Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

Marc Casals-Salvador, Federico Costa,
Miquel India, and Javier Hernando

September 2024

Table of Contents

Introduction

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-UPC Team

Marc Casals i Salvador
marc.casals@bsc.es

Federico Costa
federico.costa@upc.edu

Miquel India Massana
miquelindia90@gmail.com

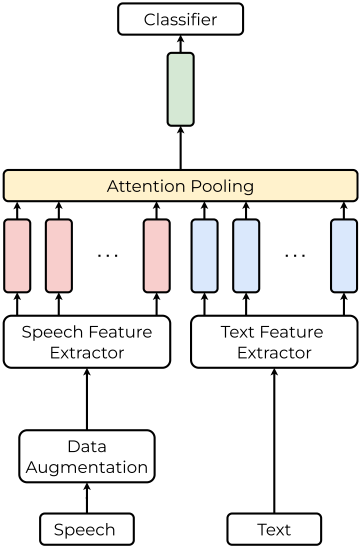
Javier Hernando Pericás
javier.hernando@upc.edu

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Architecture

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

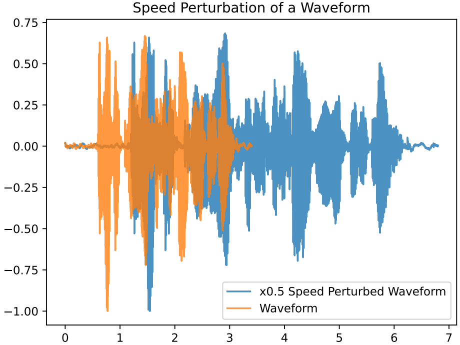
Model Architecture

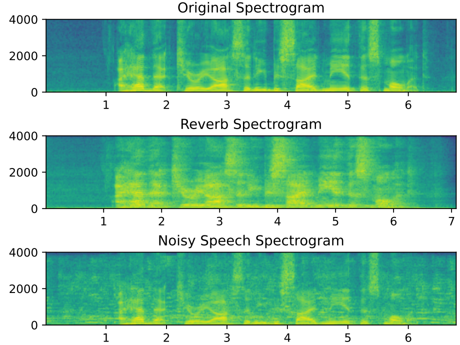


- The input of the network is **speech** and **text**.
- We applied **Data Augmentation** to the speech waveforms.
- Self-Supervised Learning Models were used as feature extractors to produce **hidden state vectors**.
- These vectors were merged into one using **Attention Pooling**.
- The final vector was projected with some **dense layers**.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Data Augmentation





Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Feature Extractors

We used speech and text pre-trained self-supervised models to extract relevant features. The ones we experimented with are the following:

Speech Feature Extractors

- **WavLM**: Trained with 80,000h. The model has 316.2M of parameters
- **XLSR-wav2vec 2.0** : Trained with 436,000h. The model has 317M of paramters
- **HuBERT**: Trained with 60,000h. The model has 300M of parameters

Text Feature Extractors

- **BERT**: Output dimension of 1,024. The model has 355M of parameters
- **XLM-RoBERTa Spanish**: Output dimension of 1,024. The model has 355M of parameters
- **BETO**: Output dimension of 768. The model has 110M of parameters.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Speech Feature Extactors

From the different feature extractors, we selected **XLSR-wav2vec2.0**.

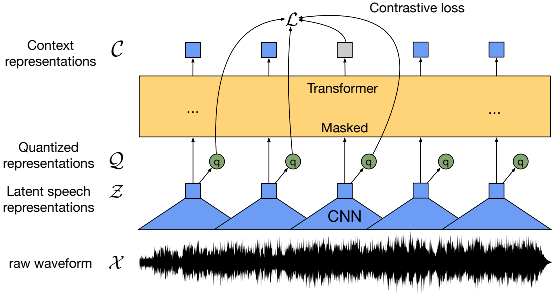
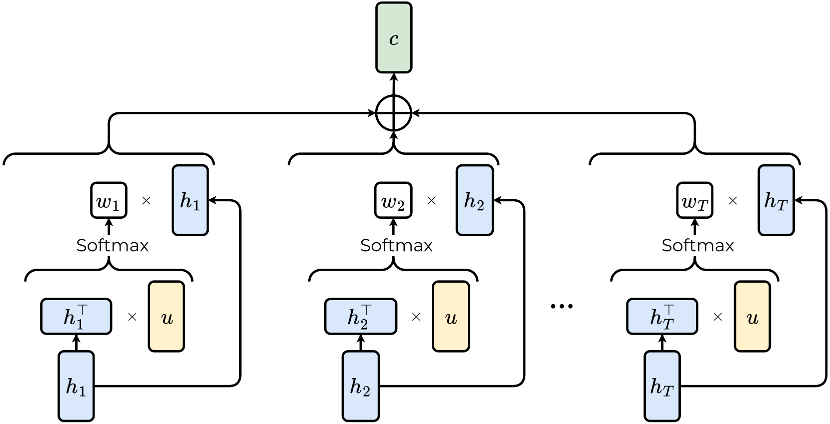


