한국어 음성 감정 인식 전처리 방식에 따른 감정 분류 성능 비교 연구

윤선민, 이다빈, 황병일, 김동주, 서영주, 김다현*
포항공과대학교 인공지능연구원
ciot28@postech.ac.kr, leedb@postech.ac.kr, bihwang@postech.ac.kr,
kkb0320@postech.ac.kr, yjsuh@postech.ac.kr, *kdhyun8011@postech.ac.kr

A Comparative Study of Emotion Classification Performance by Preprocessing Methods in Korean Speech Emotion Recognition(SER)

Yoon Sun Min, Lee Da Bin, Hwang Byeong Il, Kim Dong Ju, Suh Young Joo, Kim Da Hyun*
POSTECH Institute of Artificial Intelligence

요 약

음성 감정 인식은(Speech Emotion Recognition, SER)은 음성 신호로부터 화자의 감정 상태를 추정하는 기술로 최근 의료, 교육, 고객 서비스 등 다양한 분야에서 중요한 역할을 한다. 음성 감정 인식 분야에서는 대부분 영어 발화 중심으로 많은 연구가 이루어져왔지만, 실제로 영어와 한국어의 리듬, 강세 등 발화 패턴이 상이하므로 한국어 도메인에 특화된 더 많은 연구가 필요한 상황이다. 대부분의 기존 연구에서는 음성 전처리 방식에 있어 멜-스펙트럼 켑스트럼 계수, 멜 스펙트로그램, 웨이블릿 변환 등 주로 사용되는 전처리 방식들이 다양하게 존재한다. 이에 본 연구는 동일한 조건에서 5개의 감정(기쁨, 슬픔, 분노, 불안, 중립)으로 구분된 한국어 음성 데이터셋에 대하여 기존의 다양한 전처리 방식을 각각 적용하여 비교·검증 하였다. 특징 추출과 감정 분류 모델로는 ResNet 모델을 통일하게 사용하였다. 결과는 멜 스펙트로그램(97.15%)과 STFT(97.08%) 두 가지 전처리 방식을 사용하였을 때 가장 우수한 정확도 성능을 보였다.

Ⅰ. 서 론

음성 감정 인식(Speech Emotion Recognition, SER)은 음성 신호로부터 화자의 감정 상태를 추정하는 기술로 인간-컴퓨터 상호작용, 고객 서비스, 정신 건강 모니터링 등 다양한 분야에서 응용되고 있다.[1]. 또한 최근 사람과 직접 상호작용하는 인공지능의 발달에 따라 인간의 감정을 이해하는 기술의 수요가 빠르게 늘고 있는 추세이다[1]. 그럼에도 불구하고 다수의 선행 연구는 영어 발화를 중심으로 수행되어 왔으며, 한국어에 특화된 체계적인 비교 연구는 상대적으로 미흡하다[2]. 언어별 운율 구조와 감정 표현 방식의 차이를 고려하면, 영어에서 유효한 시스템이 한국어에서도 최선이라 보기 어렵다. 기존 음성 감정 인식(SER) 연구는 대체로 제한된 음성 전처리 방식 한두 가지를 채택하는 것이 일반적이었으며, 활용된 데이터가 연구마다 달라 음성 신호 전처리 방식에 대한 명확한 비교가 어렵다. 이에 본 연구는 한국어 환경에서 사용 가능한 음성 감정 인식 분야에서의 전처리 방식을 동일한 조건에서 비교함으로써, 언어적 특수성을 반영하고 실무 적용 가능성을 높이는 방안을 제안하고자 한다.

Ⅱ. 관련 연구

음성 감정 인식(SER)은 일반적으로 1) 음성 신호 입력, 2) 전처리, 3) 특징 추출, 4) 감정 분류 단계를 수행한다.

전처리 단계에서는 입력 음성(16kHz 단일 채널)에 대한 길이 정규화를 수행한다. 일반적으로 음성 신호의 표현으로 자주 사용되는 방식은 STFT(Short-Time Fourier Transform), 멜 스펙트로그램(Mel-spectrogram), 멜-주파수 켑스트럼 계수(Mel-Frequency Cepstrual Coefficient, MFCC), 연속 웨이블릿 변환(Continuous Wavelet Transform, CWT)등이 존재한다[3][4][5]. 특징 추출과 감정 분류 단계에서는 서포트 벡터 머신(SVM), 랜덤 포레스트(Random Forest), LSTM(Long Short-Term Memory), CNN(Convolutional Neural Network)등 다양한 모델이 사용되며, 특히 시간-주파수 형태의 2차원의 특징 스펙트럼의 경우, 일반적으로 CNN 기반의 분류 모델이 채택된다[6].

