Title: Multilingual Emotion Recognition Using Transformer-Based Audio-Text Fusion

Abstract:

Emotion recognition from multimodal data is essential for enhancing natural language interfaces and improving human-computer interaction. In this study, we propose a multilingual emotion recognition model that fuses acoustic and textual features using a transformer-based attention mechanism.

Our model integrates Mel spectrogram representations of audio signals with language-specific BERT embeddings, enabling accurate emotion classification across diverse languages. We evaluate our model on three public datasets-EmoDB, RAVDESS, and MELD-covering English, Arabic, and German.

Experimental results demonstrate that our architecture significantly outperforms conventional CNN and RNN-based models, achieving state-of-the-art accuracy in multilingual settings. This approach shows promising potential for emotion-aware virtual assistants and globalized affective computing systems.