The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.6, pp.284~290

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.6.284

Spontaneous Speech Emotion Recognition Based On Spectrogram With Convolutional Neural Network

Guiyoung Son† · Soonil Kwon††

ABSTRACT

Speech emotion recognition (SER) is a technique that is used to analyze the speaker's voice patterns, including vibration, intensity, and tone, to determine their emotional state. There has been an increase in interest in artificial intelligence (AI) techniques, which are now widely used in medicine, education, industry, and the military. Nevertheless, existing researchers have attained impressive results by utilizing acted-out speech from skilled actors in a controlled environment for various scenarios. In particular, there is a mismatch between acted and spontaneous speech since acted speech includes more explicit emotional expressions than spontaneous speech. For this reason, spontaneous speech-emotion recognition remains a challenging task. This paper aims to conduct emotion recognition and improve performance using spontaneous speech data. To this end, we implement deep learning-based speech emotion recognition using the VGG (Visual Geometry Group) after converting 1-dimensional audio signals into a 2-dimensional spectrogram image. The experimental evaluations are performed on the Korean spontaneous emotional speech database from AI-Hub, consisting of 7 emotions, i.e., joy, love, anger, fear, sadness, surprise, and neutral. As a result, we achieved an average accuracy of 83.5% and 73.0% for adults and young people using a time-frequency 2-dimension spectrogram, respectively. In conclusion, our findings demonstrated that the suggested framework outperformed current state-of-the-art techniques for spontaneous speech and showed a promising performance despite the difficulty in quantifying spontaneous speech emotional expression.

Keywords : Spontaneous Speech, Speech Emotion Recognition, Spectrogram, Convolutional Neural Network

CNN 기반 스펙트로그램을 이용한 자유발화 음성감정인식

손 귀 영† · 권 순 일††

요 약

음성감정인식(Speech Emotion Recognition, SER)은 사용자의 목소리에서 나타나는 떨림, 어조, 크기 등의 음성 패턴 분석을 통하여 감정 상태를 판단하는 기술이다. 하지만, 기존의 음성 감정인식 연구는 구현된 시나리오를 이용하여 제한된 환경 내에서 숙련된 연기자를 대상으로 기록된 음성인 구현발화를 중심의 연구로 그 결과 또한 높은 성능을 얻을 수 있지만, 이에 반해 자유발화 감정인식은 일상생활에서 통제되지 않는 환경에서 이루어지기 때문에 기존 구현발화보다 현저히 낮은 성능을 보여주고 있다. 본 논문에서는 일상적 자유발화 음성을 활용하여 감정인식을 진행하고, 그 성능을 향상하고자 한다. 성능평가를 위하여 AI Hub에서 제공되는 한국인 자유발화 대화 음성데이터를 사용하였으며, 딥러닝 학습을 위하여 1차원의 음성신호를 시간-주파수가 포함된 2차원의 스펙트로그램(Spectrogram)로 이미지 변환을 진행하였다. 생성된 이미지는 CNN기반 전이학습 신경망 모델인 VGG (Visual Geometry Group) 로 학습하였고, 그 결과 7개 감정(기쁨, 사랑스러움, 화남, 두려움, 슬픔, 중립, 놀람)에 대해서 성인 83.5%, 청소년 73.0%의 감정인식 성능을 확인하였다. 본 연구를 통하여, 기존의 구현발화기반 감정인식 성능과 비교하면, 낮은 성능이지만, 자유발화 감정표현에 대한 정량화할 수 있는 음성적 특징을 규정하기 어려움에도 불구하고, 일상생활에서 이루어진 대화를 기반으로 감정인식을 진행한 점에서 의의를 두고자 한다.

키워드 : 자유발화, 음성감정인식, 스펙트로그램, 합성곱신경망

1. 서 론

지난 수년간 코로나19 사태를 겪는 동안, 사회 전반에 걸쳐 비대면 환경으로 전환되면서 온라인 강의, 화상회의 등 비대면 학습/업무 처리가 증가하게 되었다. 이로 인하여 상대방의 상태를 파악하고 인지하기 위하여 음성에 집중하게 되었고, 상대방의 음성을 통하여 현재 상황, 감정 분석 등을 통하여 비대면 환경에서 활용 가능한 음성기반 다양한 기술개발의 수요가 점차 증가하고 있다[1].

