Double Multi-Head Attention Multimodal System for Odyssey 2024 Speech Emotion Recognition Challenge

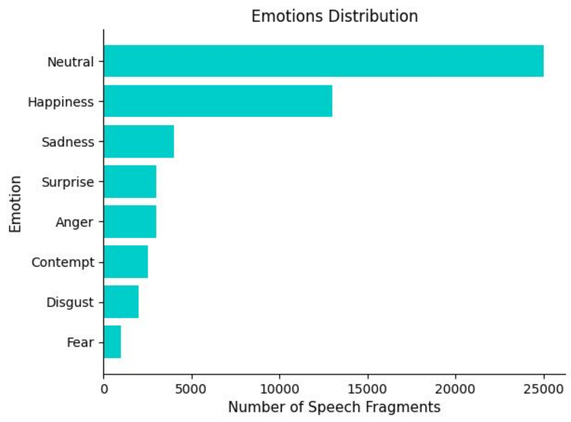
Federico Costa, Miquel India, Javier Hernando

Abstract

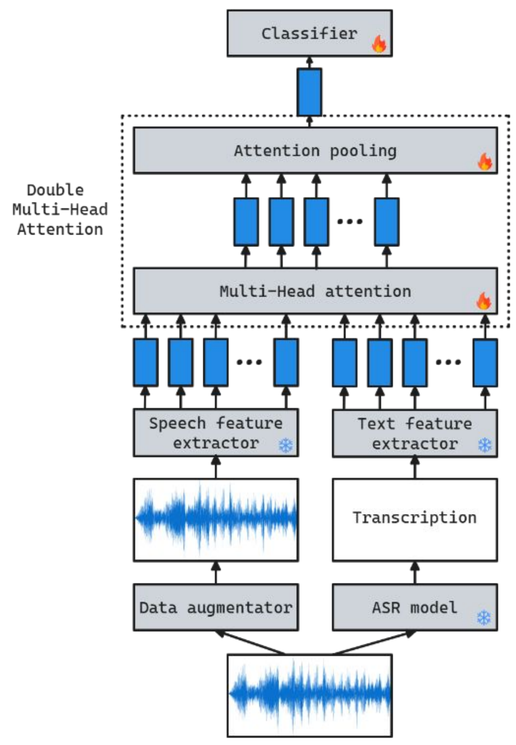
As computer-based applications are becoming more integrated into our daily lives, the importance of Speech Emotion Recognition (SER) has increased significantly. In this paper we describe the Double Multi-Head Attention Multimodal System developed for the Odyssey 2024 SER Challenge. Our proposed system achieved the third position, where 31 teams participated in total.

The Odyssey 2024 SER Challenge

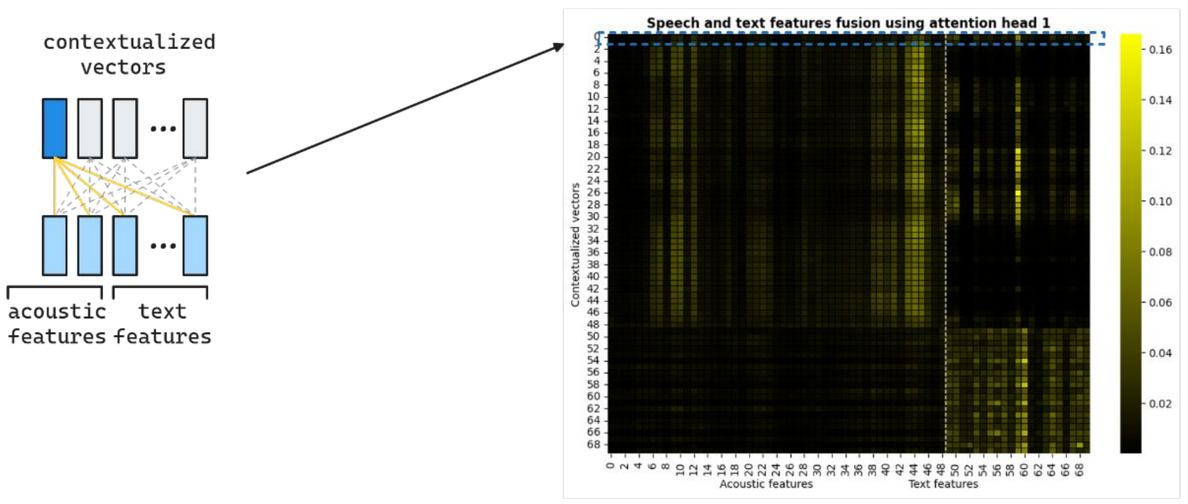
- The challenge aims to enhance innovation in recognizing emotions from spontaneous speech.
- The task it to classify speech segments into emotional classes: anger, happiness, sadness, fear, surprise, contempt, disgust and a neutral state.
- The MSP-Podcast dataset contains 240K hours of spontaneous and diverse emotional speech samples collected from podcast recordings.



Architecture



Multi-Head Attention Multimodal Fusion Visualization



Experimental Details

- 18M trainable parameters
- On-line data augmentation process
- ASR model: Whisper
- Text feature extractor: BERT
- Text features: last layer of the pre-trained model
- Speech feature extractors:
  - wav2vec2.0 and XLS-R
  - HuBERT
  - wavLM
- Speech features: (learned) weighted averaged layers of the pre-trained model
- Loss functions: Weighted Cross Entropy or Focal Loss
- Thresholds adjustment to maximize each class F1-score

Results

Configuration		Train Macro F1-score	Validation Macro F1-score ↓
Ensemble of models		34.88%	33.80%
WCE Loss	XLS-R	34.42%	33.43%
Focal Loss	XLS-R	39.15%	33.37%
WCE Loss	wav2vec2.0	34.83%	32.69%
Focal Loss	HuBERT	37.84%	32.40%
WCE Loss	HuBERT	36.73%	32.18%
WCE Loss	wavLM	32.48%	31.44%
Focal Loss	wav2vec2.0	35.75%	31.27%
Focal Loss	wavLM	33.24%	30.77%
Challenge Official Baseline		-	30.70%

Threshold adjustment was applied to every model, except for the Challenge Official Baseline. Ensemble of models combine the following three models: WCE Loss and XLS-R; WCE Loss and wav2vec2.0; WCE Loss and wavLM.

- Odyssey 2024 SER Challenge
  - 3rd place (31 teams)
  - Macro-F1 test: 34.41%
  - Baseline 10% relative improvement
- IberLEF 2024 EmoSpeech Challenge
  - 1st place (14 teams)
  - Macro-F1 test: 86.69%
  - Baseline 63% relative improvement

Conclusions

- Acoustic and text features were extracted using pre-trained self-supervised models
- Multimodal features were mixed using a Double Multihead Attention component
- 3rd place in the Odyssey SER Challenge 2024 and 1st place in the IberLEF EmoSPeech Challenge 2024
- Architecture adaptable to other multimodal classification tasks

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project.