음성특징의 다양한 조합과 문장 정보를 이용한 감정인식

서승현, 이보원

인하대학교 전자공학과

seosh7039@gmail.com, bowon.lee@inha.ac.kr

Emotion Recognition using Various Combinations of Audio Features and Textual Information

Seunghyun Seo and Bowon Lee

Department of Electronic Engineering

Inha University

요 약

본 논문은 다양한 음성 특징과 텍스트를 이용한 멀티 모드 순환신경망 네트워크를 사용하여 음성을 통한 범주형(categorical) 분류 방법과 Arousal-Valence(AV) 도메인에서의 분류방법을 통해 감정인식 결과를 제시한다. 본 연구에서는 음성 특징으로는 MFCC, Energy, Velocity, Acceleration, Prosody 및 Mel Spectrogram 등의 다양한 특징들의 조합을 이용하였고 이에 해당하는 텍스트 정보를 순환신경망 기반 네트워크를 통해 융합하여 범주형 분류 방법과 과 AV 도메인에서의 분류 방법을 이용해 감정을 이산적으로 분류하였다. 실험 결과, 음성 특징의 조합으로 MFCC Energy, Velocity, Acceleration 각 13 차원과 35 차원의 Prosody 의 조합을 사용하였을 때 범주형 분류 방법에서는 75%로 다른 특징 조합들 보다 높은 결과를 보였고 AV 도메인 에서도 같은 음성 특징의 조합이 Arousal 55.3%, Valence 53.1%로 각각 가장 높은 결과를 보였다.

1. 서론

최근 딥러닝을 이용한 연구 중 인간의 감정을 분류하는 감정인식은 영상, 음성, 생체 신호등을 이용한 다양한 분야에서 활발하게 연구가 되고 있으며 이들을 융합한 멀티 모드 감정인식 연구 또한 활발하게 진행되고 있다 [1].

이 중 음성을 통한 감정 인식을 위해서는 음성 특징과 더불어 문장 정보를 이용한 멀티 모드 딥러닝 네트워크를 이용해 감정을 인식하는 연구들이 좋은 결과를 보이고 있다 [2]. 음성을 통한 딥러닝 연구에서는 다양한 음성 특징을 음성으로부터 추출하여 이를 단일 입력으로 사용하거나 조합하여 사용한다 [3]. 이러한 특징들을 이용해 감정을 분류하는 방법으로는 '행복, 슬픔, 화남, 역겨움, 공포감, 좌절감, 흥분됨' 등의 이산적인 분류 결과를 추론하는 범주형 분류 방법과 더 나아가 2 차원 Arousal-Valence(AV) 도메인에서 Valence(1-부정적, 5-긍정적), Arousal (1-고요함, 5-흥분됨) 로 결과를 추론하는 방법이 있다 [4]. 본 논문은 음성 신호로부터 다양한 특징들을 추출하여 이를 텍스트 정보와 융합하여 범주형 분류방법과 과 AV 도메인에서의 분류 방법을 통해 감정을 분류하는 방식을 제안한다. 본 연구를 통해 딥러닝을 통한 감정인식을 위해 사용된 특징 조합 중에서는 MFCC Energy(e), Velocity(d), Acceleration(dd) 및 Prosody 특징들을 이용한 특징조합이 범주형 분류 방법 에서는 75.3%로 가장 좋은 결과를 보였고, AV 도메인 에서는 Arousal, Valence 가 각각 55.3%와 55.1%로 같은 특징 조합을 사용하였을 때 가장 높은 결과를 보였다.

2. 감정 분류 방법

감정을 분류하는 방법은 감정을 '행복, 슬픔, 화남, 역겨움, 공포감, 좌절감, 흥분됨' 등으로 분류하는 범주형(categorical) 분류 방법이 대표적이다.

이 외에도 최근 연구에서는 감정을 2 차원으로 분류하기 위해 감정을 Valence(1-부정적, 5-긍정적), Arousal (1-고요함, 5-흥분됨) 두 가지로 나누어 (-2, -1, 0, 1, 2) 5 개의 스케일로 각각 분류하는 방법이 있다 [4].