음성감정인식(speech emotion recognition, SER)은 사용자의 목소리에서 나타나는 떨림, 어조, 크기 등의 음성 패턴

※ 이 논문은 2021년도 세종대학교 교내연구비 지원에 의하여 연구되었음 (No.20211105).
† 비 회 원 : 세종대학교 소프트웨어학과 연구교수
†† 종신회원 : 세종대학교 소프트웨어학과 교수
Manuscript Received : May 2, 2024
Accepted : May 20, 2024
* Corresponding Author : Soonil Kwon(skwon@sejong.edu)

분석을 통하여 감정 상태를 판단하는 기술이다. 최근에는 전화 금융사기(voice phishing), 보험 청구 사기 등과 같은 사용자의 감정 상태가 표면적으로 파악하기 어려운 환경일 때, 음성정보에 의존하여 발화자의 감정 및 의도를 분석하여 범죄를 예방하는 기술과도 접목되어 긍정적인 효과를 얻고 있다[2].

최근 인공지능 기술개발과 함께 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network), LSTM(Long-Short Term Memory)과 같은 다양한 딥러닝 모델을 활용한 음성 감정인식연구가 활발히 진행되면서 우수한 성능을 보여주고 있다[2-7]. 기존의 음성기반 감정인식 연구는 구현된 시나리오를 제시하고, 숙련된 연기자를 대상으로 음성을 기록하여 감정인식을 진행하였다. 연기자에 의해 연기된 음성은 제한된 환경 내에서 기록되므로, 명확하게 감정이 표현된 음성을 사용했기 때문에 높은 성능을 얻을 수 있었다. 하지만, 연기된 감정은 종종 과장된 표현을 포함하기 때문에, 일상생활에서 이루어지는 자유발화 감정표현과는 매우 다르다. 이러한 까닭에, 자유발화 감정표현이 구현발화의 감정표현을 반영하지 못하기 때문에 구현발화 감정인식에 활용된 음성적 특징들은 자유발화에 활용하기에는 한계가 존재한다.

본 연구에서는 일상생활에 이루어지는 자유발화 음성 활용한 감정인식을 진행하였다. 즉, 1차원 음성신호를 이용하여 딥러닝 모델 중 하나인 CNN을 활용하여 감정분류를 진행하고 그 성능을 분석하였다. 실험에 사용된 데이터는 AI Hub[3]에서 제공되는 한국인 자유발화 대화 음성 데이터베이스로 성인과 청소년으로 구성되어있다. 먼저, 딥러닝 학습을 위하여 1차원의 음성신호를 스펙트로그램(spectrogram)의 2차원의 시간-주파수가 포함된 이미지로 변환하였고, CNN 전이학습 모델 가운데 이미지 인식에 뛰어난 VGG (Visual Geometry Group)를 사용하여 학습한 후, 그 성능을 평가하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존의 음성감정인식에 대한 선행연구를 설명한다. 3장에서는 딥러닝 모델 학습 및 검증을 위해 사용되는 CNN에 대한 개념과 구조를 설명한다. 4장에서는 딥러닝 학습을 위한 실험환경과 진행 과정을 설명하고, 5장에서는 실험 결과 및 성능을 분석한다. 마지막으로 6장에서는 결론 및 향후 연구 방향에 관해 서술한다.

2. 관련 연구

음성기반 감정인식 관련 국내외 연구는 주어진 상황에 가공된 음성인 구현발화를 중심으로 높은 감정인식 결과를 보여주고 있다[4-6](Table 1).

[4]에서는 IEMOCAP(interactive emotional dyadic motion capture database)을 사용하여 감정인식을 진행하였다. 그들은 1차원의 음성을 2차원의 스펙트로그램으로 이미지 변환 후, Time-frequency CNN-ELM(Extreme-Learning Machine)을 활용하여 4개의 감정에 대하여 70.78%의 인식성능을 보여주었다. [5]에서는 3개의 감정 음성 데이터베이스를 사용하여 스펙트로그램으로 이미지화한 후, 감정인식을 진행하였다. 기존 CNN의 확장된 CNN layer인 dilated CNN을 활용하였고, 그 결과로 EMO-DB로 7개 감정에 대하여 93.00%의 높은 감정인식 정확도를 얻었다. [6]에서는 TCN(Temporal Convolutional Network) 구조를 사용하여 심층 학습모델을 제안하였고, IEMOCAP, EMO-DB를 활용하여 각각 80.84%, 92.31%의 감정인식 성능을 확인하였다.

