오디오 데이터에 FFT를 이용한 특징 추출과 이를 이용한 오디오-텍스트 멀티모달 모델

(심사용 논문양식에는 이름, 소속, 이메일 주소를 기입할 수 없으며,
기입할 경우 심사 시 불이익을 받을 수 있습니다.)

audio-text multimodal model from Feature extraction using FFT for audio data

(심사용 논문양식에는 이름, 소속, 이메일 주소를 기입할 수 없으며,
기입할 경우 심사 시 불이익을 받을 수 있습니다.)

요 약

MFCC는 사람이 음성을 인지하고 파악하는 과정을 모사하기 위해 개발된 훌륭한 특징 추출 방법이다. 하지만 멀티모달을 위한 Feature Engineering에서 인간 음성 이외의 잡음의 특징은 제대로 활용하지 못할 수 있다. 따라서 본 논문은 FFT를 소개하며, MCB를 이용하여 오디오 데이터와 텍스트 데이터의 특징 벡터를 생성하여 감정 정보를 예측한다. 그리고 MFCC를 사용하여 MCB를 통해 특징 벡터를 생성하여 감정 정보를 예측했을 때의 멀티모달 모델과 성능을 비교 검증하였다. 실험 결과, FFT를 사용한 특징 벡터 생성 방법이 더 높은 성능을 보였다.

1. 서 론

딥러닝 알고리즘이 발전하고 컴퓨팅 성능이 발달함에 따라 인간 작업의 많은 영역에 인공지능이 도입되고 있다. 그와 더불어 인간 감정을 자동으로 인지하는 모델에 대해서도 활발한 연구가 진행되고 있다. 예를 들어, 오디오 신호로부터 감정을 인식하는 연구[1]라든가 텍스트 데이터로부터 감정 정보를 파악하는 연구[2]가 대표적이다. 최근에는 이를 이용한 멀티모달 연구가 활발히 진행되면서 기존의 단일 데이터 분석으로 복잡한 감정 표현을 고려하지 못하는 분석 방법에 한계점을 극복하고자 노력하고 있다.

멀티 모달을 위한 Feature Engineering에서 오디오 데이터를 분석하기 위한 전처리에서 문제가 발생한다. Mel-Frequency Cepstral Coefficients(MFCC)는 음성인식 분야에서 사용되는 음성 특징 추출 기술 중 하나로, MFCC는 주파수 스펙트럼의 정보를 이용하여 특징을 추출한다. 하지만 사용할 데이터의 길이마다 임베딩 데이터가 다르게 설정되는 문제가 존재한다. 많은 연구에서 오디오 데이터의 길이에 따라 길이가 긴 데이터는 동일한 길이로 자른다든지, 길이가 부족한 데이터는 패딩을 추가하는 방식으로 데이터를 전처리한다. 즉, 데이터의 길이가 길어지면 임베딩 데이터의 차원이 커지고, 짧아지면 임베딩 데이터의 차원이 작아지는 것이다. 이어서 현재의 Feature Engineering은 인간의 의사소통 상호작용을 제대로 고려하지 못한다는 한계점을 소개하고자 한다[3].

사람이 음성을 인지하고 파악하는 과정은 다음과 같다. 먼저 청각을 수용한다. 청각 기관으로부터 들어오는 소리를 수용하고 이 과정에서 음성의 파형이 고주파와 저주파로 분리되어 각각의 주파수 대역을 각기 다른 세포군에서 처리한다[4].

두 번째로 들어온 소리가 언어인지 여부를 판단하고, 어떤 언어인지, 누가 말하는 지를 인지한다. 그리고 들어온 언어를 문장 단위로 분리하고. 각 단어의 의미와 문맥을 이해한다. 이 과정에서 문법, 의미, 상황 등을 고려하여 문장을 이해한다.

세 번째로 음성 파악을 진행한다. 분석된 문장을 바탕으로 들어온 음성의 의도와 내용을 파악한다. 이 과정에서 상황, 문맥, 말하는 사람의 감정 등을 고려하여 의도와 내용을 알아낸다.

