Youtube와 Emotion API를 이용한 음성 기반 감정 인식 AI 시스템

*김수연, 김은영, 구은희
아주대학교 ICT융합학과, 아주대학교 소프트웨어학과, 아주대학교 다산학부대학
e-mail : ksyee03@ajou.ac.kr, kyy8383@naver.com, ehgoo@ajou.ac.kr

Emotion Recognition Method from Speech Using the Youtube Video and Emotion API

*Soo-Yeon Kim, Eun-Young Kim, Eun-Hee Goo
Department of Software
Ajou University, Department of Software
Ajou University, Department of Dasan University College Ajou University

Abstract

By developing AI technology and cloud computing, Speech recognition technology is actively used in various fields. However, among the methods that utilize speech recognition technology, Research on emotion recognition technology through voice has not been developed much due to absent of database. The speech recognition system using speech can be applied to text - based speech recognition technology to solve the limitations of speech such as irony, and to provide a service to read and respond to user 's feelings. In this paper, we propose a database construction method using video and image emotion recognition technology for speech based emotion recognition system. Using the database and machine learning constructed in this way, the economic aspect of the voice-based emotion recognition technology system was verified. In addition, we verified the effectiveness in terms of validation by using the actual speech voice unlike the previous studies recording the voice of the actor.

I. 서론

AI 기술과 클라우딩 컴퓨팅이 발달하면서 음성 인식 기술이 크게 발달하였다. 텍스트 추출, 음성 합성 등 음성 인식에 관한 기술은 전 사업에 유용하게 활용될 수 있으므로 음성 인식 정보를 사용한 연구들은 활발하다. 음성 인식이 쓰이는 사업은 가상 비서 서비스 등 고객 맞춤형을 제공할 수 있으며, 활용 분야가 넓기 때문에 음성 기반 감정 인식 기술에 관해 연구하게 되었다.

텍스트 기반 감정 인식은 많은 연구가 진행되었으나 음성 기반 감정 인식에 관한 연구는 데이터베이스의 부재로 많이 진행되지 못했다. 기존 음성 기반 감정 인식 관련 연구는 초기 데이터를 구축할 때 여러 사람의 음성을 녹음하여 초기 데이터를 구축하였다.[1] 하지만 이러한 방법은 학습된 데이터양이 증가할수록 정확도가 높아지는 머신 러닝 모델에 사용하기에는 많은 사람의 목소리를 녹음해야 하므로 데이터 수집을 위해 비용이 많이 든다는 문제점이 있었다.

본 논문에서는 유튜브와 MS emotion cognition API 통한 경제적인 감정 기반 음성 데이터 구축 방법과 전처리된 음성 데이터를 머신 러닝을 통해 학습시키는 효과적인 음성 기반 감정 인식 기술을 제안한다.

II. 본론

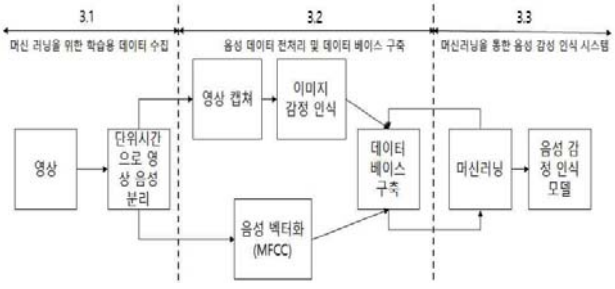
2.1 관련연구

사람의 음성에서 감정을 인식하는 기술로는 크게 두 가지가 있다. 첫째는 사람의 말을 텍스트로 바꾸어 감정 키워드에 기반을 두어 자연어 문장이 가진 의미 정보를 하는 것이다. 둘째는 사람 음성의 특징을 분석하여 감정을 인식하는 것이다. 텍스트 기반 감정인식은 현재 다양한 플랫폼에서 상용화되어 사용되고 있다. 최근 네이버 클로바 서비스와 연계된 인공지능 스피커나 시리가 그 예이다. 하지만 음성 기반 감정 인식은 아직 상용화되지 않았다. 텍스트 기반으로만 감정 인식을 하면 발생하는 대표적인 한계는 다음과 같다. 반어법으로 하는 말에 대해 감정 분석을 하지 못한다.[2] 텍스트 기반 감정 인식과 음성 기반 감정 인식을 함께 사용한다면 이 문제를 해결 할 수 있다. 하지만 아직 음성 기반 감정 인식은 상용화되지 않았다. 이 이유는 연구를 위해 데이터 베이스를 찾는 과정에 있다. 많은 양의 감정 데이터의 수집은 감정인식 시스템의 성공을 위한 필수 요건이다. 이때 수집된 데이터 베이스는 영상, 음성자료는 물론 각 자료의 감정 상태를 기술하는 라벨링을 포함한다. 이러한 라벨링은 수작업에 의해 이루어지는 데 수작업 라벨링으로 자연스러운 감정을 정확히 분류하는 것은 매우 어렵고 많은 오류를 만들 수 있으며 또한 많은 비용이 소요된다.[3] 머신 러닝은 많은 양의 이질적인 데이터로부터 결론을 내는 것이다. 따라서 데이터의 규모와 신뢰성이 중요하다.

