Multimodal Conversational Emotion Recognition

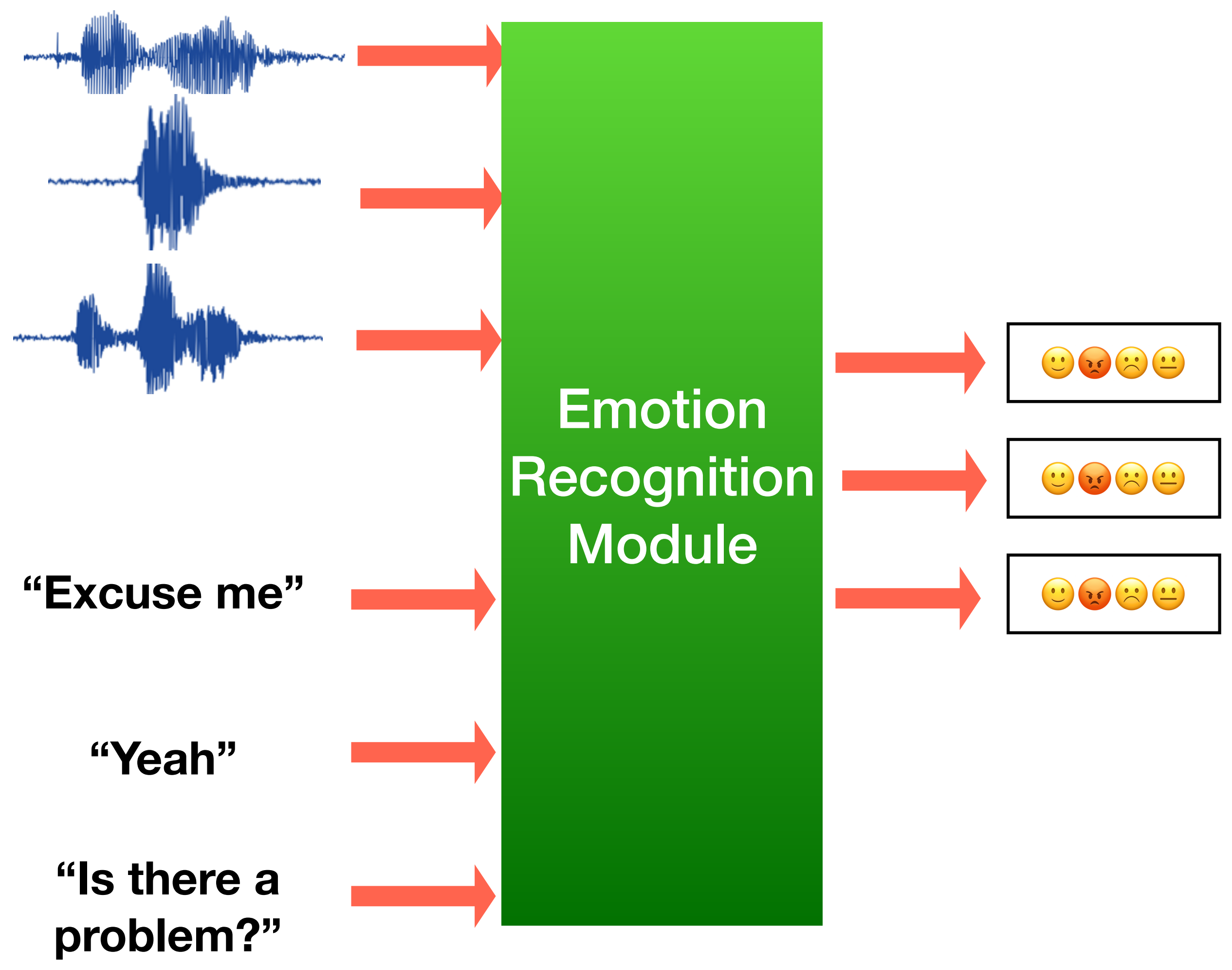
Soumya Dutta and Sriram Ganapathy

LEAP Lab, Electrical Engineering,
Indian Institute of Science, Bangalore

OPEN DAY 2023

Introduction

Recognise emotions of speakers engaging in a conversation, taking cues from multiple modalities such as speech and provided text transcriptions


