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

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

[1]*Barcelona Supercomputing Center, Eusebi Güell Square 1-3, 08034 Barcelona, Spain*

[2]*Universitat Politècnica de Catalunya, Jordi Girona 31, 08034 Barcelona, Spain*

Abstract

The domain of speech emotion recognition (SER) has persistently been a frontier within the landscape of machine learning. It is an active field that has been revolutionized in the last few decades and whose implementations are remarkable in multiple applications that could affect daily life. Consequently, the Iberian Languages Evaluation Forum (IberLEF) of 2024 held a competitive challenge to leverage the SER results with a Spanish corpus. This paper presents the approach followed with the goal of participating in this competition. The main architecture consists of different pre-trained speech and text models to extract features from both modalities, utilizing an attention pooling mechanism. The proposed system has achieved the first position in the challenge with an 86.69% in Macro F1-Score.

Keywords

Speech Emotion Recognition, Deep Learning, Attention, Transformers,

1. Introduction

Emotions are undoubtedly fundamental parts of our idiosyncrasy. They play an important role in inter-personal relationships and decision-making and generally take part in the evolution and consciousness of any mental process [1]. Moreover, there is empirical evidence that emotions immensely influence human health [2], which creates the necessity to monitor them. Affective computing is of great interest in medical health fields [3]. Consequently, developing a system capable of discerning various emotions is highly valuable. Researchers have attempted to predict emotions using machine learning approaches, but their effectiveness depends on the quality and quantity of available data.

In the field of Natural Language Processing (NLP), numerous emotion recognition models rely exclusively on text-based features. Since the creation of Transformers [4], multiple approaches [5, 6, 7] seek to use pre-trained Transformer encoders, such as BERT [8] as text feature extractors. These encoders are pre-trained models using self-supervised learning techniques on extensive datasets. This pre-training enables the models to project words in a latent space with a rich semantic representation, from which classifiers can then be employed to predict the corresponding emotion classes. On the other hand, speech is crucial for expressing emotions. Elements such as pitch, prominence, and phrasing contribute generously to providing emotion information. As explained in [9], the human brain is capable of recognizing emotions pan-culturally and independently of the language they are expressed with. Following this principle, the researchers have proposed different architectures to process and extract information from speech signals. In the realm of Speech Emotion Recognition (SER), Machine Learning (ML) and Deep Learning (DL) models often utilize hand-crafted features such as Mel-Frequency Cepstral Coefficients (MFCC) [10, 11, 12, 13]. Nevertheless, recent advances in Deep Learning allow cutting-edge architectures to combine text and speech to provide better results. Multimodal emotion Recognition (MER) is a complex task since it requires models to be able to learn complex patterns of the data. For

---

this reason, the usage of text and audio pre-trained models significantly improves the embedding representation of their features, allowing them to be combined in their latent space [14, 15, 16].

In summary, the exceptional progress that has been made in this field is appreciable, yet there also exists a disparity in the amount of work carried out in Spanish. Developing these models using Spanish data is crucial for several reasons. On the one hand, Spanish is one of the languages with the most native speakers worldwide. On the other hand, it is necessary to leverage the engineering opportunities of Spanish-speaking countries, opening the door to developing new technologies that could satisfy their population needs. However, one of the difficulties this presents is the lack of labelled data in Spanish needed to train any supervised learning approach.

In order to encourage the creation of an emotion recognition model trained with Spanish data, the Iberian Languages Evaluation Forum (IberLEF) of 2024 created the challenge EmoSPeech 2024. This competition evaluates the Macro F1-Score of the participants in two tasks: Emotion Recognition with text and with speech and text. The training corpus is Spanish MEACorpus 2023 [17]. This dataset contains 13.16 h of speech, and its transparent methodology distinguishes it from other datasets of the same task. The speech samples are collected from YouTube videos and are labelled using the categorical taxonomy proposed by P. Ekman [18], which include emotions such as surprise, disgust, anger, joy, sadness, fear, and neutral expressions. Nevertheless, it was impossible for the annotators to find any sample that contained speech expressing surprise emotion, though this class is not represented in the dataset.

This paper endeavours to leverage the research of Spanish models with cutting-edge technologies in the current state of the art by making use of the MEACorpus 2023. The current state of the art is the usage of Transformers-based pre-trained models with high capacity that, leveraging the enormous datasets they are trained with, are capable of describing complex patterns in both text and speech. Following this lead, the system combines a speech pre-trained model, the XLSR-wav2vec 2.0 [19], that is trained with 436,000h and a RoBERTA text model fine-tuned in Spanish [20]. Both models are used as feature extractors for speech and text respectively, and they output a vector with the relevant information of the utterances. The two vectors are then concatenated into a single vector that contains information about text and speech. Subsequently, this vector is reduced to lower-dimension representation by making use of an attention pooling mechanism. Finally, dense layers are utilized to project this reduced vector, determining the classification of the utterance. This approach has reached an F1-Score of 86.69%, achieving the first position in the multimodal task of the EmoSPeech 2024 challenge.

