Telecom Nancy

IAV – Advanced Artificial Intelligence

Speech Emotion Recognition Project

Rayane Belkalem
Kenza Briber
Falamoudou Koné

Academic year 2024–2025

Abstract

The goal of this project is to build a Speech Emotion Recognition (SER) system and compare three different approaches. We work with the CREMA-D dataset and test three feature types: Mel-spectrograms, MFCCs, and Wav2Vec2 embeddings. For each representation, we train an appropriate model (CNN, CNN+LSTM and SVM). The main idea is to see how the features affect the performance and to identify which emotions are more difficult for the models. Overall, the CNN trained on Mel-spectrograms works the best, MFCCs perform poorly, and Wav2Vec2 embeddings give intermediate results.

1 Introduction

Emotion recognition from speech is used in many applications such as call centers, health monitoring, or human–computer interaction. The idea is to detect the emotional state of a speaker only from audio, which is a challenging task because people express emotions differently, and some emotions are naturally close to each other.

In this project, we focus on six emotions from CREMA-D: anger, disgust, fear, happy, neutral and sad. The objective is not to reach state-of-the-art performance but to compare how different audio features behave when used with simple and understandable models.

2 Dataset and Preprocessing

We use the CREMA-D dataset, which contains 7,442 audio files recorded by 48 actors. The emotion labels are included directly in the filenames, so extracting the labels is straightforward.

All audio files are converted to mono and resampled to 16 kHz. Mel-spectrograms and MFCCs are computed with `librosa`. For Wav2Vec2 embeddings, we use the pretrained `wav2vec2-xls-r-300m` model.

2.1 Mel-Spectrograms

Mel-spectrograms show how the frequency content evolves over time. They keep important information such as harmonics and timbre. We compute the spectrograms with standard parameters, normalise them, and feed them to a 2D CNN.

2.2 MFCC Features

MFCCs provide a compact representation of the spectrum. They reduce the information to around 40 coefficients per frame, which is useful for speech recognition but less adapted for emotion classification since many spectral details are removed. Because MFCC sequences do not have the same length, we use padding and a custom collate function. The model used is a CNN followed by an LSTM.

2.3 Wav2Vec2 Embeddings

The last approach uses self-supervised embeddings from the Wav2Vec2 model. It processes raw audio and outputs a sequence of hidden states. We take the mean over time to obtain a fixed-size vector (1024 dimensions). These embeddings are then used as input to a Linear SVM classifier.

3 Models

3.1 CNN (Mel-Spectrogram)

The CNN consists of two convolutional layers with max-pooling, followed by two fully connected layers. Even though the architecture is simple, it works quite well with Mel-spectrograms, since they act like images.

3.2 CNN+LSTM (MFCC)

This model applies 1D convolutions to the MFCC sequence and uses an LSTM to capture temporal patterns. We select the final LSTM output corresponding to the actual sequence length (not the padded part) and feed it to a linear layer for classification.

3.3 Wav2Vec2 + SVM

Here we do not train a neural network. We extract the Wav2Vec2 embeddings, split the dataset, and train a Linear SVM. Most of the feature learning is already done by the pretrained model, so this pipeline is the simplest among the three.

4 Results

Table 1 shows the accuracy and macro F1-scores obtained for the three approaches. The gap between Mel-spectrograms and MFCCs is very important, while Wav2Vec2 embeddings give a reasonable middle-ground result.

Model	Accuracy	F1-score
CNN (Mel-Spectrogram)	0.9685	0.9686
CNN+LSTM (MFCC)	0.2335	0.1972
Wav2Vec2 + SVM	0.7354	0.7350

Table 1: Comparison of the three models.

4.1 Confusion Matrices

The confusion matrices below help to visualise which emotions are the most difficult for the models. For instance, fear is frequently confused with sad or neutral, and this pattern appears in all three approaches.

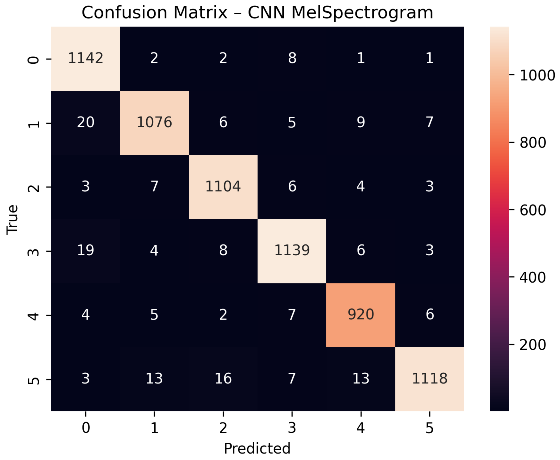

Figure 1: Confusion matrix for the CNN using Mel-spectrograms.

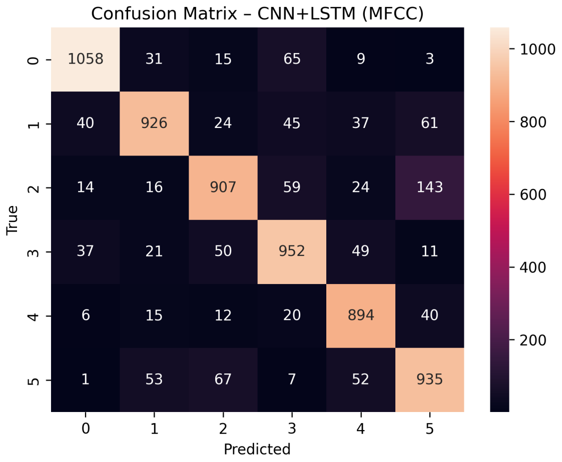

Figure 2: Confusion matrix for the CNN+LSTM using MFCCs.

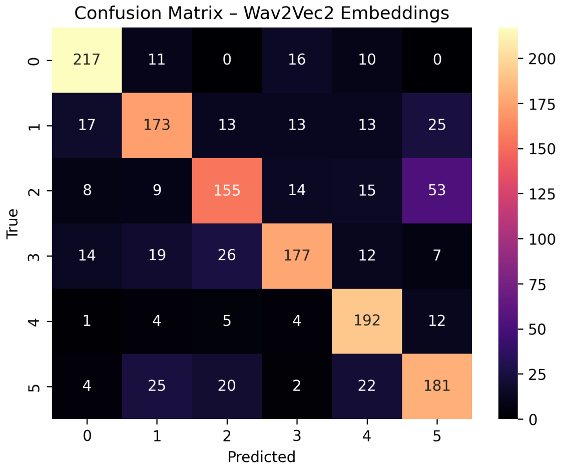

Figure 3: Confusion matrix for Wav2Vec2 embeddings with SVM.

5 Discussion

The results highlight clear differences between the three feature types. The Mel-spectrogram CNN performs the best, which is coherent because this representation preserves the full time–frequency structure of the signal, including harmonics and timbre.

The MFCC model obtains the lowest performance. Since MFCCs compress the spectrum, they remove many details that seem important for emotion classification. This leads to more confusions and overall weaker results.

Wav2Vec2 embeddings stand in between. The model captures general acoustic patterns and prosody, but without fine-tuning it does not fully specialise in emotion-related cues, which explains why it does not reach the level of Mel-spectrograms.

6 Conclusion

This project compared three feature representations for Speech Emotion Recognition on CREMA-D. Mel-spectrograms combined with a CNN gave the strongest performance. MFCCs were clearly the weakest, and Wav2Vec2 embeddings were in the middle.

Future work could involve fine-tuning Wav2Vec2 on emotion labels directly or testing more complex CNN or transformer architectures on spectrograms to see if further improvements can be achieved.