Emotion Recognition Module

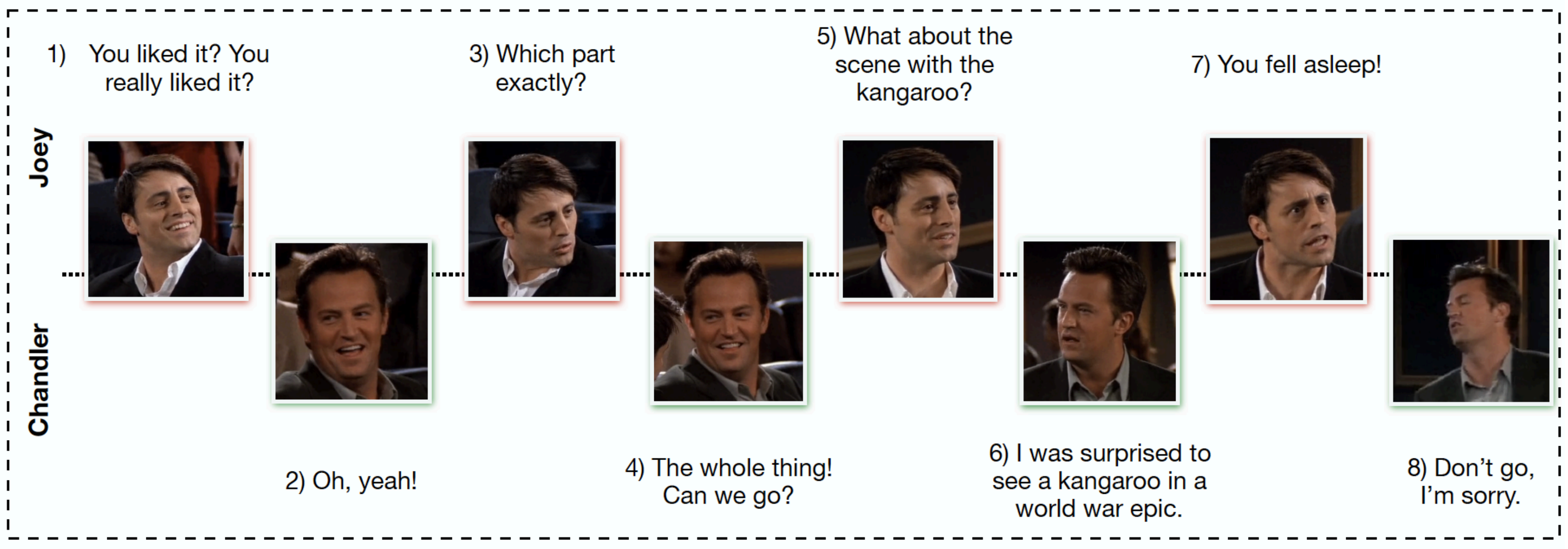
"Excuse me"

"Yeah"

"Is there a problem?"

Major Challenges

- Multiple speakers - contextual decisions are key
- Information across multiple modalities - fusion is key



1) You liked it? You really liked it?

2) Oh, yeah!

3) Which part exactly?

4) The whole thing! Can we go?

5) What about the scene with the kangaroo?

6) I was surprised to see a kangaroo in a world war epic.