현재 구현발화 감정인식 성능은 실생활에 활용 가능할 수준까지 향상되었고, 국제적인 연구 결과에서도 높은 성능에 도달했음을 알 수 있다. 이에 반해 자유발화는 일상생활에 통제되지 않는 환경에서 이루어지며, 감정표현 및 방해요인이 동시에 발생하기 때문에 감정인식 성능에서 높은 성능을 보여주는 연구가 아직 저조하다[7-10](Table 1).

[7]에서는 CNN과 양방향 LSTM을 활용하여 특징 수준(feature-level)과 모델 수준(model-level)에서 특징추출을 융

Table 1. Previous Speech Emotion Recognition Experimental Results

Expression	Paper	Database	Emotion	Feature	Method	Accuracy(%)
Acted	[3]	IEMOCAP	Anger, Disgust, Happiness, Neutral	Spectrogram	Time-frequency CNN + ELM	70.78
	[4]	IEMOCAP EMO-DB RAVDESS	Happiness, Sadness, Fear, Anger Disgust, Borden, Neutral	Spectrogram	dilated CNN	78.01 93.00 80.00
	[5]	IEMOCAP EMO-DB	Anger, Disgust, Happiness, Neutral/ Happiness, Sadness, Fear, Anger, Disgust, Excited, Neutral	Raw speech signals	Temporal Convolutional Network (TCN)	80.84 92.31
Spontaneous	[7]	AFEW5.0	Anger, Joy, Sadness, Disgust, Fear, Surprise, Neutral	spectral related low-level audio feature descriptors (LLDs)	CNN-BLSTM	35.51
	[8]	BAUM-1	Anger, joy, sad, disgust, fear, surprise	Spectrogram	CNN	48
	[9]	BAUM-1	Anger Joy, Sadness, Disgust, Fear, Surprise	Mel-spectrogram	Multi-CNN	44.6
	[10]	BAUM-1	Anger, Joy, Sadness, Disgust, Fear, Surprise	Mel spectrogram	CNN+LSTM	53.98

합하는 방법을 사용하는 알고리즘을 제안하였다. 그 결과, EmotiW2018(Emotion Recognition in the Wild), AFEW (Acted Facial Expressions in the Wild) 을 사용한 성능평가에서 7개 감정에 대하여 36.61%의 성능을 보여주었다. [8]에서는 자유발화 데이터인 BAUM-1를 사용하여 낮은 레벨의 음성특징 요소를 활용하여 CNN을 활용하여 성능평가를 진행하였다. 그 결과 7개의 감정에 있어서 48%의 정확도를 확인하였다. [9]에서는 자유발화 데이터베이스인 BAUM-1를 사용하여 Mel-Spectrum 특징을 추출한 후, Multi-CNNs 모델로 학습을 진행하였다. 그 결과, 6개 감정에 대해서 44.6%의 성능을 보여주었다. [10]에서는 CNN+LSTM을 이용하여 6개 감정에 대해서 53.98%의 성능을 확인하였다.

구현발화 감정인식은 기존 딥러닝 모델을 활용하여 다양한 모델 조합을 활용하여, 국제적으로 공인된 데이터를 사용하여 80% 이상 수준의 인식률을 보였다. 하지만, 자유발화는 최근 연구 결과에서도 구현발화보다 현저히 낮음을 알 수 있다. 이는 자유발화에 대한 국제적으로 공인된 데이터가 많지 않고, 구현발화와 달리 통제되지 않은 환경이라는 점에서 나타나는 불명확한 감정표현 및 부정확한 발음과 주변잡음 등의 방해요인들로 인해 낮은 성능을 보였다.

본 논문에서는 지금까지의 구현발화 감정인식 모델을 자유발화 모델에 적응하기에는 한계가 있음을 파악하고, 이에 대한 보완을 통하여 자유발화 감정인식 성능향상을 시도하고자 하였다. 한국어 자유발화 음성데이터를 사용하여, 음성신호 변화를 주파수 대역으로 이미지로 표현되는 스펙트로그램으로 변환 후, 영상처리에서 활용되는 CNN 전이학습 VGG를 사용하여 감정인식을 진행하였다.

