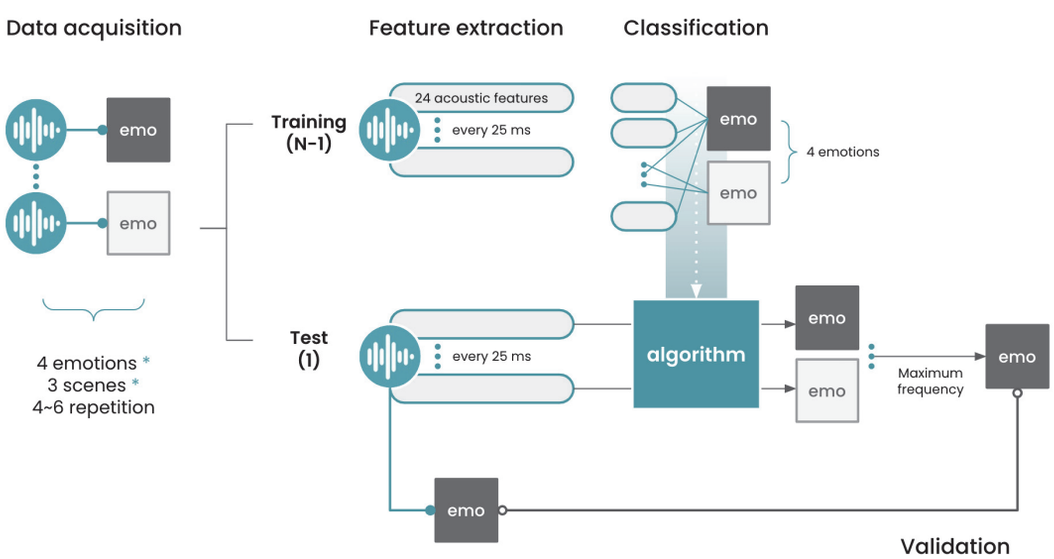

Fig. 3. Algorithm modeling and the prediction of representative emotion of each “Hi, KIA!” record

적으로 적용하였으며 알고리즘 모델 형성 시, 학습 데이터 내 cross validation은 10회를 적용하였다.

한 개의 기동어 음성 녹음에 대해 매 25ms 마다 감성 분류가 적용되었으므로, 평균 18.71회에 해당하는 감성 분류 결과를 얻게 된다. 이 때 분류 결과 중 가장 높은 빈도로 분류된 감성을 기준으로 예측 결과를 판단하였다(Fig. 3 참조).

3. 연구 결과

3.1. 기동어 학습 기반 기동어 감성 분류 과정의 피쳐 중요도 분석

기계 학습 과정에서 Boruta 함수를 이용하여 영향력이 큰 피쳐들을 선택한 후 선택된 피쳐들 만으로 알고리즘을 모델링하였는데, 다음 Fig. 4에서 알고리즘 모델링에 활용된 피쳐들의 중요도를 각 연산마다 meanImp 값으로 산출된 값을 기반으로 정리해 보았다. 녹음 길이(Duration)의 중요도가 두드러지게 높고 다섯 번 째인 거칠기(Roughness) 까지가 상위권의 피쳐임을 확인할 수 있다.

개인별 알고리즘 모델링 과정에서도 녹음 구간 중 음성이 기록된 시간(Duration_no_silence')과 후처리 된 음높이(Pitch), 소리 크기의 변화를 통한 거친 정도(Roughness'), 주요 음역대의 영역(Dom) 등이 개인에 따라 우선순위의 변동은 있으나, 고정적으로 최상위 순서에 기록되었다.