3. 네트워크 구조

네트워크 모델로는 음성과 문장 각각의 모드를 위해 모두 순환신경망 네트워크를 사용하였다. 음성을 이용한 네트워크에서는 음성에서 추출한 MFCC 데이터와 Prosody 데이터를 합쳐 입력으로 받고 시퀀스 데이터를 각 순환신경망 노드를 거쳐 출력 값을 저장한다. 문장을 이용한 네트워크 에서는 Natural Language Toolkit (NLTK)를 이용하여 문장 정보를 인코딩 하고, 각 단어를 임베딩 하여 순환신경망 노드의 인풋으로 입력 받아 결과를 출력한다. 마지막으로, 각 모드에서 출력된 결과값들을 합친 후 두 개의 완전 연결망의 인풋으로 사용하여 감정의 확률 값으로 나타낸다 [2].

본 논문에서는 두 가지 분류 방법을 위해 결과를 출력하는 마지막 단을 범주형 분류 방법에서는 7 가지 감정에 대해 확률 값을 출력하는 완전 신경망을 사용하였고, AV 도메인 분류를 위해서는 Arousal, Valence 각 두 영역의 확률 값을 출력하는 완전 신경망을 사용하였다.

표 1. <음성 특징 조합에 따른 감정 분류 정확도 결과>

Classification method	Features	Accuracy
범주형	e+d+dd+prosody	75.3%
범주형	mel+d+dd+prosody	73.4%
범주형	mel+d+dd+e+prosody	72.3%
Arousal	e+d+dd+prosody	55.3%
Arousal	mel+d+dd+prosody	52.2%
Arousal	mel+d+dd+e+prosody	53.4%
Valence	e+d+dd+prosody	53.1%
Valence	mel+d+dd+prosody	51.4%
Valence	mel+d+dd+e+prosody	50.8%

4. 실험

본 실험을 위해 데이터베이스는 IEMOCAP 데이터베이스[5]를 사용하였고 총 10039 개의 발화 문장 중 노이즈를 제거한 5539 개 문장을 사용하였다 [2]. 범주형 분류를 위해 4 개의 감정을 사용하였고, AV 도메인 실험에서도 같은 발화 문장들을 이용해 각 Arousal(1-5) Valence(1-5)로 이산적인 값을 추출하였다. 음성 특징 추출을 위해서는 MFCC Energy, Velocity, Acceleration 세 가지를 각각 13 차원으로 추출하였고 Mel Spectrogram 은 40 차원 그리고 Prosody 정보는 35 차원으로 추출하여 이 다섯 가지를 조합하여 음성 데이터 특징으로 사용하였다. 문장 임베딩은 각 단어를 300 차원으로 임베딩 하여 사용하였고 두 가지를 각각 두가지의 네트워크 입력으로 입력 받아 나온 출력 값을 합쳐 완전 연결망을 이용해 범주형 분류 방법과 AV 도메인에서의 분류 방법으로 분류하였다. 각각의 분류 방법에 대해서 1 차원 4 가지 감정에 대해 75.8%의 정확도를 얻을 수 있었고 AV 도메인 에서는 Arousal 과 Valence 에 대해서 각각 55.3%와 53.4%의 결과를 얻을 수 있었다.

5. 결론

본 논문의 실험 결과로 음성을 통한 감정인식에서 음성 특징을 조합하여 사용하는 것이 단일 음성 특징을 사용하는 것 보다 높은 성능을 보였고 그 중 MFCC Energy(e), Velocity(d), Acceleration(dd)의 조합이 가장 높은 성능을 보였다. 하지만 음성 특징을 더 많이 사용한다고 해서 더 좋은 연구 결과를 보이지는 않았다.

이 후 연구를 통해 더 다양한 음성 특징들을 이용해 감정인식에 더 효과적인 특징들에 대해 연구하고, AV 도메인에서의 데이터셋의 불균형을 해결하고 영상, 생체 등의 도메인을 추가 적용하여 모드를 확장하면 더욱 감정인식의 성능을 높일 수 있을 것으로 보인다.

참고문헌

[1] Michael Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, Maja Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge", in AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge Pages 3-10

[2] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 112-118.

[3] P. Khunarsal, C. Lursinsap, T. Raicharoen "Very short time environmental sound classification based on

spectrogram pattern matching" in Information Sciences, 2013 – Elsevier

[4] Xingfeng Li, Masato Akagi "A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech" in INTERSPEECH 2018

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S. Narayanan,

"Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.