2.2. 음성 기반 감정 인식 시스템

음성 기반 감정 인식 시스템 개요는 아래 <표1>과 같다. 유튜브를 이용하여 머신 러닝에 필요한 데이터베이스를 구현한다. 만들어진 데이터베이스를 학습시킨 음성 기반 감정 인식 모델 만든다. 이 모델에 음성이 입력되면 음성에 따른 감정이 분류되고 이 음성은 다시 분류된 감정에 따라 모델을 학습시킨다.

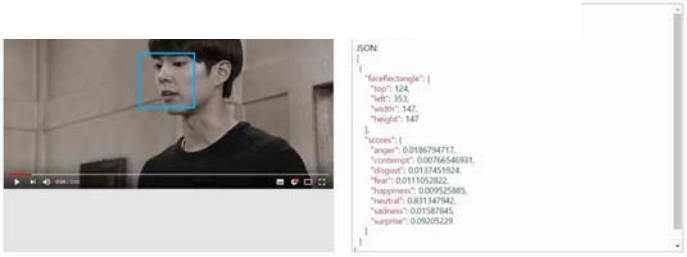
<표1. 음성 기반 감정 인식 시스템 개요>



III. 구현

3.1. 머신 러닝을 위한 학습용 데이터 수집

유튜브에서 성별, 나이를 기준으로 1인 독백 영상 10개를 얻었다. 영상을 캡처하여 음성 데이터와 이미지 데이터로 분리한다. 영상으로부터 분리한 음성 데이터를 동일하게 일정한 시간 간격으로 나눈다. 일정 시간 간격으로 나누어진 이미지 데이터는 정확도가 80~90%인 MS사에서 제공하는 MS emotion recognition API를 이용하여 감정 정보를 추출한다. MS 서버로부터 <그림1>과 같이 Jason 형태의 데이터를 얻는다. 이미지 데이터마다 추출된 감정 정보를 동일한 시간 간격으로 나눈 음성 데이터에 라벨링 한다.



<그림1. MS Emotion API의 감정 추출 결과>

3.2. 음성 데이터 전처리 및 데이터 베이스 구축

음성 데이터는 스펙트럼을 분석하여 소리의 특징을 추출하기 위해 MFCC(Mel Frequency Cepstral Coefficient) 기법을 사용한다. 수집한 음성 데이터를 20ms 단위로 잘라 MFCC 기법을 이용해 벡터로 변환한다. MFCC 기법을 이용하여 변환한 음성 데이터 벡터값을 넣을 데이터베이스를 구축한다. MS사에서 제공하는 MS emotion recognition API를 이용해 추출된 감정 정보를 <표1>과 같이 라벨링 한 후 감정에 따라 변환한 벡터값과 함께 데이터베이스에 넣는다. 이러한 방법을 이용해 <표2>과 같이 분류한 감정에 따라 MFCC 기법으로 변환된 음성 값을 넣을 초기 데이터베이스를 구축하여 5,000개의 데이터를 얻었다.

<표2. 감정 라벨링 표>

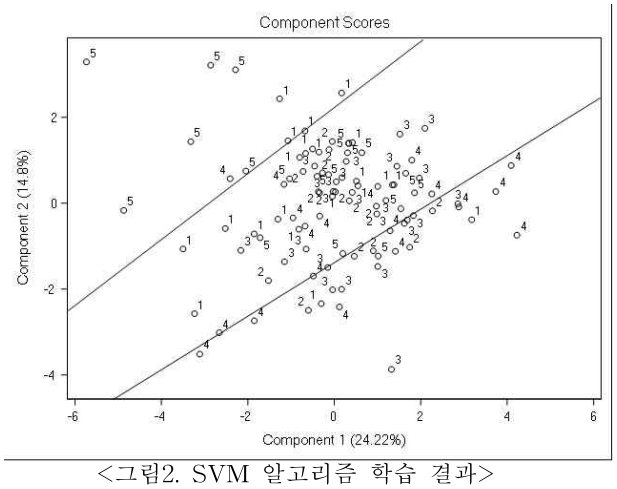
Happines	Neutral	Sadnes	Surpris	Anger
1	2	3	4	5