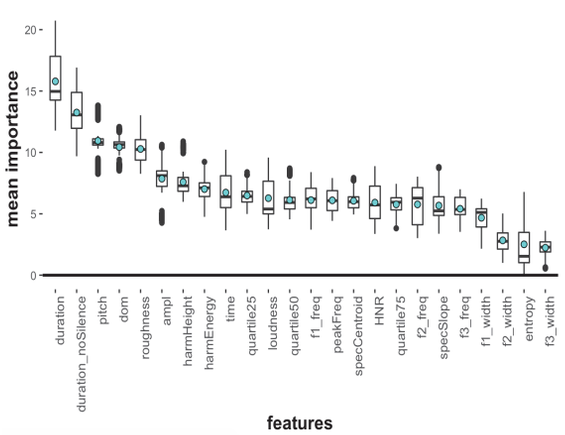

Fig. 4. The features' importance scores derived during the feature selection

3.2. 기계 학습 기반 기동어 감성 분류 결과

다음 Table 4에는 각 음성 녹음 단위로 테스트를 적용한 결과를 참여자 별로 구분하여 보여주고 있다. 예를 들어, F01 참여자의 경우 excited 상황에서 녹음한 총 18번의 "Hi, KIA!" 중 13번은 excited로, 2번은 angry, 그리고 3 번은 neutral로 예측된 결과로 나타났다. 이 때 괄호속 백분율은 민감도(Sensitivity) 수치로서 연구자가 의도한 감성 분류가 의도한 대로 네 가지 분류에 대한 전체 결과 기준 정확도는 F01의 경우 59%로 산출이 되는데, M01의 경우 81%에 달하여 개인별 차이가 크게 관찰되었다. 특히, M02는 전체 정확도가 22%에 그치는데, M02의 경우 9명의 참여자 중 경상도 남부 지역 방언을 심하게 사용하는 참여자로, 네 가지로 구분된 감성 시나리오에서 기동어 부분의 음성 표현이 타 피험자에 비하여 차이가 적게 났거나 녹음의 일관성이 다소 낮은 것으로 사료된다. 참여자의 개인별 정확도 분석 결과가 포함된 전체 자료는 부록 B에 첨부하였다.

단 정확한 분류임을 판단하는 기준은 연구자들이 설계한 시나리오에서 의도한 감성이 실제 참여자 음성 녹음에 충분히 표현하였다는 것을 전제로 하였다. 따라서, 각 음성 녹음에 대해 다수의 사람들이 실제 해당 감성을 잘 인지하였는가를 토대로 학습 데이터셋을 선별적으로 구성하여 정확도를 더 객관적으로 산출할 수 있다. 또한 피처 추출의 전문성과 알고리즘 모델링의 고도화와 같은 인공지능 기술을 접목하여 개선이 시도될 수 있다.

3.3. 기동어와 문장 전체 감성 예측 비교

총 12개의 시나리오는 본 연구에서 적용한 네 가지 감성 구분에 고르게 분포하므로 각 감성 분류마다 3개의 문장이 포함되어 있다(Table 1 참조). 본 연구에서는 기동어에 대한 감성 분류가 만족 할만큼 정확한가를 판단하기 위해서 문장 전체의 감성 인식 분류 결과를 베이스라인으로 설정하였다.

이에, 문장 전체에 대한 감성 분류 알고리즘 모델링의 경우 각 상황을 테스트 데이터 셋으로 간주하고 나머지 데이터 셋을 학습하도록 알고리즘을 모델링하였다. 각 시나리오 마다 동일한 과정을 반복하여, 정확도를 산출하였다.

Table 4. Results of wake-up words emotion classification using machine learning. a single recording file divided into every 25 ms, and emotions are predicted for each divided recording file. then, the representative emotion was classified to the highest frequency among the four categories. entire data are shown in Appendix B. Modeling methods: svmRadial

Participant	Expected emotion based on scenario	Prediction results of the **wake-up words** emotions using machine learning (% shows its sensitivity)				Top 5 weights feature calculated from the Boruta function.
		Excited	Angry	Desperate	Neutral	
F01 (25)	Excited	13 (72%)	2	0	3	Rank 1~5: “duration” > “duration_noSilence” > “pitch” > “dom” > “roughness” Accuracy: 59%
	Angry	4	6 (33%)	1	7	
	Desperate	0	1	12 (71%)	4	
	Neutral	5	1	1	11 (61%)	

Table 5. Results of wake-up words emotion classification using machine learning. modeling methods: svmRadial

