A hybrid deep framework to predict the emotions from spoken language, which consists of ConvNets, CNN-LSTM, and DNN, to extract spatial and temporal associations from the raw audio data and low-level acoustic features.

A four-layer deep neural network to fuse the features and classify the emotions, which allows global fine-tuning of the entire network.

A detailed comparison with previous work and modality-specific models.

1. Used text as input and extracted the part-of-speech tags (POS) for each sentence using Natural Language Toolkit (NLTK) [1].
2. Extracted the Mel-frequency spectral coefficients (MFSCs) from raw audio signals as input and extracted the low-level pitch and vocal related features using OpenSMILE software [2].
3. Evaluated on IEMOCAP including anger, sad, neutral, frustration, and happy (happy∗excited).

Figure 1. Overall structure of the proposed deep multimodal framework

Figure 2. Feature extraction structure for MFSC maps

Table 1. Accuracy comparison of different feature combinations (percentage)