Figure: Figure extracted from the Wav2Vec2.0 paper.

- Trained with 436,000h of **multilingual** speech.
- It has a **quantization module** which transforms continuous speech into discrete speech units.
- Unlike the Transformer, it doesn't have a **Positional Encoding**. It uses **1-D Convolutions** that acts as relative positional embedding.
- It uses **contrastive loss**.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Attention Pooling



Attention Pooling VS Basic single-query attention

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

Basic single-query attention.

$$w_t = \frac{\exp\left(\frac{q^\top k_t}{\sqrt{E}}\right)}{\sum_{i=1}^T \exp\left(\frac{q^\top k_i}{\sqrt{E}}\right)}$$

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

where k_i, q_i, v_i are trainable parameters.

Attention Pooling

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

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

where u is a trainable parameter

Experimental Setup and Results

Waveform processing

Audio Cropping

In training, waveforms are randomly cropped in windows of 5.5 seconds.

Data Augmentation

The data augmentation is applied on the fly, allowing each batch of data to be augmented dynamically. It was proven that 0.3 was the best value. The intrinsic details of the transformations are the following:

- **Speed Perturbation:** The waveform's speed was randomly modified x0.9 or x1.1.
- **Reverberation:** It was used OpenSLR [13, 20, 26]
- **Background noises, music or voices:** MUSAN.

Feature Extractors

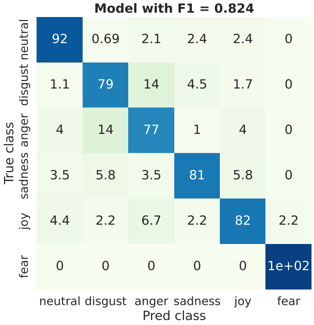
To choose the best-performing feature extractors we made different tests combining them.
We created a **Validation Set** from the training set to avoid sending submissions.

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	76.033%
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%

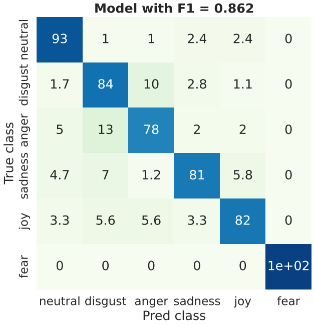
All of these configurations had their corresponding hyperparameter tuning, and the best of each one was selected

Hyperparameter Tuning

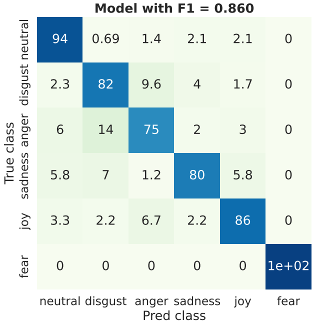
We tested different model configurations with a wide variety of **hyperparameters**. The ones that gave better results are the following:



(a) Hidden dense layers = 3
Weight decay = 0.1



(b) Hidden dense layers = 2
Weight decay = 0.01



(c) Hidden dense layers = 2
Weight decay = 0.1

Figure: Confusion matrices in the test set. The drop-out was set to 0.1 and the data augmentation probability to 0.3.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Results over the Test Set

We combined the best three models using **hard voting**. These are the results obtained over the test set for each submission in Codalab:

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%

The **Model Ensemble** obtained the best performance in the multimodality part of the competition.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you for your attention!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Additional Slides

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Attention

In Slide 11 we used Single Query Attention. The Attention algorithm follows this formula

$$w_t = \frac{\exp\left(\frac{q_j^\top k_t}{\sqrt{E}}\right)}{\sum_{i=1}^{T} \exp\left(\frac{q_j^\top k_i}{\sqrt{E}}\right)}$$

$$c_j = \sum_{t=1}^{T} w_t v_t$$

where k_i, q_i, v_i are trainable parameters.