Participant	Expected emotion based on scenario	Prediction results of the **wake-up words** emotions using machine learning (% shows its sensitivity)				Top 5 weights feature calculated from the Boruta function. All subjects show same results.
		Excited	Angry	Desperate	Neutral	
All subjects	Excited	107 (65%)	30	3	24	Rank 1~5: “duration” Rank 2: “duration_noSilence” Rank 3~5: “pitch”, “dom”, “roughness” (Rank 3~5 changes on each subjects) Accuracy: 60% (60.19%)
	Angry	35	87 (53%)	10	33	
	Desperate	4	11	112 (71%)	31	
	Neutral	23	26	28	84 (52%)	

다음 Table 6에는 문장을 기준으로 개인별로 문장을 녹음한 내용을 기계 학습하여 테스트 문장에 대한 감성 분류 결과를 수집한 다음 9명 전체 참여자의 결과를 집계하여 제시하고 있다. 전체적으로 기동어 학습 기반 기동어의 감성 분류를 한 경우보다(Table 5) 문장을 학습하여 문장의 감성을 분류한 경우 정확도가 더 낮게 나타남을 살펴볼 수 있다. 다음 Table 7에서는 각 참여자 별로 기동어와 문장이 얼마만큼의 정확도로 감성 분류되었는가를 요약하고 있다.

한편, 기동어가 포함된 문장 전체에 대한 감성 분류는 본 실험에 사용된 문장에 한하여 기계학습을 접목한 알고리즘에 제한될 필요가 없다. 오히려 전문가 음성을 학습하여 딥러닝 알고리즘으로 분류된 알고리즘을 활용할 수 있겠다. 더불어, 문장 단위의 감성 분류는 본 연구에서처럼 음향적 속성에 한정된 분석이 아니라 STT 변환으로 취득한 문장에서 어휘나 문맥을 파악한 바를 기반으로 음향적 속성과 함께 감성적 특징을 판단할 수 있다. 그 경우 문장에 기반한 감성 분류는 그 정확도가 기동어의 음향적 속성만으로 판단한 분류 결과에 비하여 현격히 높을 수 있다. 하지만 말하는 내용

Table 6. Results of whole sentences. after learning the scenario recordings except itself, each recording are predicted its emotion. modeling methods: svmRadial

Participant	Expected emotion based on scenario	Prediction results of the **whole sentences** emotions using machine learning (% shows its sensitivity)				Top 5 weights feature calculated from the Boruta function. All subjects show the same results.
		Excited	Angry	Desperate	Neutral	
All subjects	Excited	97 (59%)	18	2	47	Rank 1~5: “duration_noSilence”, “duration”, “pitch”, “dom”, “roughness” Accuracy: 42% (41.51%)
	Angry	39	87 (53%)	21	18	
	Desperate	14	34	35 (24%)	63	
	Neutral	70	12	34	45 (28%)	

Table 7. Results of sensitivity scores across the four emotion categories and all participants

Participant	Results of sensitivity compare expected emotion based on scenarios using 1 out of N method Emotion classification using Wake-up Words (Left) Emotion classification using Whole Sentences (Right) (Modeling methods: svmRadial)							
	Excited		Angry		Desperate		Neutral	
	Wake-up Word	Whole Sentences	Wake-up Word	Whole Sentences	Wake-up Word	Whole Sentences	Wake-up Word	Whole Sentences
F01	72%	56%	33%	67%	71%	12%	61%	31%
F02	78%	50%	83%	0%	50%	8%	68%	56%
F03	76%	71%	72%	61%	86%	31%	76%	69%
F04	81%	71%	76%	29%	84%	7%	37%	0%
M01	89%	94%	78%	67%	78%	0%	78%	28%
M02	28%	17%	6%	72%	39%	11%	17%	17%
M03	83%	83%	56%	56%	94%	33%	50%	17%
M04	61%	72%	50%	67%	67%	50%	61%	33%
M05	17%	17%	17%	61%	72%	61%	33%	6%
Average	65%	59%	53%	53%	71%	25%	52%	27%

을 시스템이 실시간으로 모두 듣고 해석하는 것에 대해 충분히 거부감이 있을 수 있다. 사용자에 대한 더 많은 정보를 시스템이 취득하는 것이 도움 될 수도 있지만 프라이버시를 침해당하는 부정적 감성으로도 이어질 수 있어, 사용자 중심의 감성 시나리오에 대한 연구가 인공지능을 이용한 정확한 알고리즘 개발에 앞서 선행되어야 한다.