3. 제안방법

스펙트로그램(spectrogram)은 1차원의 음성데이터에 내재된 시간의 흐름에 따른 주파수 분포의 특징요소를 이미지로 표현할 수 있기 때문에 음성기반 분류학습에 많이 사용되고 있으며, CNN은 이미지 학습에 우수한 성능을 보여주고 있다. 본 장에서는 딥러닝 학습을 위하여 사용된 스펙트로그램과 CNN에 대하여 설명한다.

3.1 스펙트로그램(spectrogram)

스펙트로그램(spectrogram)은 단시간 푸리에 변환(Short Time Fourier transform, STFT) 기법을 사용하여 파형으로 표현되는 음성데이터에서 시간 축의 구간을 아주 짧은 단위로 나누어 푸리에 변환을 적용하여 시간 도메인을 유지하면서 주파수 도메인 정보를 구하는 기법으로 음성 및 신호의 세기가 시간에 따라서 각 주파수 대역마다 변화하는 것을 시각적으로 표현한 것이다[11]. 음성데이터의 스펙트럼 특징을 사용하는 것이 일반적이며, 이를 이용하여 감정인식을 진행하여 그 효과도 증명되었다[3,4,11,12]. 본 연구에서는 CNN 전이학습을 사용하기 위하여 음성데이터의 특징을 이미지화하는 과정이 필요하므로, 1차원의 음성신호를 시간-주파수로 표현되는 2차원의 스펙트로그램(spectrogram)으로 변환하였다[11]. 1개의 음성신호에 대하여 슬라이딩 윈도우(sliding window)을 적용하지 않고, 단일의 이미지로 추출했다. 감정별 추출한 이미지는 Fig. 1과 같다.

3.2 Convolutional Neural Network(CNN)

딥러닝 모델 중 하나인 CNN은 데이터 특징을 추출하여 특징들의 패턴을 파악하는 구조로 되어있다[13]. 여기에서 합성곱 계층(convolution layer)은 다양한 필터를 사용하여 이미지의 특징을 도출하고, 풀링 계층(pooling layer)은 이미지의 특징을 유지하면서 차원을 축소한다. 기존 CNN 모델은 주로 컴퓨터비전 분야에서 이미지를 인식하는 데 많이 사용되고 있지만, 최근에는 1차원 음성신호와 같은 시계열 데이터의 특징을 추출하기 위해서도 널리 사용되고 있다[3,4,8,9].

본 논문에서는 CNN 전이학습 모델 중 이미지 분류에 우수한 정확도를 제공하는 VGG를 사용하였다[13,14]. Fig. 2를 보면, VGG는 총 16개의 층(layer)으로, 특징을 추출하는 13개의 합성곱(convolution) 계층과 추출한 크기의 특징을 줄여주는 3개의 풀링(pooling) 계층으로 구성되어있다.

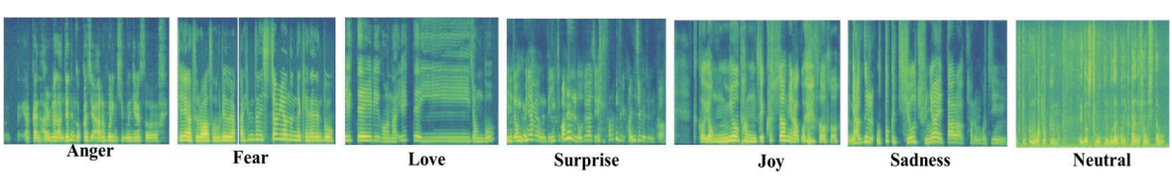

Fig. 1. Spectrogram Example

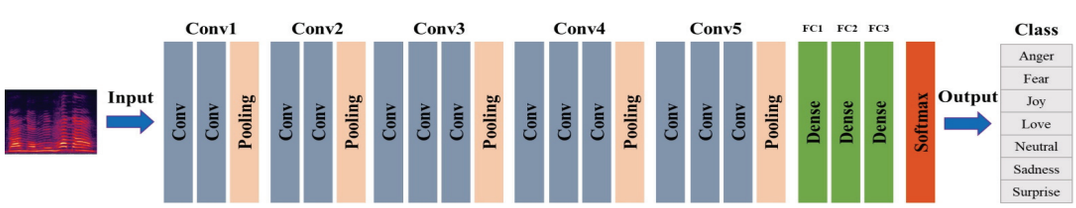

Fig. 2. CNN Architecture

VGG는 네트워크의 깊이와 모델 성능 영향에 집중한 모델로 네트워크의 깊이가 깊어질수록 이미지 분류정확도가 높아지게 된다[14]. 이러한 연구 결과에 따라 최근 음성기반 감정인식에서도 1차원의 음성신호를 2차원 시간-주파수의 이미지로 변환시킨 후 CNN 전이모델에서도 음성신호 처리가 가능하게 되었다[15,16]. 특히, VGG를 활용한 연구로 우수한 성능을 보여주었으므로 본 연구에서 VGG를 선정하여 실험을 진행하였다[17,18].

