발화 음성을 기반으로 한 감정분석 시스템

정준혁*, 박수덕*, 김민승*, 박소현**, 한상곤***, 조우현*
*부경대학교 컴퓨터공학과
**부경대학교 인쇄정보공학과
***주식회사 페이보리
junny@pukyong.ac.kr, psd634@pukyong.ac.kr, mshkim77@naver.com
sigmadream@gmail.com, whcho@pknu.ac.kr

Context sentiment analysis based on Speech Tone

Jun-Hyeok Jung*, Soo-Duck Park*, Min-Seung Kim*, So-Hyun Park**,
Sang-Gon Han***, Woo-Hyun Cho*
*Dept of Computer Engineering, Pukyong Nat'l University
**Dept of Graphic Arts Information Engineering, Pukyong Nat'l University
***favorie Corporation

요 약

현재 머신러닝과 딥러닝의 기술이 빠른 속도로 발전하면서 수많은 인공지능 음성 비서가 출시되고 있지만, 발화자의 문장 내 존재하는 단어만 분석하여 결과를 반환할 뿐, 비언어적 요소는 인식할 수 없기 때문에 결과의 구조적인 한계가 존재한다. 따라서 본 연구에서는 인간의 의사소통 내 존재하는 비언어적 요소인 말의 빠르기, 성조의 변화 등을 수치 데이터로 변환한 후, "플루칙의 감정 쳇바퀴"를 기초로 지도학습 시키고, 이후 입력되는 음성 데이터를 사전 기계학습 된 데이터를 기초로 kNN 알고리즘을 이용하여 분석한다.

1. 서론

최근 머신러닝과 딥러닝 기술의 비약적인 발전과 함께 인공지능 분야 또한 상당히 빠른 속도로 발전하고 있으며, 그에 따라 삼성의 Bixby와 Apple의 Siri, Amazon의 Alexa 등 각 회사에서는 앞다투어 인공지능 음성 비서를 개발하고 있다.

하지만 대부분의 인공지능 음성 비서는 발화자의 문장 내 존재하는 단어들을 분석하여 단순한 결과만 반환할 뿐 인간의 복잡한 감정을 인식하지는 못하며, 이는 인공지능 비서가 사용자의 발화 내용을 파악하는데서 정확도와 이해도가 낮다는 것을 방증하는 것이기도 하다.

따라서 의사소통 연구자들의 "인간의 의사소통에서 메시지의 60~95%는 발화된 내용과 같은 언어적 요소보다 강세, 억양, 목소리의 톤 등 비언어적 요소로 전달된다."[1] 라는 연구에 따라, 발화자의 강세, 억양, 목소리의 톤 등을 수치 데이터로 변환한 후, 이를 "플루칙의 감정 쳇바퀴"를 기반으로 분석하고자 한다.

그리고 비언어적 요소에서 추출된 감정과 문맥적 요소에서 추출된 감정들을 위 의사소통 연구자들이 주장하는 가중치 비율에 따라 조합함으로써 감정 분석 결과에 대한 정확도를 높여보고자 한다.

2. 관련 연구

2.1. 감정의 분류

감정이라는 것은 상당히 주관적인 부분이어서 정의를 내리기가 어려우나, 학자들은 감정을 정신활동을 통해 고도의 즐거움과 불만으로 실체화 되는 의식의 경험으로 정의하고 있으며[2][3][4], 그 구성요소를 인지판단, 신체증상, 기분 등을 포함한 5가지로 분류하였다.[5]

2.2. Robert Plutchik's Wheel of Emotion

1980년도 Robert Plutchink에 의해 제안된 Wheel of Emotions는 인간의 감정을 8개의 기초 감정으로 분류할 뿐만 아니라, 각 기초 감정의 강도와 여러 기초 감정과의 합성을 통해 더욱 세밀한 감정을 표현할 수 있는 감정 분류 모델이다.[6]

본 연구에서는 Plutchik의 Wheel of Emotions를 기초로 하여 발화자의 비언어적 요소 중 톤에 대한 감정과 문맥 내 존재하는 감정단어에 대한 감정을 분류하는 시스템을 구성한다.

*본 논문은 2017년 한이음 ICT멘토링 프로젝트의 결과물입니다.

<그림 1> Plutchik의 Wheel of Emotion

2.3. 음성의 특징 추출 방법

음성은 인간이 정보를 전달하기 위해 사용하는 가장 보편적인 수단으로, 최근에는 사람 간의 정보 전달 뿐만 아니라 음성 인식 기술을 내장하고 있는 가전제품, 전자제품, 차량 등의 기능을 사용하는 등. 그 사용 영역이 점차 확대되고 있다.

