Speech Emotion Recognition with Co-attention Based Multi-Level Acoustic Information

Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, Eng Siong Chng

Name: Harsukh Sagri, **Roll Number:** 210428

October 22, 2024

1 Introduction

The report includes the experiment and result of an implementation of a Speech Emotion Recognition experiment using a multi-level acoustic feature fusion strategy, based on MFCC, spectrogram, and wav2vec2 (W2E) embeddings with a co-attention mechanism. The experiment was conducted on the IEMOCAP dataset; performance is evaluated through two cross-validation strategies: 5-fold leave-one-session-out and 10-fold leave-one-speaker-out.

2 Dataset Description

The IEMOCAP dataset consists of 5531 utterances that have been categorized into four primary emotions, angry, sad, happy, and neutral. From this dataset, multiple modalities were provided to us, but for this experiment, we have used only the audio modality, having extracted the following features from the audio signals:

- **MFCC**: capture low-level acoustic information.
- **Spectrogram**: capture more detailed spectral information.
- **W2E**: Wav2vec2 embeddings, pre-trained on raw audio signals to capture high-level information.

These features were fused using a co-attention mechanism which is done to improve emotion recognition performance.

3 Implementation Details

The implementation of the model was predominantly on the wav2vec2.0 model that is pre-trained from the place Hugging Face. This is the named feature of the multi-level combination, which is the co-attention mechanism over MFCC, and spectrogram and W2E features. The experiment duration was 50 epochs and the repetition of the configurations was two-fold for both 5- and 10-fold cross-validation strategies. Loss, weighted accuracy (WA), and unweighted accuracy (UA) were the metrics used for performance evaluation.

3.1 Environment Setup

The environment setup involved installing the necessary libraries from the GitHub repository:

- Python 3.10
- PyTorch

- Pre-trained Wav2Vec2.0 from Hugging Face

The primary code for cross-validation was done by executing the files in **main_code** folder from the repository, which in turn executes the other files.

4 Results

The results obtained from both cross-validation strategies are summarized below. Each experiment was repeated twice, and the average results were reported.

4.1 5-fold Leave-One-Session-Out

- 1st run: Loss = 1.9917, WA = 67.99%, UA = 70.07%
- 2nd run: Loss = 1.1591, WA = 72.56%, UA = 72.95%

Average Performance: WA = 70.28%, UA = 71.51%

4.2 10-fold Leave-One-Speaker-Out

- 1st run: Loss = 0.9081, WA = 70.74%, UA = 72.55%
- 2nd run: Loss = 1.0027, WA = 68.94%, UA = 72.17%

Average Performance: WA = 69.84%, UA = 72.36%

The following table provides a comparison between our results and the averages reported in the reference paper.

Table 1: Performance Comparison

Cross-validation	**WA (%)**	**UA (%)**	**Research Paper (%)**
5-fold leave-one-session-out	70.28	71.51	WA: 69.80, UA: 71.05
10-fold leave-one-speaker-out	69.84	72.36	WA: 71.64, UA: 72.70

4.3 Normalised Confusion Matrix

The averaged confusion matrix for 5-fold leave-one-session-out is shown below:

	Angry	Sad	Happy	Neutral
Angry	0.7989	0.0143	0.0605	0.1264
Sad	0.0153	0.7293	0.1531	0.1024
Happy	0.0613	0.0416	0.7876	0.1096
Neutral	0.0331	0.2062	0.2160	0.5446

5 Conclusion

The test outcomes of the study show that the suggested way for recognizing the emotion of speech utilizing MFCC, the spectrogram, and the W2E with co-attention is feasible. The incorporation of several levels of the theory of speech with the co-attention technique gives stable outcomes. Here, baseline performance is matched or surpassed.

Some additions in the analysis could involve exploring different fusion strategies (different combinations of acoustic features with or without co-attention) or combining this approach with additional modalities such as video or text.