4. 논의

4.1. 주요 발견점

본 연구에서는 차량 내 시스템을 구동하기 위해 사용되는 기동어가 사용자의 감성 상태를 내포할 수 있는가에 대하여 탐색하고자 하였다. 차량 내에서 기동어로 시스템을 호출하는 다양한 감성 시나리오를 자유롭게 탐색하였고 excited, angry, desperate, neutral의 네 가지 분류로 수렴, 각 분류 마다 세 가지 감성 시나리오를 고안하였다. 참여자 9명은 시각화된 감성 시나리오를 제시하고 해당 상황을 연상하면서 기동어로 시작하는 문장을 녹음하였다. 녹음은 기동어 부분을 발췌한 음성과 기동어가 포함된 문장 전체로 구분하여 두 개의 음성 데이터 셋을 구성하였다. Soundgen 패키지를 이용하여 총 24개의 음향 속성과 관련된 시계열 피처를 추출한 후, Caret 패키지에 포함된 svmRadial 기계학습을 이용하여 기동어와 문장 각 데이터 셋마다 각 감성 분류의 민감도 및 정확도를 산출하였다. Table 6에 제시된 바와 같이 개인 별 기계 학습으로 모델링된 알고리즘이 예측한 분류 결과의 민감도를 기동어와 문장의 경우로 나누어 비교하였는데, 전체 평균값을 기준으로 excited, desperate, neutral 에서 기동어를 학습하여 기동어의 감성을 분류한 민감도가 문장을 학습하여 문장의 감성을 분류한 경우보다 두드러지게 높은 것으로 나타났다. Angry의 경우는 기동어와 문장이 동일한 분류 결과를 보여주었다. 특히 문장의 분류 예측은 desperate와 neutral을 예측하는 민감도가 각각 25%와 27% 인데, 감성이 네 가지로 분류된 점을 감안하면 약 1/4에 해당하는 민감도는 변별력이 있었다고 판단하기 어려운 수치이다. 따라서 다양한 어휘와 문장 길이를 학습하여 새로운 문장의 감성을 예측하는 경우에 비해서 비록 0.6초 내외의 짧은 길이의 단순한 음절임에도 불구하고 기동어를 학습하여 새로운 어투의 기동어의 감성을 예측하는 것이 안정된 정확도를 기대할 수 있다고 사료된다.

기계 학습 과정에서 Boruta 패키지를 사용하여 의미있는 영향을 미치는 피처만 선택하고자 하였으나, 알고리즘 모델링 과정에서 24종류의 피처 대부분이 실제로

활용되었다. 반면, 기동어와 문장, 그리고 9명 모두 중요도가 높게 반영된 피처의 종류는 동일한 경향성이 발견되기도 하였다. 이렇듯 기계 학습 과정에서, 여러 피쳐 중 어떠한 요인들이 결과 판독에 비중 있는 역할을 파악하는 것은 알고리즘의 연산 양을 최소화할 필요가 있는 실용적 가치에서 더욱 의미가 있다. 비중이 높은 피쳐들에 대한 분석을 고도화하는 등의 후속 연구 방향을 제시할 수 있어, 기계 학습을 포함한 인공지능 연구 분야에서 '화이트 박스' 혹은 '해석적 기계 학습' 이 주목받고 있는 측면과 맥락을 같이 한다.