3.3 머신 러닝을 통한 음성 감성 인식 시스템

음성 데이터가 특징이 되고 감정 라벨을 목적 값으로 하여 지도학습인 SVM(Support Vector Machine) 모델을 만든다. 비교적 적은 학습 데이터 양으로도 좋은 분류 성능을 나타내기 때문에 본 논문에서 구현한 데이터베이스의 유효성을 확인하기에 적합하였다.[4]

음성을 입력 받으면 그 음성이 어떤 감정인지 분류해 주는 서버를 구축하였다. 클라이언트는 사용자의 음성을 단위 시간만큼 녹음한 뒤 MFCC 계수로 전처리한 후 서버 컴퓨터에 전송한다. 서버 컴퓨터는 입력된 음성 데이터를 감정에 맞게 분류하여 그 결과를 전송하고, 이 데이터는 머신 러닝을 위한 새로운 훈련 데이터가 되어 데이터베이스에 업로드 된다. 이는 서비스의 정확도가 높아지는 동시에 사용자에게 개인화된다.

<그림2>는 본 논문의 방법과 같이 구축한 음성 데이터베이스를 주성분 분석을 사용하여 2차원 평면에 나타낸 뒤 임의로 비슷한 무리로 분류한 것이다. SVM 모델은 <그림2>와 같이 특징 공간에서 비선형 경계를 찾아 분류한다. 비선형 경계이기 때문에 감정-음성 데이터가 학습됨에 따라 감정 경계선이 바뀌게 된다.



<그림2. SVM 알고리즘 학습 결과>

3.4 연구 결과 및 고찰

5,000여 개의 음성-감정 데이터를 모아 머신 러닝 모델을 만들었다. 학습된 모델의 예측과 분류 인식률을 검증하기 위해 교차 검증을 시행하였다. 100개의 데이터로 홀드 아웃 교차 검증을 추정한 결과 모델의 인식률은 64%가 나왔다.

타 연구[5]에서 감성 기반 서비스를 위한 음성 기반 감정 인식률은 62.42%이다. 이 연구의 경우 8명에게 5가지 감정을 연기를 통해 녹음하여 1,000개의 데이터를 수집하였다. 수집한 데이터는 '감정 생존 함수' 라는 심리학 연구에 기초하여 감정의 지속 시간을 고려한다. Tilted-Time Window 감정 인식의 누적된 결과를 기반으로 감정을 추정한다. 이 방법을 Tilted-Time Window 라고 한다. 이 방법을 통해 기존의 시간의 흐름을 고려하지 않은 연구 방법의 경우 평균 50.9%가 나왔던 감정인식 정확도를 평균 62.42%로 올렸다.

이 연구[5]를 본 논문과 비교할 때 음성 기반 감정 인식 모델이 연기한 음성 데이터로 학습된다는 점을 짚어볼 수 있다. 본 논문의 음성 기반 감정인식 모델은 실제 상황에서 발화되는 음성 데이터로 학습된다. 하지만 타 연구[6]에서는 이러한 데이터 수집 방식에 신뢰성 문제를 제기하고, 피실험자에게 목적하는 감정과 같은 감정을 느끼도록 상황을 설계하여 녹음하였다. 그리고 녹음한 파일을 일반인에게 평가시킨 후 데이터베이스를 구축하였다.

본 논문은 소프트웨어를 통해 말하는 사람의 얼굴로부터 감정을 읽어 내기에 객관성이 보장된다. Validation 측면에서 음성 기반 감정 인식은 실제 상황으로 학습된 본 모델이 더 효과적이다. 실제로 본 연구의 모델이 Tilted-Time Window 기법을 사용하지 않았음에도 불구하고 인식률이 더 높게 나왔다는 사실이 이를 입증한다. 또한, 향후 본 모델에 Tilted-Time Window 기법을 사용하면 상용화 할 수 있을 정도로 음성 기반 감정 인식률을 더 높일 수 있을 것이다.