4. 실험환경

4.1 실험데이터

본 연구에서는 자유발화 감정음성 데이터는 AI Hub에서 제공되는 '감정이 태깅된 자유대화(성인, 청소년)' 한국인 대화 음성데이터를 학습 및 평가에 사용하였다[19, 20]. 데이터는 한국어를 모국어로 하는 청소년과 성인을 대상으로 수집되었고, 2인으로 구성된 발화자들의 대화를 수집한 음성(16kHz, 16bit)으로, .wav 형식으로 구성되어있다. 발화자들은 교육, 건강, 취미 등 총 11개의 주제에 따라, 분노(anger), 두려움(fear), 기쁨(joy), 사랑스러움(love), 중립(neutral), 슬픔(sadness), 놀람(surprise)의 7가지 감정에 대하여 자유롭게 녹음하였다.

성능평가를 위하여 총 데이터의 80%는 학습을 위한 학습데이터로 사용되었고, 20%는 성능평가를 위한 검증데이터로 나누어 실험을 진행하였다. 실험에 사용된 데이터 정보는 Table 2와 같다.

4.2 실험환경

음성신호를 이용한 감정인식 성능평가를 위하여 전이학습 모델인 VGG로 실험을 진행하였다. 본 연구에서는 성능향상을 위하여 파인튜닝(fine tuning)을 하였으며, 사용한 하이퍼파라메터(hyperparameter)는 Table 3과 같다. 이와 더불어, 더 정확하고 신뢰성 있는 성능평가를 수행하기 위해 5겹 교차검증(5-fold cross validation)을 적용했다. 먼저, 학습은 총 50회 수행했으며, 학습률(learning rate)을 성인과 청소년 각 0.00005, 0.000006으로 설정하였다. 활성화 함수로는 ReLu (rectified linear unit)함수를 사용하였고, 마지막 완전연결계층(fully connected layer)는 다중 클래스 분류 문제에 많이 쓰이는 Softmax 함수를 사용하였다. 옵티마이저(optimizer)는 Adam을 사용했다. 모든 실험은 아나콘다(anaconda)와 파이썬(python) 환경에서 진행되었다. 실험에 사용한 딥러닝 프레임워크(framework)는 텐서플로우(tensorflow)[21]를 사용하였고, 그래픽처리장치(graphic processing unit, GPU)는 NVIDIA Geforce 3080, 20GB를 사용하였다.

4.3 평가지표(Evaluation metrics)

성능평가를 위하여 성능 지표로써 평가샘플 전체에 대한 정확도 weighted average(WA)와 클래스마다 성능 평균 unweighted accuracy(UA)를 사용하였다[22]. 또한, 감정인식 결과의 분포를 분석하기 위하여 혼동행렬(confusion matrix)을 출력하고 예측, 실제 성능으로 나누어 분류성능을 비교 분석하였다.

5. 실험결과

5.1 성인

Table 4는 성인 자유발화 음성의 7개 감정에 대한 VGG를 활용한 감정인식 결과로 정밀도, 재현율, F1-score를 보여준다. 표에 따르면, 7개의 감정인식에 대하여 85%의 분류정확도를 확인할 수 있다. 각 감정에 대해서는 행복, 중립, 슬픔의 순으로 가장 높은 정확도를 얻었으나, 이에 반해 사랑스러움에 대한 감정에 대해서는 가장 낮은 정확도를 보였다.