4.2. 기동어 감성 분류에 따른 차량 내 시스템 연동

다양한 감정적 상황에 따라 기동어가 다르게 표현된다는 것을 전제로 효과적인 음성 수집을 위해 12가지의 상황 시나리오를 선택하여 사용한 바 있다(2.2장). 이 12가지는 초안 53가지에서 선택과 수렴이 된 결과인데, 그 과정에서 각 시나리오 별로 차량 탑승 직후에 빈번하게 경험하는 종류인지, 그리고 유사한 감성과 연결되는지를 고려한 바 있다. 12가지의 상황 시나리오는 결과적으로 excited, angry, desperate, 그리고 neutral 이라는 감성 분류로 카테고리 범주로 구분이 되었는데, Russell(1980)의 감성 평면 기준으로 그 분포를 살펴보았을 때는 긍정(pleasure, positive valence)과 침착함(calm, negative arousal)에 해당하는 4사분면의 감성은 독립적인 카테고리로 지정되지 않았다. 감성 평면에서 4사분면은 나른함이나 편안한 휴식과 같은 상황에 대응되는 상황으로서, 차량 탑승 직후 운전자가 경험하는 빈번한 상황에는 해당하지 않는다고 판단한 데 있다. 만약 기동어 화자가 동승객이거나 차량이 자율주행 4단계 이상의 완전 자율 주행 차량이라면, 차량 탑승 후 휴식을 원하는 감성적 니즈가 매우 두드러지리라 예상한다.

본 연구를 통해 도출된 결과를 바탕으로 보자면, Table 6에 제시된 excited와 desperate의 시나리오에서 기동어의 감성 예측 정확도가 각 65%, 71%로 꽤 높은 편이며, 특히나 desperate의 경우 전체 문장의 정확도가 25%임에 비해 높은 정확도가 관찰되었다. 이는 두 감성이 차량을 이용하는 상황에서 가장 자연스럽게 유도되는 상황이기 때문으로 생각된다. Nass et al.(2005)의 연구에서도 두 감성을 음성 어시스턴트를 활용할 때 가장 두드러지는 감성이라고 소개한 바 있다. 이와 대조되게, angry의 경우 강한 감정임에도 불구하고 그 정확도가 낮게 관찰되었는데, 이는 A3에 제시된 상황이 다른 시나리오와 비교하였을 때, 화가 난 상황보다 다급한 상황에 더 가까워 감정 유도가 힘들었을 수도 있다는 점에서 그 정확도가 낮아졌을 가능성이 존재한다. 뿐만 아니라, 본 연구에서는 자동차 내에서 빠른 감정 판단을 위하여 soundgen을 활용해 빠르게 머신 러닝 방식을 이용하였는데, 이 과정에서 추출된 feature들이 anger를 판단하기에 충분하지 않았기 때문일 것으로 사료된다. 본 결과만을 통해 본다면, 사용자차량의 탑승객을 위한 서비스를 제공할 때, 탑승객의 excited된 상태와 desperate된 상태에 맞는 피드백을 제공하는 것이 가장 기본적인 감성 피드백 시스템의 방향성으로 생각할 필요가 있다.

이와 같은 맥락에서 기동어 만으로도 감성 분류를 잘 하는 연구의 방향은 기술적 고도화 뿐만 아니라, 사용자 중심의 감성 시나리오를 고안하고 선택적 구현을 하는 감성 기술이 선행되어야 하는 것이다. 즉 감성을 다루는 기술이 전략적 방향을 제시하고 목적성에 맞게 기술이 최적화되어 접목되어야 한다. 2장에서 시나리오를 도출하는 연구 과정과 같이 시스템이 사용자의 감성을 기동어를 기반으로 정확히 이해를 했을 때 과연 사용자는 어떠한 경험을 원하는가에 대한 탐색이 진행되어야 한다.

4.3. 연구 한계점 및 후속 연구

본 연구의 목적이 기동어에 담긴 감성적 표현만으로도 문장 전체를 통해 파악할 수 있는 감성 구분이 가능할 것인가를 탐색하는 데 있다. 따라서, 시스템이 취득한 기동어의 감성 분류에 적용할 알고리즘은 기동어에 한정하여 학습된 모델에 한정될 필요는 없다. Kim et al.(2020), Park et al.(2019), Nordström & Laukka(2019) 등의 연구에서와 같이 기존의 타 연구에서 제안된 음성 감성 인식 알고리즘을 적용하여 기동어의 감성적 분류가 정확도 높게 도출될 수 있다면 차량 시스템에 최적화하여 상용화가 충분히 가능할 것이다. 특히 SER 분야의 기술적 발전이 딥러닝 방식의 분석 기법에 접목되어 빠른 속도로 그 성과를 보여주고 있다. 더 쉽고 직관적인 그러나 사용자 입장에서는 비침습적인 방식

으로 VUI를 효과적으로 사용할 수 있는 사용자 중심 시나리오가 가능할 것이다.

