BSC-UPC at EmoSPeech-IberLEF2024: Attention Pooling for Emotion Recognition

Marc Casals-Salvador[1], Federico Costa[1], Miquel India[2], and Javier Hernando[1,2]

Barcelona Supercomputing Center[1], Universitat Politècnica de Catalunya[2]

Speech Emotion Recognition (SER) has emerged as a dynamic field in machine learning with impactful applications in daily life. The **IberLEF 2024** hosted a competitive challenge using the Spanish MEACorpus 2023 dataset. This work leverages pre-trained speech and text models with attention pooling for feature extraction, achieving **first place** over 14 teams with an **86.69% Macro F1-Score**.

Architecture

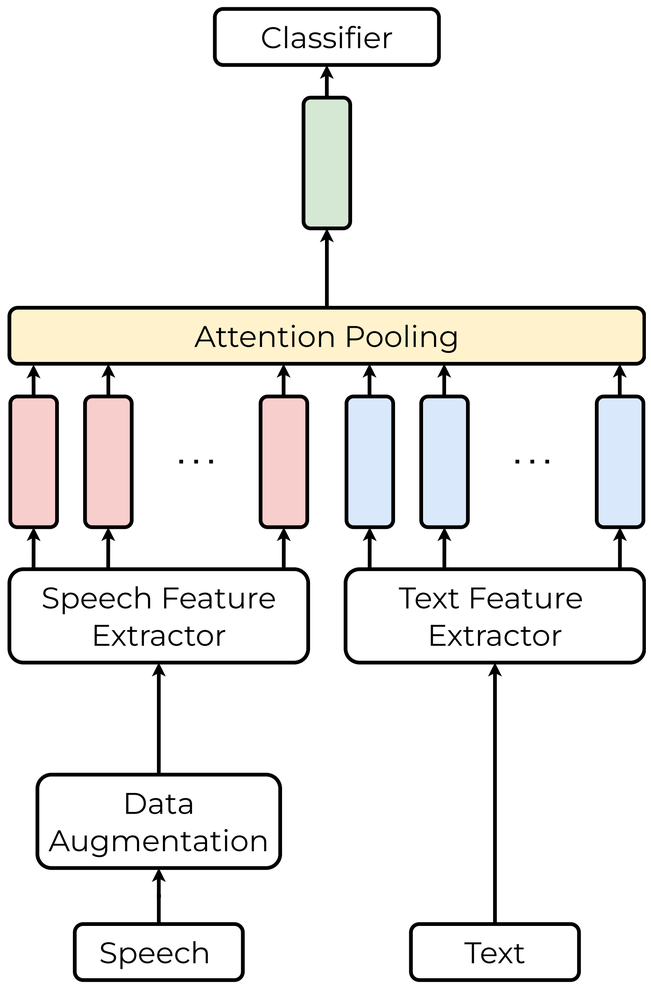

Figure 1: Diagram Attention Pooling for the Multimodal Emotion Recognition System. The speech utterances are represented in red and the text is represented in blue.

Attention Pooling

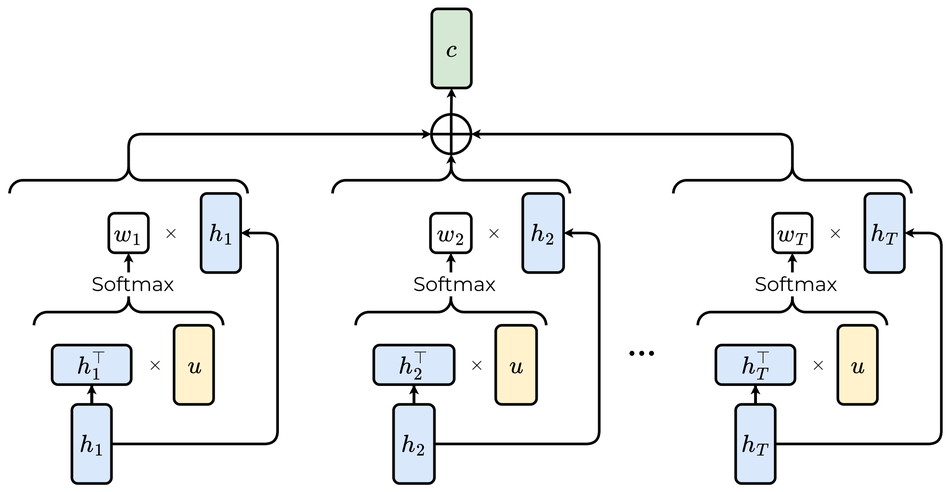

Figure 2: Diagram of the Attention Pooling operation.

Let $\{h_t \in \mathbb{R}^E | t = 1, ..., T\}$ be the hidden states of dimension E. We define the Attention Pooling as:

$$w_t = \frac{\exp\left(\frac{u^\top h_t}{\sqrt{E}}\right)}{\sum_{i=1}^{T} \exp\left(\frac{u^\top h_i}{\sqrt{E}}\right)}$$

$$c = \sum_{t=1}^{T} w_t h_t$$

where u is a trainable parameter

Speech Feature Extractor

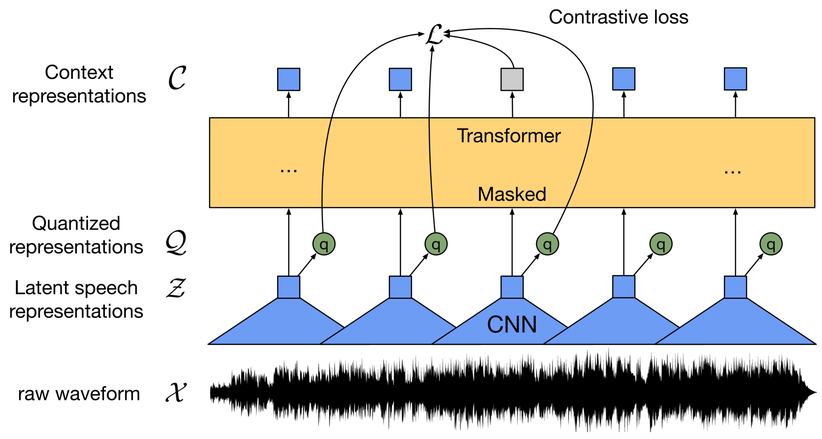

Figure 3: Diagram of wav2vec 2.0. Figure extracted from Baevski et al.

