# A multimodal architecture with shared encoder that uses spectrograms for audio

**Sai Anuroop Kesanapalli, Riya Ranjan, Aashi Goyal, Wilson Tan**
[kesanapa, riyaranj, aashiarv, wtan1167] @usc.edu
University of Southern California

## 1 Problem definition

Multimodal learning aims to create models that process and relate information from multiple modalities. Prior works such as (Buddi et al., 2023) provide architectures that have one encoder tailored per modality. Human communication is multimodal by nature which limits the performance of unimodal models. A shared encoder architecture may be capable of fusing multimodal information while providing better synergy between modalities compared to architectures that use separate encoders. Although (Li et al., 2022) propose one such sensor fusion model, it is designed for Radar and Lidar data, both of which are visual in nature. We thus propose to work on a novel bimodal architecture that features a shared encoder for audio and video channels, and utilizes the spectrogram representation of audio data to help achieve uniformity with the video data. As a proof of concept, we wish to test this architecture for emotion recognition on CREMA-D dataset (Cao et al., 2014), given its simplicity and aptness for our bimodal use-case.

## 2 Problem relevance

Multimodal fusion is critical for developing artificial agents that can jointly understand the verbal, non-verbal and contextual cues present in human communication. By aligning multimodal features better, the proposed architectures would be able to implicitly capture these cues that are subtly manifested across modalities in human communication.

Furthermore, a shared encoder architecture could lead to improved performance on identifying basic emotions, while allowing the model to identify more complex emotions in social communication such as jealousy or empathy.

## 3 Datasets

We propose utilizing the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (Cao et al., 2014) for our work, offering a rich multimodal experience, integrating audio and video for enhanced emotion analysis. Evaluated by over $2,400$ individuals, CREMA-D includes $7,442$ video clips with performances by $91$ actors, providing a diverse exploration of emotional expression. Within the dataset, each actor presents $12$ sentences, expressing $6$ emotions at different intensity levels. Each video clip is brief, lasting less than $5$ seconds. Importantly, the dataset includes the number of ratings for each emotion, offering valuable insights into the perceived emotional content of the performances.

## 4 Planned analysis

We propose to work on a novel audio-visual learning paradigm where audio data is represented as spectrograms, in order for the embeddings to be used with an encoder that is shared between audio and video data. The architecture is visualized in Fig. 1.
Our proposed work is divided into three phases as described below:

- **Video-pipeline**: This is a standard video inference pipeline that shares the same architecture as that of the audio-pipeline except for the spectrogram generation phase, as described next.
- **Audio-pipeline**: In this pipeline audio data first gets converted to spectrograms, which are then passed through an embedding layer to generate embeddings, that get fed to a shared audio-video encoder, and the subsequent latent features are then passed through a fully-connected layer coupled with softmax to generate logits.
- **Bimodal-pipeline**: This is a merger of the aforementioned two pipelines where the audio and video embeddings get fused before being passed on to the shared encoder, and can be

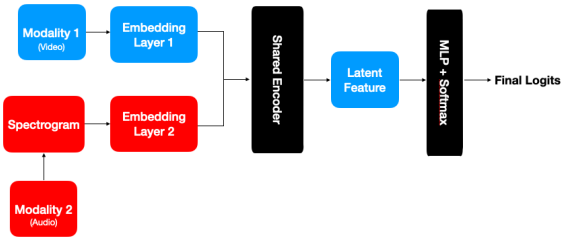


Figure 1: Architecture

categorized under early fusion.

By adopting this unimodal to bimodal development approach, we will be tackling the problem in increasing order of complexity.

## 5 Expected outcomes

- Implement the proposed architecture phase-wise (Sec. 4 and Fig. 1).
- Test the implemented architecture on CREMA-D audio-visual dataset (Sec. 3).
- Compare the performance of the proposed architecture on various metrics against baselines (Lei and Cao, 2023).
- Discuss potential applications for the proposed architecture.

## 6 Division of work

- Dataset processing - SAK, AG
- Video-pipeline - RR, WT
- Audio-pipeline - SAK, WT
- Bimodal-pipeline - AG, RR
- Reports - All

## References

Sai Srujana Buddi, Utkarsh Oggy Sarawgi, Tashweena Heeramun, Karan Sawnhey, Ed Yanosik, Saravana Rathinam, and Saurabh Adya. 2023. Efficient multimodal neural networks for trigger-less voice assistants.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Yuanyuan Lei and Houwei Cao. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969.

Yu-Jhe Li, Jinhyung Park, Matthew O'Toole, and Kris Kitani. 2022. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 918–927.