특히 위의 음성 인식 기술을 구현하기 위해서는 음성 신호로부터 특징을 추출 가능한 알고리즘을 사용하여야 하며, 과거부터 멜 캡스트럼, 루트 캡스트럼, HMM 등의 알고리즘이 제안되었다.[7][8]

또한 실질적으로 음성으로부터 특징을 추출하기 위해서는 마이크로부터 입력되는 노이즈와 주변의 배경음을 제거할 수 있어야하기에, 이를 제거하고 음성만을 깨끗하게 추출하기 위해서는 필터의 사용이 필수적이다.[9][10]

2.4. Mel-Frequency Ceptrum Coefficients

MFCC(Mel-Frequency Cepstrum Coefficients)는 신호에 FFT(Fast-Fourier Transform)을 적용하고, 신호의 세기를 Mel 단위로 변환한 후, 각각의 mel 단위의 신호에 log 함수를 적용하고, 마지막으로 DCT(Discrete Cosine Transform)를 적용한 것으로, 신호 내 구성요소들의 크기를 가지는 스펙트로그램 그래프이다. 보통 음성의 특징을 추출하여 음성인식에 이용되거나, 음악의 특성을 추출하여 음악의 장르 등을 알아낼 때 사용한다.[11]

2.5. kNN 분류 알고리즘

kNN (k-Nearest Neighbors)알고리즘은 지도 학습에서 자주 사용되는 데이터 분류 기법으로, 유클리드 거리, 맨하탄 거리 등의 그래프 상에서 최 근접(Nearest) 이웃(Neighbors)을 찾기 위한 알고리즘을 사용하여, 다수결에 따라 데이터를 분류하는 모델이다.[12]

본 연구에서는 앞서 이뤄진 음성 신호의 주파수 구성 요소의 분석과 MFCC 분석 등을 통해 도출된 음향적 특징을 해당 음성에 대한 감정과 함께 미리 학습시킨 후, 이후 입력되는 음성 데이터를 분류한다.

2.6. Messaging Queue

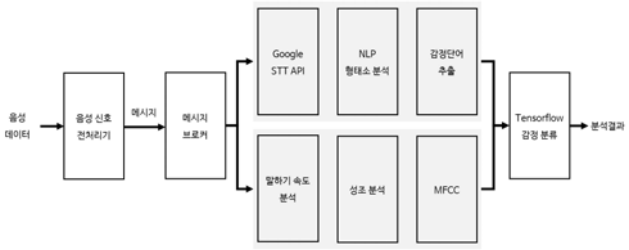
메시징이란 우편 또는 이메일과 비슷한 방식으로서, 중간에서 중계자 역할을 하는 브로커 프로세스가 큐를 생성한 후, 발행자가 생성한 메시지를 소비자에게 라우팅 및 전달 해주는 방식을 의미한다.[13][14]

메시징 방식은 클라이언트-서버 구조와 같이 동기 방식으로 상호작용 하는 것 보다 더욱 유연한 비동기 시스템을 구성할 수 있다.

본 연구에서는 음성 신호로부터 특징을 추출하고, 특징에 대한 감정을 추출하는 부분과 유저와 상호작용 하는 부분을 서로 비동기적으로 처리하고, 추후 서버의 유연성 또한 높일 수 있도록 RabbitMQ 메시징 큐 브로커 서비스를 사용하였다.

3. 톤에 대한 감정 분류 시스템 설계 및 구현

3.1. 전체 시스템 구성



<그림 2> 시스템 구조도

전체 시스템은 위의 그림 2와 같이 구성하였으며, 업로드 된 음성에 대해 먼저 전처리 과정을 거치고, 메시징 큐로 “감정 단어 분석 요청” 메시지와 “음성 특징 분석 요청” 메시지를 발행한다. 발행된 메시지들은 메시지 브로커를 통해 각 프로그램으로 전달되며, 두 프로그램은 동시에 비동기적으로 실행된다.

먼저 문맥 내 감정단어 추출 프로그램의 경우, 먼저 Google의 STT(Speech To Text) API를 통해 음성을 텍스트 데이터로 변환하며, NLP 분석 과정을 통해 텍스트를 형태소 단위로 분리한다. 그리고 감정 단어를 추출하여, 감정을 분석한 후, 음성 데이터로부터 감정을 추출할 때 가중치로 사용한다.