더불어 본 연구에 참여한 남녀 대학생들은 1차 인터뷰와 평가를 통해서 일반인들 중 상대적으로 감성적 음성 표현이 정확하고 풍부하다고 연구자가 판단하여 섭외한 대상자들이다. 그러나 20대 초중반의 연령대에 한정된 점, 한국어가 모국어인 점, 표준어 사용자인 점 등 음성의 다양성을 포용하는데 한계가 있다. 인공지능 스피커나 가전제품, 차량에서 현재 사용되고 있는 기동어의 대부분은 영문식 이름이나 '하이' 와 같은 영어 호출이 이용되고 있지만 화자의 모국어에 따라 나름의 엑센트와 억양이 반영될 수 밖에 없다. Table 3에서와 같이 실험에 참여한 한정된 인원만으로 구성한 전체 데이터 학습 알고리즘보다 더 범용성이 높고 판별 결과의 안정성이 뛰어난 알고리즘을 모델링하기 위해서 다양한 언어권의 다양한 인적 구성을 통해 수집된 데이터 셋은 큰 가치가 있을 것이다.

또한, 본 연구에 사용된 기동어는 'Hi, KIA'로 제시되었으나, '아리야', '알렉사'와 같이 단순히 에이전트를 부르거나, 'Ok, Google', 'Hey, Kakao'와 같이 다른 호출 방식을 통해 호출하는 경우도 있다. 이에 본 연구가 하나의 기동어만을 이용해 확인해 보았다는 점에서 그 한계가 있으나, 기동어는 각 브랜드별 아이덴티티와 브랜드가 추구하는 전략에 맞춰 바뀌므로, 본 연구를 활용하고자 하는 제조사들은, 각 제조사에 맞춘 데이터 셋을 확보하여 본 연구를 응용할 필요가 있다. 본 연구의 연구 방법과 기동어를 이용한 감정 측정을 통해 VUI를 개발해 기업 나름의 로직으로 자사의 타깃 고객의 특성에 최적화된 감성 분류와 분류된 감성에 대응되는 서비스 및 콘텐츠에 대한 활발한 연구 및 상용화가 이루어지기를 기대한다.

5. 결론

차량 시스템을 호출하는 기동어를 기계 학습하여 알고리즘을 모델링하고 이를 이용하여 기동어 음성에 대해 네 가지 감성(Excited, Angry, Desperate, Neutral)으로 분류하는 정확도를 탐색해 보았다. 남녀 대학생 9명이 기동어로 시작하는 문장을 70여 회 녹음한 후 기동어 부분과 문장 전체로 구분하여 데이터 셋을 구성하였다. 시계열 음향 속성 추출과 기계 학습은 오픈 소스 기반 R 언어의 Soundgen과 Caret 패키지를 이용하였고 svmRadial 기법으로 데이터 셋을 학습시켰다. 문장 전체에 대한 감성 분류와 비교하여 약 0.6초 길이의 기동어의 감성 분류가 angry는 동일한 수준을, 나머지 세 가지 감성 분류에서는 기동어의 경우가 두드러지게 높은 예측률을 확인하였다. 네 가지 감성 분류를 모두 포함한 전체 정확도(Accuracy)의 경우, 기동어와 문장에 대한 결과가 각각 60.19%와 41.51%로 나타났는데, 정확도와 민감도(Sensitivity) 모두 개인별 차이가 크게 발견되었다. 정확도의 경우 22%~81%의 범위에서 개인별 차이가 있었지만, 전체적으로는 기동어의 감성 예측 결과가 감성 분류 전체에 걸쳐 문장에 비해서 안정적으로 예측됨을 확인하였다. 문장에 대한 감성 분류는 문맥에 대한 이해를 추가하거나 기동어와 문장 모두 딥러닝 기반의 고도화된 분석 기법을 이용하여 정확도를 개선할 수 있어 후속 연구에서는 분석 방법의 개선을 기대할 수 있다. 동시에, 사용자 중심의 감성 시나리오가 탐색되어야 하며 인공 지능 기술이 이를 뒷받침할 때 더욱 가치 있는 서비스로 완성될 수 있다.