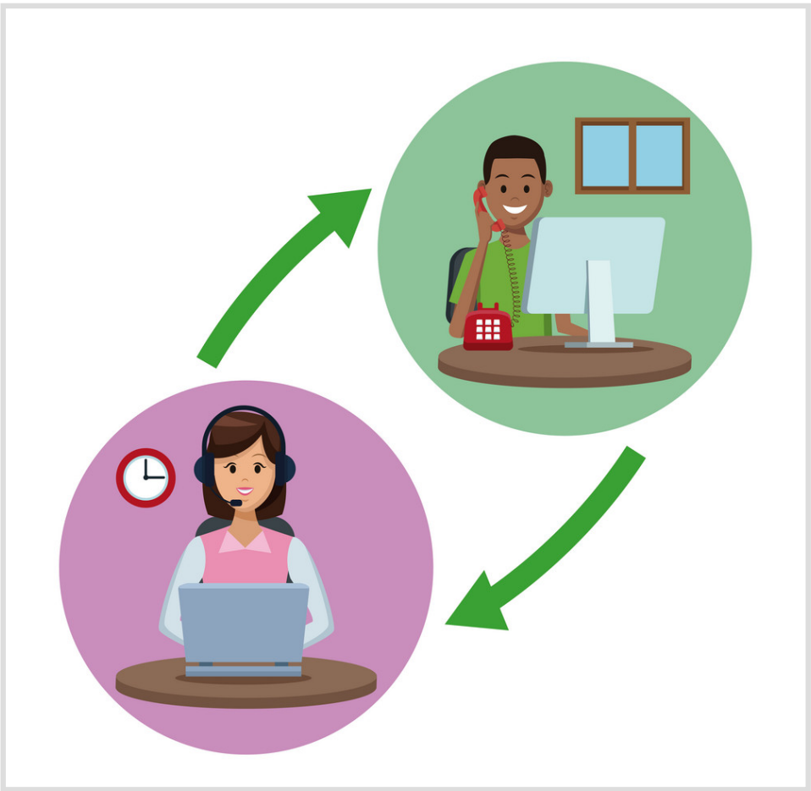
7) You fell asleep!

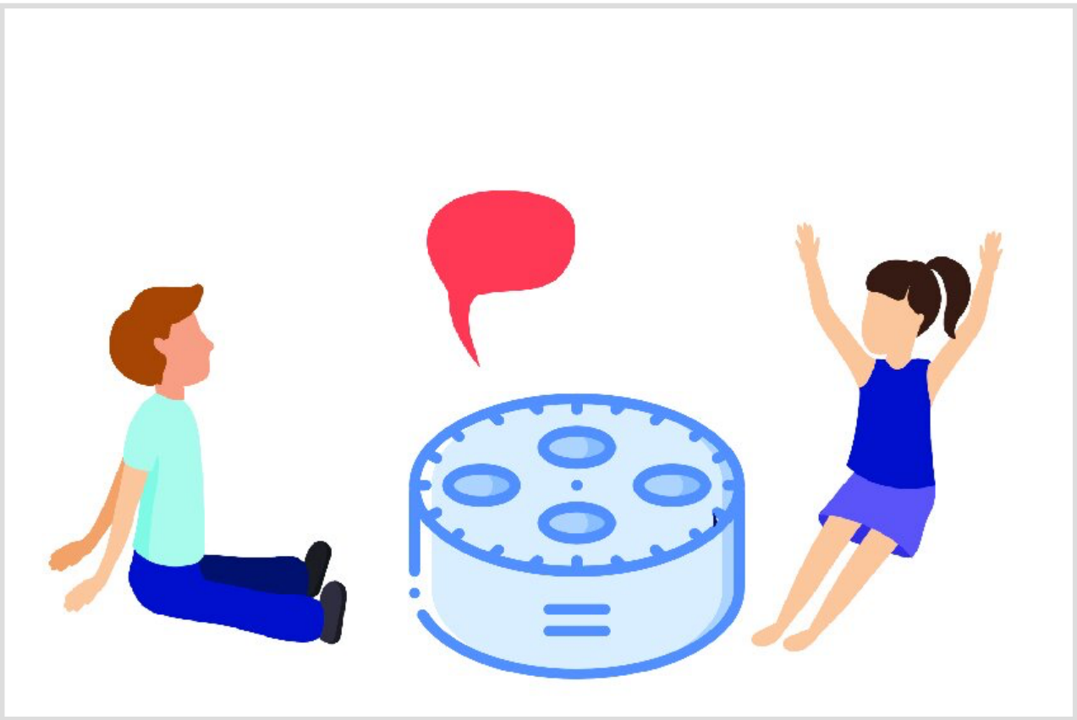
8) Don't go, I'm sorry.