5,000개의 데이터는 10개의 동영상에서 얻은 것이다. 데이터는 많으나 사람의 감정이 빠르게 변하지 않기 때문에 데이터의 다양성이 적다는 한계를 가졌다. 더 많은 영상으로부터 데이터베이스를 구축하여 모델을 학습시키면 모델의 정확도가 상승할 것이다.[7]

얼굴로부터 감정을 인식하는 것은 현재 인식률이 80%~90%로 매우 높은 편이다. 하지만 이것을 그대로 사용하면 보안 측면에서 문제가 있다. 머신 러닝을 위해 이용자를 실시간으로 모니터링 하면 사생활이 침해라는 부작용이 생긴다. 카메라가 얼굴만 찍는 것이 아니라 이용자 얼굴이 있을 수 있는 공간을 계속해서 감시한다고 할 때 문제는 더욱 심각하다. 또한, 기술 적용 부분에서도 문제가 있다. 얼굴 인식을 위해서는 일정 화소 이상의 사진 데이터가 머신 러닝 서버로 전송되어야 한다. 이 기술은 빠른 응답을 필요로 하는 데 사진은 데이터가 크기 때문에

데이터를 서버로 전송하는 데 오버로드가 일어난다. 데이터 전송에 있어 무선 인프라의 한계가 개선되기 전까지는 빠른 응답을 필요로 하는 감정 인식에는 음성을 기반으로 한 감정 인식이 적합하다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 음성으로부터 감정 인식을 하기 위해 필요한 데이터베이스 구축 방법을 제시하였다. 영상을 소리와 이미지로 분리한 후 이미지 감정 인식을 사용하여 소리에 대한 감정을 추출했다. 이렇게 구축한 데이터베이스가 음성 기반 감정 인식 머신 러닝 모델을 학습시키는 데 사용될 수 있음을 확인했다.

본 논문에서는 음성으로부터 감정인식을 하기 위해 필요한 데이터베이스 구축 방법을 제시하였다. 영상을 소리와 이미지로 분리한 후 이미지 감정 인식을 사용하여 소리에 대한 감정을 추출했다. 이렇게 구축한 데이터베이스가 음성 기반 감정 인식 머신 러닝 모델을 학습시키는 데 사용될 수 있음을 확인했다. 더불어 기존 모델과 비교할 때 학습에 쓰인 데이터가 객관적이기에 실생활에 더 적합하였고 데이터베이스를 경제적으로 구축할 수 있었다.

본 논문의 음성 기반 감정 인식률이 실생활에 적용하기 어렵다고 생각할 수 있다. 하지만 기존 연구의 음성 기반 감정 인식 알고리즘을 사용하면 인식률을 올릴 수 있다. 또한 감정 인식은 사람도 하기 어렵기에 인식률이 낮아도 서비스를 보조하는 도구로서 사용될 수 있다.

향후 연구에서는 더 많은 데이터를 모아 딥러닝을 이용하거나, Speech-to-Text 기법을 이용하여 음성 데이터를 텍스트로 변환하여 텍스트 기반 감정 인식 기술을 함께 이용한다면 감정 인식도가 높아질 것으로 예상한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 (2015-0-00908)

참고문헌

[1] 김원구. 음성 인식 정보를 사용한 감정 인식. 한국지능시스템학회 학술발표 논문집, 18(1), 425-428. 2008.

[2] Burkhardt, Felix, Alexandra Steinhilber, and Benjamin Weiss. "Ironic Speech - Evaluating Acoustic Correlates By Means Of Speech Synthesis."

[3] 김남수, “감정인식 기술의 현황과 전망,” Telecommunications Review, 제 19권, 5호, May 2009.

[4] 김창근, 박정원, 허강인. SVM 음성인식기 구현을 위한 강인한 특징 파라메터. 전자공학회논문지-SP 41.3 195-200. 2004

[5] 방재훈, 이승룡. 감성기반 서비스를 위한 통화 음성 기반 감정인식 기법. 정보과학회논문지 : 소프트웨어 및 응용, 41(3), 208-213. 2014.

[6] 손진훈, 박지은, 박정식. 자연스러운 감정표현 음성 데이터 수집과 이를 이용한 음성 기반 감정 인식. 대한인간공학회 학술대회논문집, 548-548. 2016.

[7] 박소은, 김대희, 이철, 권순일, 박능수. RNN 기반 음성 기반 감정인식 기계학습 알고리즘. 정보 및 제어 논문집, 152-153. 2017.