MFCC는 사람이 음성을 인지하고 파악하는 과정을 모사하기 위해 개발된 훌륭한 특징 추출 방법이다. 하지만 Mel 척도를 사용하여 주파수 대역을 분석하기 때문에, 인간 음성 이외의 잡음에 민감할 수 있다[5]. 그러나 텍스트 데이터에는 감정인식에 도움되는 오디오가 포함될 수 있으며, 그런 상황 또한 오디오 데이터에 포함될 수 있다[6]. 즉 MFCC는 음성 데이터 전처리에 유용한 방법 중 하나이지만, 분석 목적과 데이터 특성에 따라 적절한 방법을 선택해야 한다.

Fast Fourier Transform는 주파수 대역 뿐만 아니라,

시간 정보까지 고려할 수 있어. 오디오 신호의 다양한 정보를 추출할 수 있다. 이러한 장점 때문에, 일부 논문에서는 MFCC의 한계를 극복하기 위해 FFT를 사용하는 방법도 제안되고 있다[7]. 본 논문은 오디오 데이터를 FFT를 통해 전처리하고, 이를 임베딩한 텍스트 데이터와 결합하여 Multimodal Compact Bilinear(MCB) 방법을 사용하여 특징 벡터를 생성하는 감정 분류 모델을 제안한다.

2. 연구 방법

2.1. 데이터 전처리

본 논문에서는 KEMDy19[6]를 사용했다. 2개 이상의 감정 Label은 “;” 앞의 감정으로 통일했다. annotation file의 wav_start, end를 기준으로 차이가 30초 이상의 데이터는 제거했으며, ’을’, ’를’, ’이’, ’가’, ’은’, ’는’, ’에’와 같은 불용어를 제거했다. 하지만 ‘c/’, ‘n/’과 같은 음성 이외 사운드 상황에 대한 태깅 등은 제거하지 않고 텍스트 데이터로서 사용하였다.

MFCC와 FFT를 이용한 데이터 추출을 하기 위해 오디오 데이터는 sampling rate, 22500으로 load했으며, MFCC는 30초 이상의 데이터 중 가장 긴 데이터를 기준으로 padding하여 (128, 138)의 shape로 불러왔으며, FFT는 768의 데이터 크기로 불러왔다.

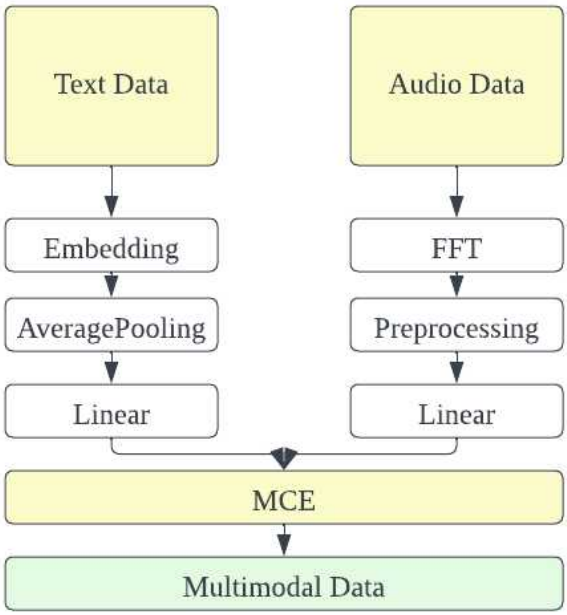
전처리 이후 데이터의 분포는 <표 1>과 같다. 대다수의 데이터 Emotion이 ‘Natural’인 불균형한 데이터이다.

<표 1> 전처리 이후 데이터의 Emotion 분포

Emotion	Nunber of sample
Natural	4303
Happy	1911
Angry	1441
Surprise	997
Sad	766
Fear	413
Disgust	398

2.2. 멀티모달 모델

본 연구는 오디오 데이터를 FFT를 통해 전처리하고, 이를 임베딩한 텍스트 데이터를 MCB 방법을 사용하여 결합하고 감정 정보를 예측한다. <그림 1>은 오디오와 텍스트가 서로 다른 정보이지만, MCB 방법을 사용하기 위해 오디오는 FFT를 사용한 후 전처리를 통해 구조를 만들고 텍스트는 임베딩을 사용한 후, GlobalAveragePooling를 통해 구조를 만들었다. 이러한 구조를 통해, 한 문장이 MCB 방법을 통해 특징벡터가 생성되는 것을 확인할 수 있다.