Table 3. Hyperparameter Setting

Parameter	Value
Batch size	128
Learning rate	0.00005(Adult)
	0.000006(Young)
Epoch	50
Dropout	0.5
Optimizer	Adam
Activation function	ReLu

Table 2. The Description of all the Spontaneous Emotional Speech Data

Emotion	Group(N)					
	Adult			Young		
	Train	Test	Total	Train	Test	Total
Anger	7,074	1,769	8,843	9,033	2,259	11,292
Fear	6,036	1,510	7,546	9,751	2,438	12,189
Joy	9,668	2,418	12,086	9,245	2,397	11,642
Love	2,793	699	3,492	8,633	1,584	10,217
Neutral	9,636	2,409	12,045	9,551	2,388	11,939
Sadness	9,557	2,390	11,947	8,775	1,239	10,014
Surprise	3,043	761	3,804	8,081	1,890	9,971
Total	**49,831**	**9,932**	**59,763**	**39,100**	**14,195**	**53,295**

Table 4. The Class-wise Emotion Recognition Results in Term of Precision, Recall, Weighted, Unweighted, and F-1 Score for Adult

Emotion	Precision(%)	Recall(%)	F1-Score(%)
Anger	0.78	0.81	0.80
Fear	0.84	0.71	0.77
Joy	0.91	0.94	0.93
Love	0.76	0.69	0.72
Neutral	0.93	0.90	0.92
Sadness	0.81	0.87	0.84
Surprise	0.77	0.81	0.79
Weighted	**0.85**	**0.85**	**0.85**
Unweighted	**0.83**	**0.82**	**0.82**

Table 5. The Class-wise Emotion Recognition Results in Term of Precision, Recall, Weighted, Unweighted, and F1-score for Young

Emotion	Precision(%)	Recall(%)	F1-Score(%)
Anger	0.70	0.68	0.69
Fear	0.68	0.68	0.68
Joy	0.76	0.78	0.77
Love	0.72	0.62	0.67
Neutral	0.82	0.83	0.83
Sadness	0.64	0.69	0.67
Surprise	0.72	0.75	0.73
Weighted	**0.73**	**0.73**	**0.73**
Unweighted	**0.72**	**0.72**	**0.72**