REFERENCES

Alcamo, J. (2008). Chapter six the SAS approach: combining qualitative and quantitative knowledge in environmental scenarios. *Developments in integrated environmental assessment*, *2*, 123-150. DOI: /10.1016/S1574-101X(08)00406-7

Davitz, J. R. (1964). *The communication of emotional meaning*. Oxford, England: McGraw Hill.

Jang, K., & Kim, T. (2005). The pragmatic elements concerned with the sounds of utterance. *Korean Semantics*, *18*, 175-196.

Jones, C. M., & Jonsson, I. M. (2005). Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. *In Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (pp. 1-10), Narrabundah, Australia, Nov. 2005. DOI: 10.5555/

1108368.1108397

Jones, C. M., & Jonsson, I. M. (2007). Performance analysis of acoustic emotion recognition for in-car conversational interfaces. *In International Conference on Universal Access in Human-Computer Interaction* (pp. 411-420). Berlin, Heidelberg, DOI: 10.1007/978-3-540-73281-5_44

Këpuska, V. Z., & Klein, T. B. (2009). A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation. *Nonlinear Analysis: Theory, Methods & Applications*, *71*(12), e2772-e2789. DOI: 10.1016/j.na.2009.06.089

Kim, Y., Kim, T., Kim, G., Jeon, H., & Suk. H. J. (2020). Hi Kia~, hi... kia..., HI KIA!! *Proceeding of Fall Conference of Korean Society for Emotion and Sensibility* (pp. 21-22), Daejeon.

Nass, C., Jonsson, I. M., Harris, H., Reaves, B., Endo, J., Brave, S., & Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. *In Proceedings of CHI '05 Extended Abstracts on Human Factors in Computing Systems 2-7* (pp. 1973-1976), Portland, Oregon, USA. DOI: 10.1145/1056808.1057070

Nordström, H., & Laukka, P. (2019). The time course of emotion recognition in speech and music. *The Journal of the Acoustical Society of America*, *145*(5), 3058-3074. DOI: 10.1121/1.5108601

Ogilvy, J. (2011). *Facing the Fold: Essays on Scenario Planning (*pp. 11-29). Devon: Triarchy Press.

Park. J., Park, J., & Sohn, J. (2013). Acoustic parameters for induced emotion categorizing and dimensional approach. S*cience of Emotion and Sensibility*, *16*(1), 117-124.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161-1178. DOI: 10.1037/h0077714

Schuller, B., Lang, M., & Rigoll, G. (2006). Recognition of spontaneous emotions by speech within automotive environment. *Proceedings of German Annual Conference of Acoustics,* Braunschweig, Germany, Mar, 2006.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93-120. DOI: 10.1007/s10772-018-9491-z

Voicebot. ai. (2020). In-car voice assistant consumer adoption report. Retrieved from https://voicebot.ai/wp-content/uploads/2020/02/in_car_voice_assistant_consumer_adoption_report_2020_voicebot.pdf

Wiegand, G., Mai, C., Holländer, K., & Hussmann, H. (2019). InCarAR: A Design Space Towards 3D Augmented Reality Applications in Vehicles. *In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, Netherlands* (pp. 1-13), DOI: 10.1145/3342197.3344539