영어 중심의 음성 감정 인식 연구는 IEMOCAP, RAVDESS, EMO-DB 등의 대규모 데이터가 표준적으로 사용되어왔다[6]. 이러한 연구에서 전처리 과정은 주로 멜 스펙트로그램과 멜-주파수 켑스트럼 계수를 기반으로 한 전통적 신호 처리 기법을 활용하며, 이는 인간 청각 시스템을 모방한 멜 스케일 변환을 통해 감정 관련 음성 패턴을 효과적으로 포착하기 때문이다. 멜-주파수 켑스트럼 계수는 짧은 시간 프레임 내에서 로그 멜 스펙트럼의 이산 코사인 변환을 통해 주파수 왜곡을 보정하고, 감정 표현의 미세한 뉘앙스를 강조하는 데 적합하다. 그러나 이러한 방법론이 한국어와 같은 비영어 언어에 적용될 때 억양, 음운구조 등의 차이로 인해 적합하지 않을 수 있다. 이에 한국어 데이터셋을 활용하여 동일한 조건에서 여러 전처리 방식을 비교하여 한국어 특화 최적 전략을 도출하고자 한다.

Ⅲ. 실험

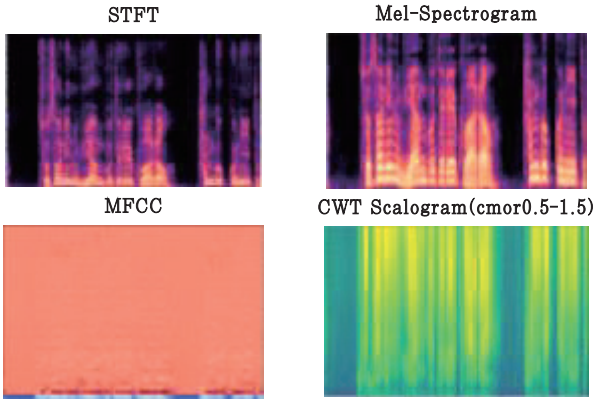
1. 데이터셋

본 논문에서는 실험은 AI-Hub의 <감성 및 발화 스타일별 음성합성> 데이터 중심으로 진행되었다. 이는 50명 이상의 전문 성우가 7개 감정(기쁨, 슬픔, 분노, 불안, 상처, 당황, 중립)에 대해 사전에 정의된 문장을 녹음한 음성으로 구성된다. 데이터셋 내 7개의 감정은 모두 음성 데이터가 균형 있게 분포되어 있다. 그러나 상처, 당황 감정은 다른 감정(예: 슬픔, 불안)과 혼동되기 쉽고, 주관적 레이블링의 변동성이 크다는 점을 고려하여 우선 제외하고 실험을 진행하였다. 원시 오디오는 샘플레이트 16kHz 형

식이며, 총 발화 수는 326,024건이다. 모든 음성 파일은 3초 단위로 자르거나(clip), 부족한 부분은 제로 패딩(zero-padding) 처리하였으며, 8:1:1의 비율로 학습/검증/테스트로 분할하여 최종 데이터셋을 구성하였다.

2. 모델 학습 및 결과 분석

실험은 전부 ResNet 모델을 특징 추출 및 분류로 고정하여 전처리 방식을 비교하였다. 이는 일반적으로 ResNet이 잔차 학습(residual learning)을 통해 깊은 네트워크에서 성능이 우수하다는 점을 고려하였다[7].



[그림 1] STFT, Mel-spectrogram, MFCC, CWT 시각화

비교를 위한 전처리 방법으로 음성 감정 인식에서 주로 사용되는 전통적 신호 처리 방식인 STFT, 멜 스펙트로그램(Mel-spectrogram), 멜-주파수 켑스트럼 계수(MFCC), 연속 웨이블릿 변환(CWT)를 활용하였다. [그림 1]은 '기쁨' 감정 라벨을 가진 발화 샘플에 대해 STFT, 멜 스펙트로그램, 멜-주파수 켑스트럼 계수, 연속 웨이블릿 변환을 적용하여 시각화한 결과이다. 스펙트럼 기반 방식인 STFT, 멜 스펙트로그램, 멜-주파수 켑스트럼 계수는 공통 파라미터 n_fft=512, hop_length=160, win_length=400으로 설정하였으며, 멜 스펙트로그램은 n_mels=64, 멜-주파수 켑스트럼 계수는 n_mfcc=40으로 설정하였다. 연속 웨이블릿 변환(CWT)는 중심 주파수 1.5, 대역폭 0.5인 모를렛 웨이블릿(Morlet Wavelet) cmor0.5-1.5을 사용하고, 스케일 수를 64로 설정하여, 64개의 서로 다른 스케일(주파수 대역)을 사용해 음성 신호를 분석하였다. 이는 음성 처리에서 일반적으로 사용되는 파라미터 범위를 고려하여 설정하였다. 모든 실험은 배치 256, 에포크 5 조건에서 학습되었으며, Adam(Adaptive Moment Estimation) 옵티마이저(학습률 1e-3), Cross-Entropy 손실함수를 사용하였다.