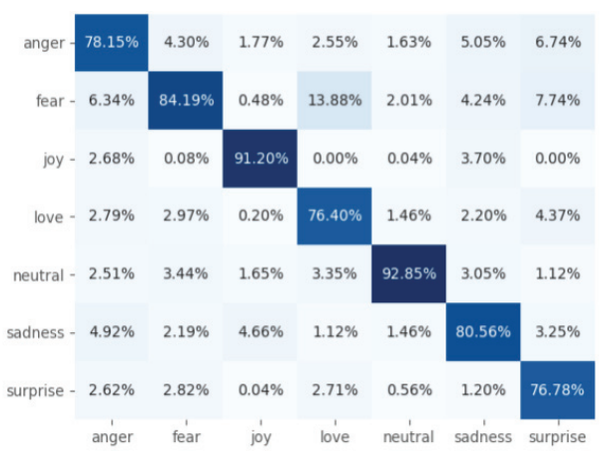

Fig. 3. A Confusion Matrix Between the Actual and Predicted Labels for Adult

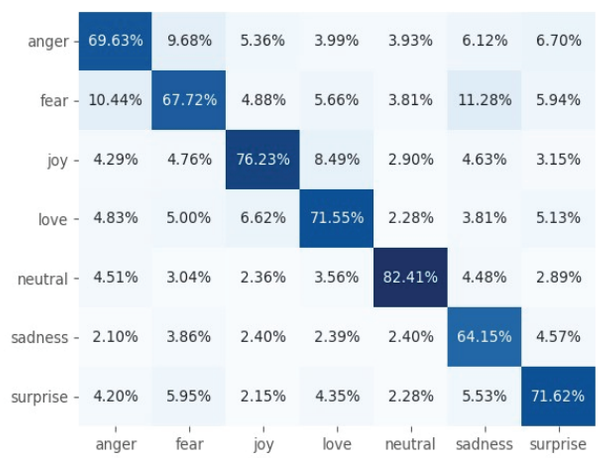

Fig. 4. A Confusion Matrix Between the Actual and Predicted Labels for Young

Fig. 3을 보면, 성인의 자유발화 음성에 대한 7개의 감정인식 결과를 혼동행렬(confusion matrix)로 나타내는 것으로, 혼동행렬재현율이 중립 92.85%, 행복 91.20%, 두려움 84.19% 순서로 높은 성능을 보여주었다. 그러나, 평가 감정 중 사랑스러움에 대한 재현율은 69%로 가장 낮게 예측되었다. 이는, 사랑스러움은 사회적 감정 중 하나로, 일반적으로 연인, 가족 등 특정 상대와 대화에서 많이 나타나고, 성별, 대상의 연령 등에 따라 표현방식도 상이하므로 높은 성능을 얻을 수 없는 요인이 될 수 있다 이와 더불어, 사랑스러움은 놀람의 감정으로 오분류가 가장 많이 되었다. 이는 일반적으로 놀람은 긍/부정 범주에 포함되지 않는 중립으로 높은 각성의 상태에서 감정이 발현된다고 볼 수 있다. 이는, 과한 사랑스러움의 표현이 일부 높은 각성에서 나타날 수 있는 놀람으로 표현되었을 가능성도 있다. 또한, 사랑스러움은 데이터 수의 측면에서도 다른 감정보다 데이터의 불균형으로 나타났기 때문에, 이로 인하여 성능에 영향이 미칠 것으로 추측해 볼 수 있다. 즉, 학습데이터 수가 적은 사랑스러움은 학습 횟수가 상대적으로 작아서 학습률이 저조하기 때문이라고 유추해 볼 수 있다.

5.2 청소년

Table 5는 청소년의 자유발화 데이터를 이용하여 7개의 감정에 대해 감정인식 결과로 정밀도, 재현율, F1-score를 보여준다. 청소년은 전체 감정에 대해서 73%의 감정인식률을 얻었으며, 각 감정별로 보면, 중립, 행복, 놀람의 감정에서 높은 정확도를 얻었으며, 이에 반해 슬픔에 대한 감정에 대해서는 가장 낮은 성능을 보였다.

Fig. 4를 보면, 청소년 자유발화에 대하여 7개의 감정인식 결과를 혼동행렬(confusion matrix)로 나타내는 것으로, 그 결과로 혼동행렬재현율이 중립 82.41%, 행복 76.23%, 놀라움 71.62% 순서로 높은 성능을 보여주었다. 그러나, 슬픔과 두려움 감정에 대해서는 각각 64.15%, 67.72%로 가장 낮게 예측되었다. 이를 추측해 보면, 슬픔과 두려움의 경우에는 음성 발화시 나타날 수 있는 슬픔으로 인한 목소리의 흐느낌과 두려움에서 나타날 수 있는 목소리의 떨림 현상이 감정인식을 서로 방해하는 요인으로 작용하였을 것으로 추측해 볼 수 있다.

성인과 청소년의 자유발화에 대한 감정인식률을 보면, 청소년이 성인과 비교해 전체적으로 감정인식에 대한 성능이 낮

음을 확인할 수 있다. 이는 청소년은 감정을 표현에 있어서, 성인보다 감정을 표현하는 방식, 강도 등이 서툴고, 일련의 규정되지 않은 다양한 표현으로 개인 간의 편차가 더 크게 나타난 것으로 추측해 볼 수 있을 것이다.

6. 결 론

본 논문은 한국어 자유발화 음성데이터를 활용하여 CNN 전이학습 모델인 VGG를 활용하여 감정인식 성능을 평가하였다. 결과적으로, 7개 감정(기쁨, 사랑스러움, 화남, 두려움, 슬픔, 중립, 놀람)에 대해서 성인 83.5%, 청소년 73.0% 감정인식 분류정확도를 확인하였다. 이는 기존 국제적으로 연구된 자유발화 감정인식보다 성능이 우수하다는 것을 확인할 수 있었다. 기존 구현발화기반 감정인식보다는 성능이 낮지만, 자유발화는 감정에 대한 명확한 음성적 특징 및 감정표현에 대한 객관적인 규정이 어려움에도 불구하고, 일상생활에서 자유롭게 수집된 음성을 활용했다는 점에서 유의한 결과라 할 수 있다.

향후 연구에서는 자유발화에서 나타나는 감정표현에 대한 보다 정확하고 명확한 표현방식과 카테고리의 재정립이 필요하다. 이와 더불어, 다양한 데이터와의 비교분석을 통하여 성능검증과 최신 딥러닝 모델 활용을 통한 성능향상을 기대할 수 있을 것이다. 또한, 최근 활발하게 연구되고 있는 텍스트, 얼굴, 생체신호 등과 융합을 통한 멀티모달 감정인식을 통하여 성능향상을 도모해 볼 수 있을 것이다.