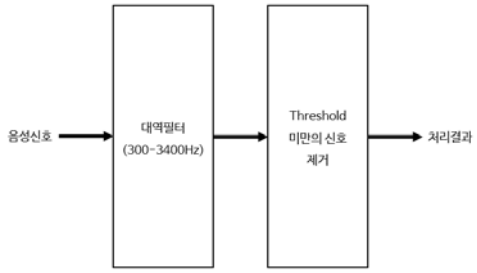
음성 데이터로부터 특징을 추출하는 프로그램은 먼저 말하기 속도를 분석하고, FFT(Fast-Fourier Transform) 과정을 통해 신호의 요소를 분리하여, 전체적인 성조와 변화량을 추출한다.

그리고 MFCC 분석을 통해 특징 벡터를 추출하고 텐서플로우를 통해 해당 값과 앞의 감정단어로부터 추출된 가중치를 이용하여 감정을 분류한 후, 결과를 반환한다.

3.2. 음성 데이터 전처리기 설계

앞의 2.3절에서 언급한 연구에 따라, 본 연구에서는 시끄러운 환경에서도 입력된 음성 신호로부터 더욱 정확한 감정을 추출하기 위해 먼저 신호를 사람의 음성 대역(300~3400Hz)만을 통과시키도록 설계된 대역 필터(Bandpass Filter)에 입력하고, 그 결과에서 미세한 잡음

또한 제거하기 위해, 전체 음량에서 지정된 Threshold 미만의 데이터를 제거하는 과정을 전처리 과정으로 설계하였다.



<그림 3> 음성의 전처리 과정

3.3. 음성 특징 추출

입력된 음성 데이터로부터 특징을 추출하기 위해서 NumPy, SciPy 라이브러리를 사용하였으며, 먼저 입력받은 데이터를 ffmpeg를 사용하여 wav파일로 변환한 후, SciPy를 통해 리샘플링 하였다. 그리고 앞의 3.2절에서 설계하였던 전처리기를 Scipy의 Butter worth filter를 사용하여 구현하였다.

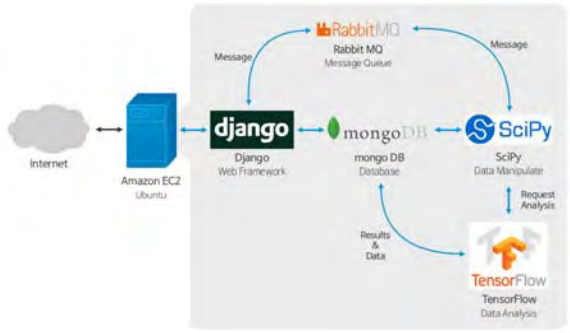
또한 위 결과를 SciPy 내에서 음성 처리를 위한 함수만 선별해놓은 구현체인 librosa 라이브러리를 사용하여 음성의 특징을 추출하고, Mel-Frequency Ceptrum 분석을 진행하였다.

3.4. 기계학습 및 감정 분류기

위의 알고리즘 통해 추출된 음성의 특징을 Plutchik의 Wheel of Emotions를 기반으로 분류하고, 이를 Google의 Tensorflow를 이용하여 지도학습 시켰으며, 이후 분석을 위해 입력되는 음성은 앞서 지도 학습한 데이터들을 기반으로 kNN(K-Nearest Neighbors) 알고리즘을 이용하여 분류하였다.

4. 구현결과

4.1. 구현환경



<그림 4> 전체 서버 시스템의 구조

전체적인 시스템은 Apple의 Siri, Samsung의 Vixby등을 모티브로 하여, 웹 사이트에서 음성을 녹음하여 전송하거나 파일을 업로드 할 수 있도록 설계하였으며, 서비스는 Python 3 언어를 사용하여 작성되었다.

먼저 음성의 녹음과 업로드 등의 유저가 직접 사용하는 Frontend 부분은 Django를 이용하여 구현되었으며, Django의 uWSGI를 사용하여, 웹 서버인 Nginx가 요청을 받아, Django로 전달 및 Django의 응답을 대신하도록 하였고, Nginx와 클라이언트의 연결에 HTTPS를 적용함으로서 보안을 더욱 향상시켰다.

음성이 업로드 되고 처리하는 과정에서 사용자에게 처리과정을 실시간으로 보여주고, 상태 업데이트 함께 다른 작업 또한 함께 진행하기 위해 RabbitMQ 메시징 큐를 사용하여 비동기적으로 처리되도록 구현하였으며, 업로드 된 음성과 분석 결과를 저장하기 위해 MongoDB를 사용하였다.

4.2. 구현결과

본 연구에서는 발화 음성에서 추출 된 특징으로부터 감정을 추론할 수 있는 시스템을 개발했다. 먼저 국내의 드라마와 예능 프로그램으로부터 약 1600개 정도의 샘플 음성 데이터를 추출하였으며, 각 음성 데이터에 대한 감정을 Plutchik의 Wheel of Emotions을 이용하여 분류한 후, 그 중 1200개를 학습시켰다.

