**2. Emotion recognition in conversations**

Each dataset has a number of conversations. Each conversation has a number of utterances each of which has an emotional label, e.g., happy, angry, sad, neutral.

**3. Our approach**

- A learnable frontend used for audio feature extraction
- Effective context addition using self-attention with Bi-GRU network
- Multimodal transformers used for fusion of modalities
- Model trained in a hierarchical manner

**4. Proposed Model**

- LEAF-CNN training results in emotionally discriminative audio features
- BERT-BiGRU training results in emotionally discriminative text features
- LEAF-CNN and BERT-BiGRU networks are frozen

**5. Audio feature extraction**

- Each blue block represents a CNN filter with batch normalisation and ReLU activation
- LEAF-CNN training results in emotionally discriminative audio features

**6. Text feature extraction**

- BERT-BiGRU training results in emotionally discriminative text features
- BERT-base pre-trained model is not frozen during this training

**7. Multi-utterance self-attention**

- Separate BiGRU networks used for text and audio for adding context
- Self-attention employed across utterances for better context modelling

**8. Dataset**

- IEMOCAP has 151 recordings - divided into 5 sessions
- Each utterance is labeled one of the four categories - happy, angry, sad and neutral

**9. Results**

- Common test settings - CV5 - 5-fold cross validation, CV10 - 10 fold cross validation, Session 5 as test

**10. Results with ASR transcripts**

- Provided transcripts replaced by Google speech-to-text (42% WER)
- Models trained with provided transcripts, tested with ASR transcripts

**References**

- Wu et al. "Emotion recognition by fusing time synchronous and time asynchronous representations," ICASSP 2019