<그림 1> 오디오-텍스트 멀티모달 모델의 아키텍처

3. 실험 및 결과

3.1. 실험 세팅

본 논문의 텍스트, 음성, 멀티모달 기반 모델에 사용된 배치(batch) 크기는 64, 학습률(learning rate)은 학습횟수(epoch)는 10, 옵티마이저(optimizer)는 adam을 사용하였다. 실험에 사용된 GPU는 rtx4080, 11.8의 CUDA version을 사용하였다.

MCB는 두 가지 종류의 데이터를 하나의 특징 벡터로 만들어 줌으로써, 다양한 정보를 함께 고려할 수 있게 한다. 이는 두 가지 데이터의 차원을 줄여서 모델의 계산량을 줄이고, 과적합을 방지하는 데도 도움이 된다[8].

3.2. 실험 결과

<표 2>는 성별을 분류하기 전, 오디오 데이터에 MFCC를 이용해 특징을 추출한 데이터와 텍스트 데이터의 멀티모달 모델과, 오디오 데이터에 FFT를 이용해 특징을 추출한 데이터와 텍스트 데이터의 멀티모달 모델의 성능을 비교한 것이다.

<표 3>은 성별을 분류한 후에 각각의 모델 성능을 비교한 것이다.

<표 2> 발화자의 감정 평가 결과

Input Data	Precision	Recall	F1-score
MFCC	0.31	0.23	0.43
FFT	0.21	0.19	0.43
MFCC+Voice	0.64	0.59	0.67
FFT+Voice	0.64	0.53	0.67

<표 3> 발화자의 성별별 감정 평가 결과

Input Data	Sex	Precision	Recall	F1-score
MFCC	Male	0.28	0.23	0.42
	Female	0.37	0.26	0.48
FFT	Male	0.20	0.18	0.43
	Female	0.25	0.19	0.46
MFCC+Voice	Male	0.65	0.58	0.66
	Female	0.69	0.59	0.68
FFT+Voice	Male	**0.63**	**0.63**	**0.68**
	Female	**0.69**	**0.63**	**0.71**

<표 2>의 발화자의 감정 평가 결과에서는 MFCC-텍스트 멀티모달 모델이 비슷하게 나온 반면, <표 3>의 발화자의 감정 평가 결과에서는 FFT-텍스트 멀티모달 모델이 2~3% 가량 결과가 더 좋게 나왔다. 이는 사람의 남성과 여성의 음성 데이터에서 주파수 영역이 크게 다른 특징[9]이 FFT 전처리 데이터에서 MFCC 전처리 데이터보다 좀 더 두드러지게 표현된 것이라 추론할 수 있다.

4. 결론

본 연구는 텍스트 데이터에 포함된 비언어적 표현을 이용하고자 제안된 Feature Engineering 방법이다. 기존의 멀티모달 감정인식 방법에서 사용되는 MFCC를 이용하여 특징 추출하는 방법과 FFT를 사용하여 특징을 추출하는 방법을 비교하여, 후자의 방법을 사용하면 결과가 5% 정도 개선됨을 실험을 통해 확인하였다.

이를 통해 텍스트 데이터와 음성 데이터를 결합할 때, 음성 데이터에 포함된 상황의 노이즈와 텍스트 데이터 안의 비언어적 표현이 발화자의 감정인식에 도움을 줄 수 있음을 확인하였다. MFCC보다 더 높은 성능을 보여주는 FFT 기반의 특징 추출 방법을 사용함으로써, 감정인식 분야에서 보다 정확한 결과를 도출할 수 있음을 제시하였다.

참 고 문 헌

[1] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345. (2019).

[2] Nandwani, P., & Verma, R. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(1), 81. (2021).

[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80. (2001).

[4] Gazzaniga, M. S. The split brain in man. Scientific American, 217(2), 24-29. (1967).

[5] Singh, N., Khan, R. A., & Shree, R. MFCC and prosodic feature extraction techniques: a comparative study. International Journal of Computer Applications, 54(1). (2012).

[6] Noh, K. J., Jeong, C. Y., Lim, J., Chung, S., Kim, G., Lim, J. M., & Jeong, H. Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets. Sensors, 21(5), 1579. (2021).

[7] Moritz, N., Anemüller, J., & Kollmeier, B. Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5492-5495). IEEE. (2011, May).

[8] Pham, N., & Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 239-247). (2013, August).

[9] Titze, I. R. Physiologic and acoustic differences between male and female voices. The Journal of the Acoustical Society of America, 85(4), 1699-1707. (1989).