원고접수: 2020.11.27
수정접수: 2020.12.16
게재확정: 2020.12.21

Appendix A. Features Extracted by Soundgen Package of R

Type	Feature name	Description	Unit
Time	Time	Time of the middle of each frame	ms
	Duration	Total duration	s
	Duration_ noSilence	Duration from the beginning of the first non-silent STFT frame to the end of the last non-silent STFT frame	s
Loudness	Ampl	Amplitude. Root mean square of amplitude per frame, calculated as sqrt(mean($frame^2$))	dB
	HarmEnergy	The amount of energy in upper harmonics, namely the ratio of total spectral mass above 1.25 x F_0 to the total spectral mass below 1.25 x F_0 (dB)	dB
	HarmHeight	How high harmonics reach in the spectrum, based on the best guess at a pitch	
	HNR	Harmonics-to-noise ratio	dB
	Loudness	Subjective loudness, in sone, corresponding to the chosen SPL_measured	
	Roughness	The amount of amplitude modulation	
Pitch	Dom	Lowest dominant frequency band (Hz)	
	Entropy	Weiner entropy of the spectrum of the current frame. Close to 0: pure tone or tonal sound with nearly all energy in harmonics; close to 1: white noise	0~1
	f1 ~ f3_freq	The frequency and bandwidth of each n-th formants per STFT frame	
	f1 ~ f3_width		
	PeakFreq	The frequency with maximum spectral power	Hz
	Pitch	Post-processed pitch contour based on all F_0 estimates	
	Quartile25, Quartile50, Quartile75	The 25^{th}, 50^{th}, and 75^{th} quantiles of the spectrum of voiced frames	Hz
	SpecCentroid	The center of gravity of the frame's spectrum, first spectral moment	Hz

Appendix B. Results of Wake-up Words Emotion Classification Using Machine Learning

Participant	Expected emotion based on scenario	Prediction results of the **wake-up words** emotions using machine learning (% shows its sensitivity)				Top 5 weights feature calculated from the Boruta function.
		Excited	Angry	Desperate	Neutral	
F01 (25)	Excited	13 (72%)	2	0	3	"duration" > "duration_noSilence" > "pitch" > "dom" > "roughness"
	Angry	4	6 (33%)	1	7	
	Desperate	0	1	12 (71%)	4	
	Neutral	5	1	1	11 (61%)	Accuracy: 59%
F02 (27)	Excited	14 (78%)	1	0	3	"duration" > "duration_noSilence" > "pitch" > "dom" > "roughness"
	Angry	1	15 (83%)	1	1	
	Desperate	0	2	9 (50%)	7	
	Neutral	2	1	4	10 (59%)	Accuracy: 68%
F03 (26)	Excited	13 (76%)	2	0	2	"Duration" > "duration_noSilence" > "roughness" > "dom" > "pitch"
	Angry	2	13 (72%)	1	2	
	Desperate	0	1	12 (86%)	1	
	Neutral	2	1	1	13 (76%)	Accuracy: 77%
F04 (27)	Excited	17 (81%)	3	0	1	"duration" > "duration_noSilence" > "roughness" > "pitch" > "dom"
	Angry	3	16 (76%)	1	1	
	Desperate	0	2	16 (84%)	1	
	Neutral	5	3	4	7 (37%)	Accuracy: 70%
M01 (24)	Excited	16 (89%)	1	0	1	"duration" > "pitch" > "duration_noSilence" > "dom" > "roughness"
	Angry	0	14 (78%)	0	4	
	Desperate	0	0	14 (78%)	4	
	Neutral	0	0	4	14 (78%)	Accuracy: 81%
M02 (28)	Excited	5 (28%)	9	2	2	"duration" > "duration_noSilence" > "roughness" > "pitch" > "dom"
	Angry	10	1 (6%)	2	5	
	Desperate	1	5	7 (39%)	5	
	Neutral	2	7	6	3 (17%)	Accuracy: 22%
M03 (23)	Excited	15 (83%)	2	0	1	"duration" > "duration_noSilence" > "pitch" > "dom" > "roughness"
	Angry	3	10 (56%)	3	2	
	Desperate	0	0	17 (94%)	1	
	Neutral	2	4	3	9 (50%)	Accuracy: 71%
M04 (22)	Excited	11 (61%)	3	1	3	"duration" > "pitch" > "duration_noSilence" > "roughness" > "dom"
	Angry	5	9 (50%)	1	3	
	Desperate	2	0	12 (67%)	4	
	Neutral	1	3	3	11 (61%)	Accuracy: 60%
M05 (22)	Excited	3 (17%)	7	0	8	"duration" > "duration_noSilence"> "roughnessVoiced" > "pitch" > "dom"
	Angry	7	3 (17%)	0	8	
	Desperate	1	0	13 (72%)	4	
	Neutral	4	6	2	6 (33%)	Accuracy: 35%