나머지 400개를 알고리즘의 테스트를 위해 위의 시스템에 입력 및 분석해본 결과, 아래의 표와 같은 정확도를 얻을 수 있었다.

<표 1 > 각 감정에 대한 정확도

감정 명	정확도
Rage	86.65%
Vigilance	67.12%
Ecstasy	58.71%
Admiration	60.11%
Terror	79.84%
Amazement	74.68%
Grief	82.43%
Loathing	84.58%

Rage, Grief, Loathing, Terror 등. 부정적인 감정에 관련된 감정 요소들의 정확도는 높게 나온 반면, Ecstasy, Admiration 등 긍정적인 감정과 관련된 감정 요소들의 정확도는 비교적 낮았다.

해당 결과에 대해 그 원인을 면밀히 파악해본 결과, 학습을 위해 입력한 데이터의 1200개 중 약 38%는 Rage가 포함되어 있는 것으로 파악되었으며, 특히, Rage와 Loathing이 함께 나타나는 감정이 많이 존재하였다.

이는 부정적인 감정에 대해 학습 데이터가 Over Fitting 된 것으로 추정되며, 각 감정 분류에 대한 정확도를 개선하고 Over Fitting 문제를 해결하기 위해서는 더욱 다양한 데이터를 학습시켜야할 것이다.

5. 결론 및 향후연구

본 논문에서는 기계학습을 통해 음성 신호로부터 추출된 특징을 Plutchik의 Wheel of Emotions을 기반으로 분류할 수 있는 시스템을 제안하였다. 비록 Over Fitting 문제가 발생하였으나, 실험 결과 중 샘플 데이터가 충분하였던 감정인 Rage의 경우 86.65%의 정확도를 보였으며, 이는 감정을 매우 정확하게 분류한다고 볼 수 있다.

향후 연구에서는 rNN 유전 알고리즘을 사용하여 적은 학습 양이더라도 정확도가 더욱 향상된 감정 분석 결과를 도출해 내도록 시스템을 개선할 수 있겠다.

또한, 본 연구에서 이용한 발화자의 톤뿐만 아니라 다른 비언어적 요소인 표정과 몸짓 등으로 부터 감정을 추출하고 분석 및 분류할 수 있는 시스템에 대해 연구를 수행할 것 이다.

참고문헌

[1] Birdwhiste, R. (1972). “Paralanguage twenty-five years after Sapir.” Communication in face. to face Interaction, ed. by J. Laver/S. Hutchenson, Penquin.

[2] Damasio, AR (1998). “Emotion in the perspective of an integrated nervous system.”. Brain research. Brain research reviews. 26 (2-3): 83 - 6.

[3] Davidson, edited by Paul Ekman, Richard J. (1994). “The Nature of emotion : fundamental questions” New York: Oxford University Press. pp. 291 - 93.

[4] Cabanac, Michel (2002). “What is emotion?” Behavioural Processes 60(2): 69-83.

[5] Scherer, K. R. (2005). “What are emotions? And how can they be measured?” Social Science Information. 44 (4): 693 - 727.

[6] Plutchik, R. (2001), “The Nature of Emotions” American Scientist vol. 89. pp. 344-350.

[7] J. S. Choi (Oct. 2016), “Recognition Algorithm using MFCC Feature Parameter”, Proceedings of Conference on Information and Communication Engineering Vol. 20, No. 2, pp. 343.

[8] 박재홍, 이광호, 안동순. (1999). “HMM에 기반 음성인식을 위한 Toolkit의 구성요소” 한국정보과학회 학술발표논문집, 26(1B), 472-474.

[9] J. S. Choi (2012), “Classification Algorithm of Male and Female in the Noisy Environment”, Journal of KIIT, Vol. 10, No. 11, pp. 63-68.

[10] S. F. Boll (1979), “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoust., Speech, Signal Processing, Vol. 27, No.2, pp. 113-120.

[11] Min Xu; et al. (2004). “HMM-based audio keyword generation” In Kiyoharu Aizawa; Yuichi Nakamura; Shin'ichi Satoh. Advances in Multimedia Information Processing - PCM 2004: 5th Pacific Rim Conference on Multimedia.

[12] Altman, N. S. (1992). “An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician 46 (3): 175 - 185.

[13] Vinoski, S. (2006). "Advanced Message Queuing Protocol". IEEE Internet Computing. 10 (6): 87 - 89.

[14] Curry, Edward. (2004). “Message-Oriented Middle - ware” In Middleware for Communications, ed. Qusay H Mahmoud, 1-28. Chichester, England: John Wiley and Sons.