Applications

Continuous monitoring of emotions - better recognition of mental health issues

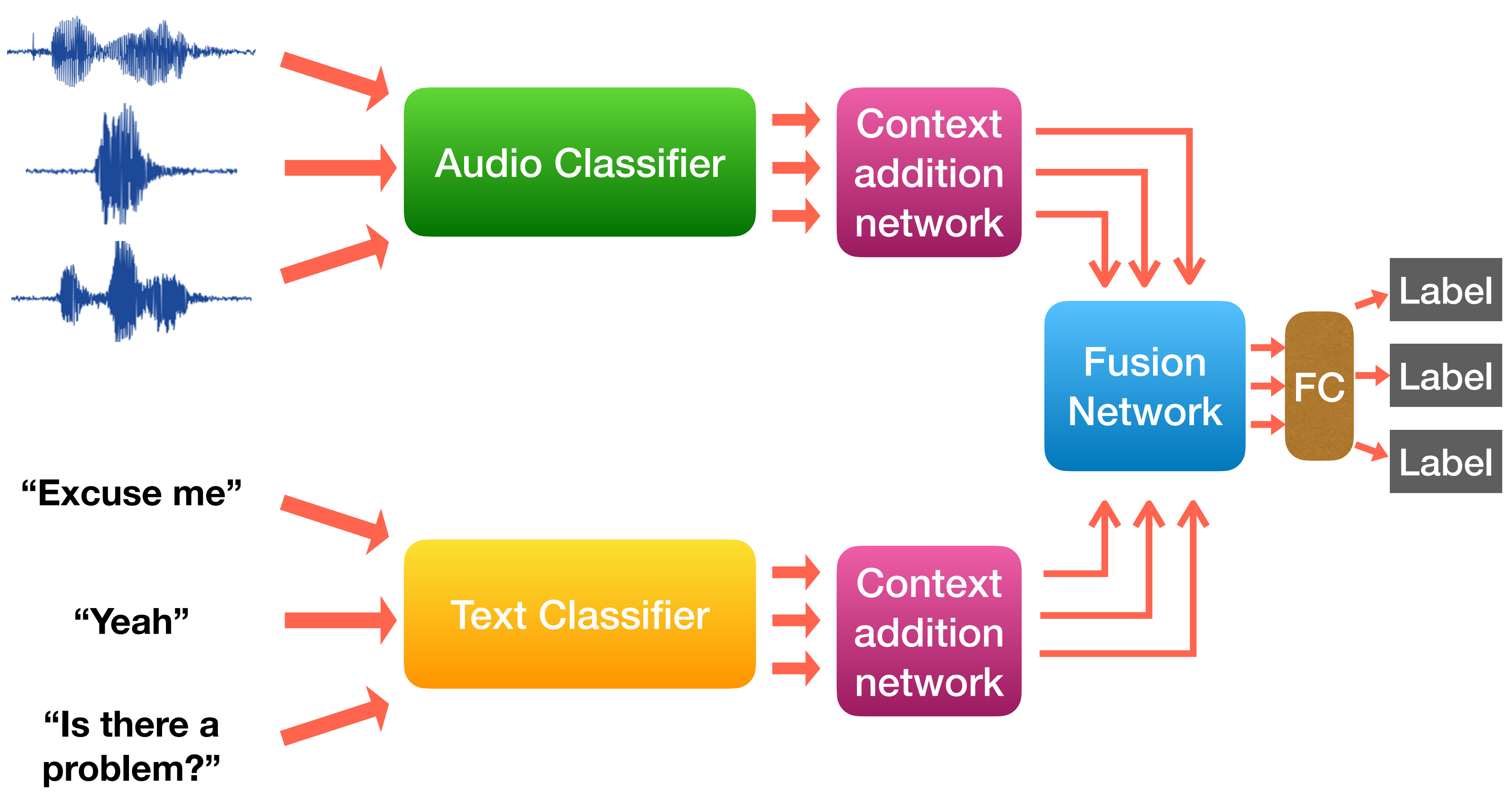
Analysis of customer care call centre conversations





Improving human machine interaction

Approach


Audio Classifier

Context addition network

Fusion Network

FC

Label

Label

Label

"Excuse me"

"Yeah"

"Is there a problem?"