References

[1] Eunji Lee(2022). ASTI MARKET INSIGHT 67: Speech recognition service.

[2] Integrated Data Analysis Center, “Development of the world’s first ‘voice phishing voice analysis model’,” Ministry of the Interior and Safety, 2023.02.22.

[3] AI 기술 및 제품·서비스 개발에 필요한 AI 통합 플랫폼 [Internet], https://aihub.or.kr/

[4] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, “Speech emotion recognition with local-global aware deep representation learning,” In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7174-7178), IEEE, 2020.

[5] S. Kwon, “MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach,” *Expert Systems with Applications*, Vol.167, pp.114177, 2021.

[6] M. Ishaq, M. Khan, and S. Kwon, “TC-Net: A modest & lightweight emotion recognition system using temporal convolution network,” *Computer Systems Science & Engineering*, Vol.46, No.3, pp.3355-3369, 2023.

[7] J. Cai, et al., “Feature-level and model-level audiovisual fusion for emotion recognition in the wild,” In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp.443-448). IEEE, 2019.

[8] G. Chen, S. Zhang, X. Tao, and X. Zhao, “Speech emotion recognition by combining a unified first-order attention network with data balance,” *IEEE Access*, Vol.8, pp.215851-215862, 2020.

[9] S. Zhang, X. Tao, Y. Chuang, and X. Zhao, “Learning deep multimodal affective features for spontaneous speech emotion recognition,” *Speech Communication*, Vol.127, pp.73-81, 2021.

[10] A., Amjad, L., Khan, N., Ashraf, M. B., Mahmood, and H. T. Chang, “Recognizing semi-natural and spontaneous speech emotions using deep neural networks,” *IEEE Access*, Vol.10, pp.37149-37163, 2022.

[11] Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, Vol.20, No.1, pp.183, 2019.

[12] M. Khan, M. Ishaq, M. Swain, and S. Kwon, “Advanced sequence learning approaches for emotion recognition using speech signals,” *In Intelligent Multimedia Signal Processing for Smart Ecosystems* (pp.307-325). Cham: Springer International Publishing, 2023.

[13] H. C. Shin, et al., “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Transactions on Medical Imaging*, Vol.35, No.5, pp.1285-1298, 2016.

[14] O. Russakovsky, et al., “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, Vol.115, pp.211-252, 2015.

[15] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.22, No.10, pp.1533-1545, 2014.

[16] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” *IEEE Access*, Vol.7, pp.19143-19165, 2019.

[17] A. Aggarwal, et al., “Two-way feature extraction for speech emotion recognition using deep learning,” *Sensors*, Vol.22, No.6, pp.2378, 2022.

[18] S. Akinpelu, S. Viriri, and A. Adegun, “Lightweight Deep Learning Framework for Speech Emotion Recognition,” *IEEE Access*, Vol.11, pp.77086-7709, 2023.

[19] 감정이 태깅된 자유대화(성인) [Internet], https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71631

[20] 감정이 태깅된 자유대화(청소년) [Internet], https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71632
[21] M. Abadi, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprintarXiv: 1603.04467*, 2016.
[22] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," In *Australasian Joint Conference on Artificial Intelligence* (pp.1015-1021). Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

손 귀 영

https://orcid.org/0000-0003-0505-5863
e-mail : sgy1017@sejong.ac.kr
2015년 성균관대학교 독어독문학과(석사)
2020년 연세대학교 인지과학협동과정 (컴퓨터과학전공)(박사)
2015년 ~ 2018년 세종대학교 디지털콘텐츠학과 전임연구원
2018년 ~ 2023년 세종대학교 소프트웨어학과 선임연구원
2023년 ~ 현 재 세종대학교 소프트웨어학과 연구교수
관심분야 : Brain-computer Interfaction(BCI), Human Computer Interfaction(HCI), Affective Computing

권 순 일

https://orcid.org/0000-0001-5451-8815
e-mail : skwon@sejong.edu
2000년 Southern California University, Electrical Engineering(석사)
2005년 Southern California University, Electrical Engineering(박사)
2005년 ~ 2006년 삼성전자 책임연구원
2006년 ~ 2009년 한국과학기술연구원 선임연구원
2009년 ~ 현 재 세종대학교 소프트웨어학과 교수
관심분야 : Audio Digital Signals Processing, Human-computer Interaction(HCI), Speech Recognition, Affective Computing, Audio Processing