- Trained with 436,000h of **multilingual** speech.
- The **quantization module** transforms continuous speech into discrete speech units.
- It uses **1-D Convolutions** that acts as relative positional embedding.
- It uses **contrastive loss**.

Dataset

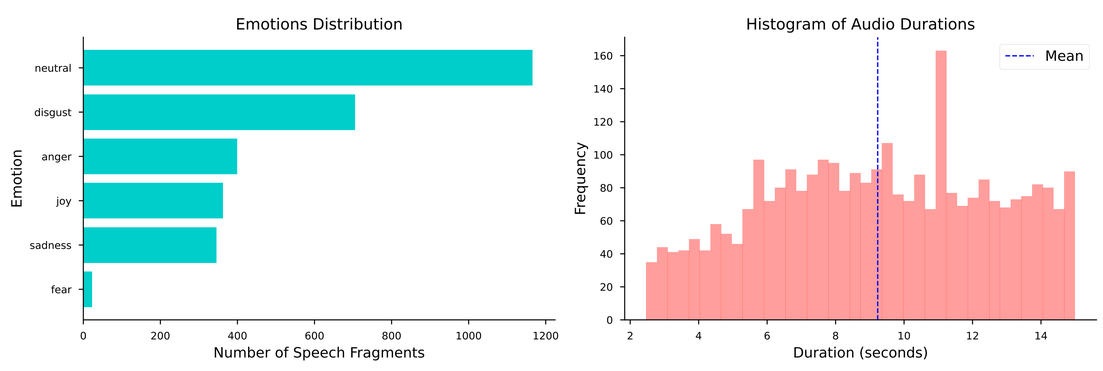

Figure 4: Some of the characteristics of the MEACorpus 2023. On the left is the distribution of the number of speech fragments over emotions. On the right, a histogram of the durations of the audio fragments.

Data Augmentation

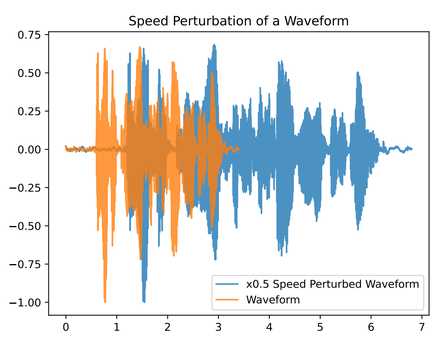

Figure 5: In orange, the original waveform. In blue, the same waveform half the speed

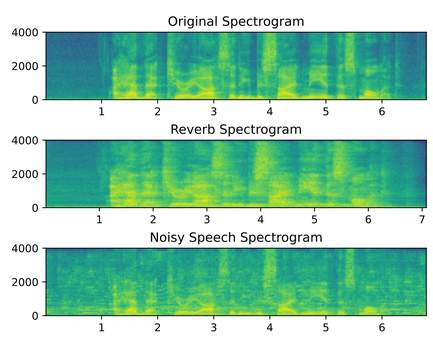

Figure 6: Spectrograms of distinct transformations

Experimentation and Results

- The audio files were **cropped** using a **5.5s window**, randomly slicing the waveform.
- The optimal **data augmentation probability** was **0.3.**
- A **weighted cross-entropy loss** was used, with weights as the inverse of class frequencies.

Different configurations of Feature Extractors were evaluated in the development sub-dataset.

Text Model	Audio Model	Output Dimensions	Validation F1-Score
RoBERTa	WavLM LARGE	1,024	80.04%
RoBERTa	**XLSR-wav2vec 2.0**	**1,024**	**89.73%**
RoBERTa	HuBERT LARGE	1,024	76.03%
BERT Large Uncased	WavLM LARGE	1,024	83.27%
BERT Large Uncased	XLSR-wav2vec 2.0	1,024	86.59%
BETO	WavLM BASE PLUS	768	74.79%
BETO-EMO	WavLM BASE PLUS	768	73.19%

Different hyperparameter configurations were trained to ensemble different models.

Model Name	Hidden dense layers	Weight Decay	Test F1-Score
Top 1 Model	2	0.01	86.20%
Top 2 Model	2	0.1	85.96%
Top 3 Model	3	0.1	82.43%
Model Ensemble	-	-	**86.69%**
Baseline	-	-	53.08%

Conclusions

- What distinguishes **XLSR-wav2vec 2.0** and **RoBERTa** from their counterparts is their training on a large corpus, which results in better performance.
- The top-3 models had very similar architectures and hyperparameter configurations; hence, ensembling provides little variance in the predictions.
- **Attention pooling** offers an efficient method for merging speech and text features, yielding better results than vanilla attention and other pooling strategies.
- **Few-parameter attention** models can still outperform **transformer-based** models in certain domains.
- The system scored **86.69% Macro F1-Score**, achieving the **first place** over 14 teams.

Read the paper here!

Download the code here!

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