[표 1] STFT, Mel-spectrogram, MFCC, CWT 별 성능 평가 결과

Model	Accuracy	F1-score
CWT	0.6843	0.6723
MFCC	0.8981	0.8987
STFT	0.9708	0.9709
Mel-spectrogram	**0.9715**	**0.9716**

실험 결과, 멜 스펙트로그램과 STFT가 각각 97.15%, 97.08%의 정확도로 가장 높은 성능을 보였다. 이어서 멜-주파수 켑스트럼 계수(MFCC)가 정확도 89.91%, 연속 웨이블릿 변환(CWT) 정확도 68.43%로 그보다 낮은 성능을 기록했다. 이는 시간-주파수 분포가 한국어 발화의 특성을 효과적으로 포착하는 것을 보여준다. 멜-주파수 켑스트럼 계수는 특징 표현 방식은 멜 스케일과 이산 코사인 변환으로 인해 고주파 정보 손실이 발생하여 상대적으로 낮은 성능을 보인 것으로 해석되며, 연속 웨이블릿 변환은 복잡한 계산과 스케일 64개의 높은 해상도로 인한 최적화가 어려웠을 가능성이 있다. 따라서 스케일 수, 웨이블릿 종류 등 파라미터 조정을 통해 성능 향상의 여지가 있다.

Ⅳ. 결론

본 논문에서는 한국어 환경 음성 감정 인식(SER)의 전처리 방식을 동일한 조건에서 체계적으로 비교하였다. 모든 실험은 16kHz 음성을 3초 단위로 정제한 뒤, STFT, 멜 스펙트로그램(Mel-spectrgram), 멜-주파수 캡스트럴 계수(MFCC), 연속 웨이블릿 변환(CWT)을 각각 2차원 표현으로 변환하여 ResNet 모델에 입력하였다. 결과는 멜 스펙트로그램과 STFT 전처리 방식이 각각 97.15%, 97.08%의 정확도로 가장 우수했으며, 연속 웨이블릿 변환 방식이 68.43%로 가장 낮았다. 이는 한국어 발화에서 시간-주파수 분포를 보존하는 표현 방식이 효과적인 전처리 방식임을 보여준다. 이는 한국어 도메인의 음성 감정 인식 분야에서의 전처리 방법의 선택 근거를 제공하여, 향후 실무 확장성을 높이는 데 기여할 것으로 기대된다.

다만, 본 연구는 잡음이 적고, 감정이 비교적 명확한 발화 데이터셋을 기반하므로 일반 발화로의 전이에서 성능 저하가 발생할 수 있다는 점에서 한계가 있다. 이에 AI-Hub의 <감정인 태깅된 자유대화>를 추가로 활용해 일반(자연) 발화 도메인 적응을 후속 연구에서 검증할 예정이다.

ACKNOWLEDGMENT

이 연구는 2025년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업(RS-2022-NR070870) 지원을 받아 수행되었으며, 과학기술정보통신부·경찰청이 공동 지원한 '폴리스랩 2.0 사업'(RS-2023-00281072) 및 2025년도 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 지원을 받아 수행된 2025년 산학연 Collabo R&D 사업 (RS-2025-02323085)의 지원을 받아 수행된 연구입니다.

참 고 문 헌

[1] 신도경, 김영대. "딥러닝 기반 음성 감정 인식 학습 데이터 확장을 위한 감정 특화 오디오 데이터 증강 방법", 2025, 한국방송·미디어공학회 하계학술대회 논문

[2] 최지예, 최아영. "한국어 데이터셋을 활용한 음성 감정 인식에서의 데이터 증강 효과 연구", 2025, 한국컴퓨터종합학술대회

[3] 서재진, 강태인, 곽일엽. "사전 학습 모델과 앙상블 기법을 통합 음성 감정 인식", 2024, 한국데이터정보과학회지

[4] K. V.Krishna Kishore, et al. "Emotion Recognition in Speech Using MFCC and Wavelet Features", 2024, IEEE

[5] Vidhi Sareen, Seeja K.R. "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks(CNN)", 2025, ELSEVIER

[6] Taiba Majid Wani, at al. "A Comprehensive Review of Speech Emotion Recognition Systems", 2021, IEEE

[7] Minjeong Lee, Miran Lee. "Performance Improvement of Speech Emotion Recognition Using ResNet Model with Data Augmentation-Saturation",2025, MDPI