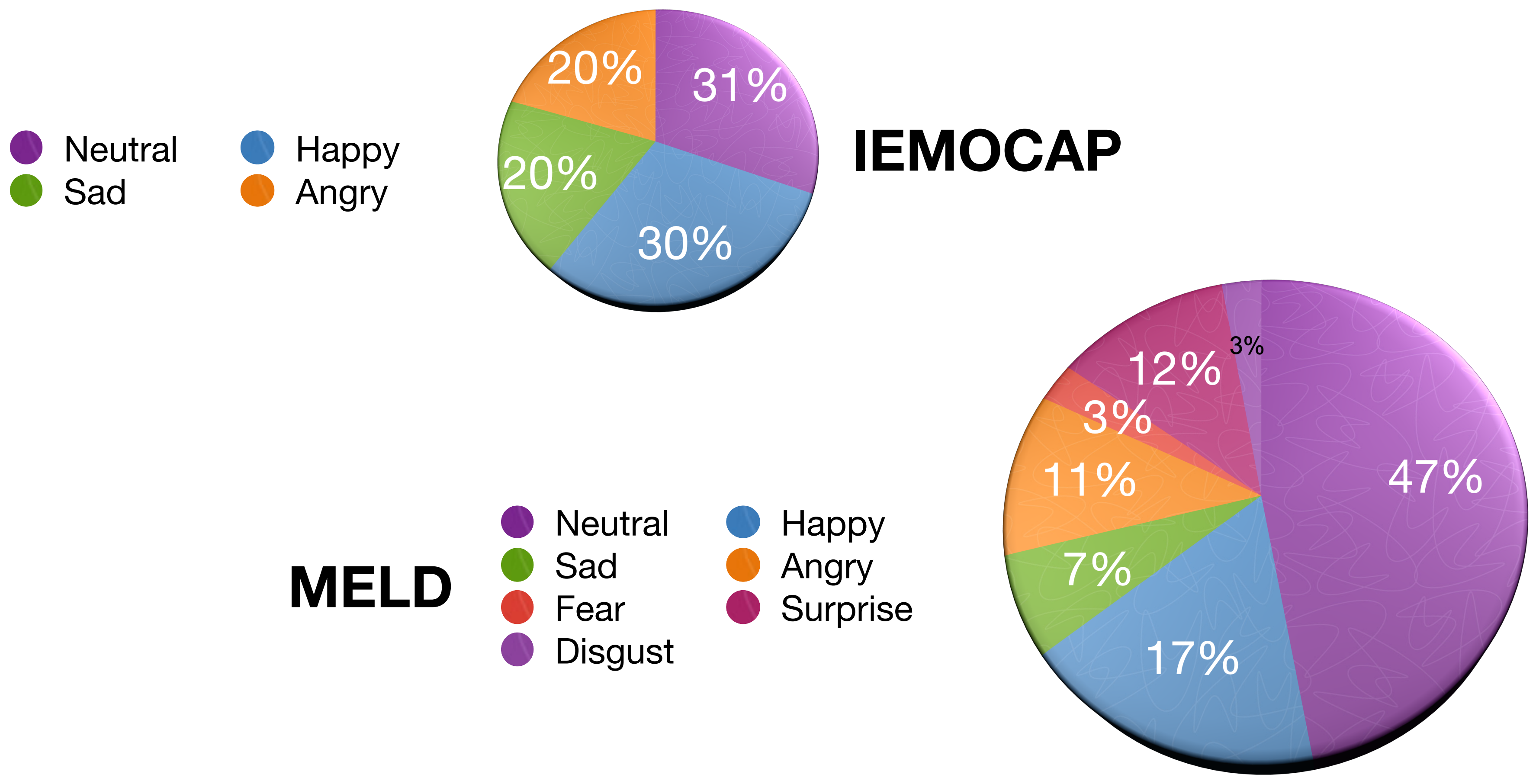
Text Classifier

Context addition network

- This is a hierarchical model which first trains the audio and text classifier to classify individual speech and text utterances
- Context from other utterances in conversations is added by neural networks suited for temporal data - RNNs, LSTMs, GRUs
- Fusion of the two modalities is done by aligning the audio and text representations by means of similarity - transformer

Datasets and Results

- Major and most popular Datasets - IEMOCAP and MELD
- Unbalanced datasets - Metric is **Weighted F1 score**



- IEMOCAP - recognition score ~85%; MELD recognition score ~ 66%
- Further research required for better recognition of sparsely represented classes

References

- Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database." *LREC* 42.4 (2008): 335-359.
- Poria et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations." *arXiv preprint arXiv:1810.02508.*
- Dutta et al. "Multimodal Transformer with Learnable Frontend and Self Attention for Emotion Recognition." *ICASSP 2022*.