Implementation of Speech Emotion Recognition using Dual-Sequence LSTM paper

Jayanthi Sri Manichandana, 210469

1 Dataset Description

Multimodal Emotion Recognition IEMOCAP The IEMOCAP dataset consists of a total of 302 videos across the dataset with 2 speakers per session. The 9 emotions (angry, excited, fearful, sad, surprised, frustrated, happy, disappointed, and neutral) as well as valence, arousal, and dominance are noted in each segment. We used only the audio modality for training. The dataset includes a total Utterances of 5,531 utterances (audio clips), Emotions: Happy (29.5%), Neutral (30.8%), Angry (19.9%), Sad (19.5%). Sample Rate of 16 kHz and an Audio Length which Varies from 2 to 40 seconds

2 Model Implementation

The DS-LSTM (Dual Stream LSTM) model processes audio data. It uses both time-domain and frequency-domain features and enhances speech recognition. It processes two types of data i.e.; MFCC features, and Mel-spectrograms. The MFCC features are extracted using Librosa. 13 MFCCs make up each frame. To represent temporal dynamics, it incorporates delta (first derivative) and delta-delta (second derivative) characteristics. We have implemented LSTM for MFCCs. CNN for Mel-Spectrograms and DS-LSTM for both. Two mel-spectrograms with varying time-frequency resolutions are calculated. This offers complimentary spectrum data at different levels of detail. We trained the model using the Adam optimizer with a learning rate of 0.0001 for 20 epochs. The loss function used is cross-entropy.

3 Results

We evaluated the model using 5-fold cross-validation on the IEMOCAP dataset. The performance metrics are shown in Table.

Table 1: Weighted Accuracy (WA) and Unweighted Accuracy (UA) of the DS-LSTM Model

Model	**Mean WA (%)**	**Mean UA (%)**
Baseline 1 (MFCC LSTM)	64.7	65.5
Baseline 2 (CNN + LSTM)	63.5	64.5
Proposed DS-LSTM	**72.7**	**73.3**

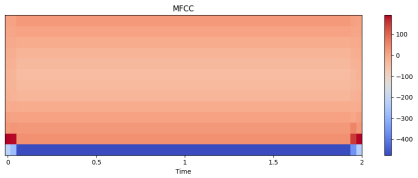

Figure 1: MFCC features extracted from an audio clip.

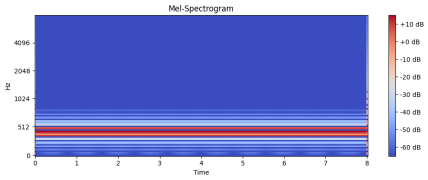

Figure 2: Mel-spectrogram extracted from the same audio clip.