2. Challenge

The EmoSPeech 2024 [21] is a Challenge proposed by the Iberian Languages Evaluation Forum (IberLEF) of 2024 [22], which is a Spanish workshop hosted by the *Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN). This event is dedicated to producing models in the frame of the Iberian peninsula, encompassing different national languages such as Spanish, Portuguese, Catalan, Basque, or Galician.

To develop the competition, the organisers proposed a dataset named Spanish MEACorpus 2023 [17]. The dataset, comprising 13.16h of speech divided into 5,129 audios, was meticulously labelled by the research team of the article. As explained in their paper, the procedure to extract the audio files is as follows: The authors of the dataset selected YouTube videos according to their topic and extracted audio segments considering the noise of the audio files and the silence gaps. Once this part was done, the audio files were classified using the emotion taxonomy developed by P. Ekman [18]. It comprehended the basic emotions of disgust, anger, joy, sadness, fear, surprise, and neutral emotion. Nevertheless, despite the efforts, finding any speech audio that contained the surprise emotion was impossible.

As is common in the field, the dataset exhibits an unequal representation of the emotion classes. Figure 1 (left)shows that neutrality and disgust are the most prevalent emotions, while fear is notably scarce. Another important aspect is the length of the audio files, which can directly affect the performance of the network by providing more contextual information about the speech. Therefore, audio duration is
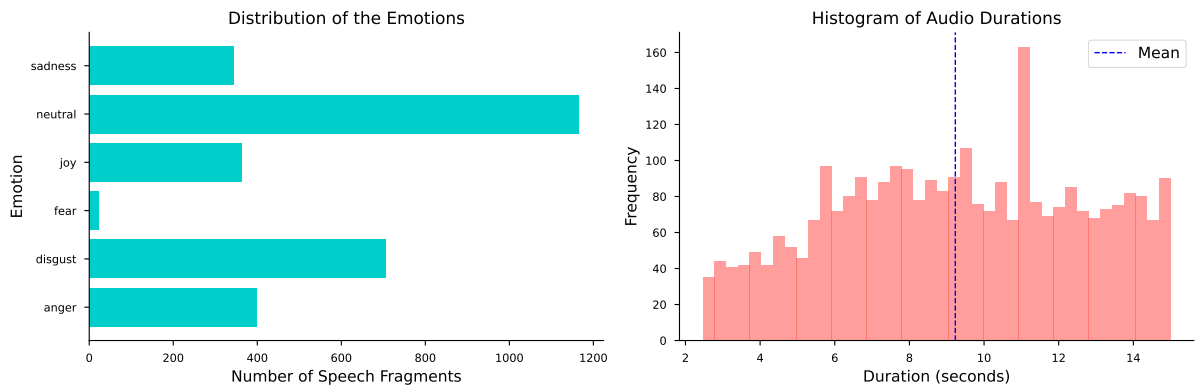
Figure 1: Some of the characteristics of the MEACorpus 2023. On the left is the distribution of the number of speech fragments over emotions. On the right, a histogram of the durations of the audio fragments.

an essential characteristic of any speech dataset and must be considered a quality metric. A histogram of the audio file duration is shown in Figure 1(right). The mean of the duration is 9.24 s. Lastly, another fundamental consideration is the variability of the recordings. The fact that this competition uses third-party audio files makes it more difficult to control other parameters, such as noise or the magnitude of the audio. Some videos are recorded outdoors and are more likely to have background noise or exhibit poorer recording quality, while others are studio-recorded and of higher quality. Furthermore, the dataset's paper detailed that 46% of the speech segments are attributed to female voices, with the remainder belonging to males. The paper affirmed that the text transcriptions were extracted directly from the raw audio file using Whisper [23] followed by a manual revision by the researchers.

3. Architecture

The system created is a multimodality model that combines text and speech and, trained with the Spanish MEACorpus 2023 dataset [17], effectuates a classification of the emotion. In Figure 2, it is possible to see a global representation of the system. As can be seen, both text and speech are fed to the network and processed with a self-supervised learning (SSL) model that works as a feature extractor. Then, the model concatenates these features and pools them into one single vector. Finally, a classification is performed by processing this vector with multilayer perceptrons that serve as a classifier.

3.1. Speech

It is a standard practice to apply regularisation techniques to preprocess the data before its utilisation in a machine learning model. In the case of this system, the overall mean and standard deviation of all the audio files were calculated and used to normalise the values in the dataset. This technique is widespread in Deep Learning literature for speech due to the fact that it makes backpropagation more efficient and reduces the impact of the outliers. The normalisation process was extended to the validation and test sets, whereby the mean and standard deviation values obtained from the training set were utilised. It is essential to mention that, during the training stage, the samples were randomly cropped with a window of 5.5 seconds. The shorter utterances were enlarged using repetition padding.

Given the limited size of the dataset, it became necessary to use data augmentation to mitigate overfitting to the training set. The specific techniques utilized included speed perturbation, which alters the speed of the audio; reverberation, which simulates a reverberant environment; and background noise, which adds ambient sounds to the audio. PyTorch implements data augmentation using the following scheme: first of all, defining the transformations used to create synthetic data. Then, for each sample, a probability decides whether the model will see the original data or the one created after applying these transformations. This process occurs in every epoch, so data augmentation is applied to
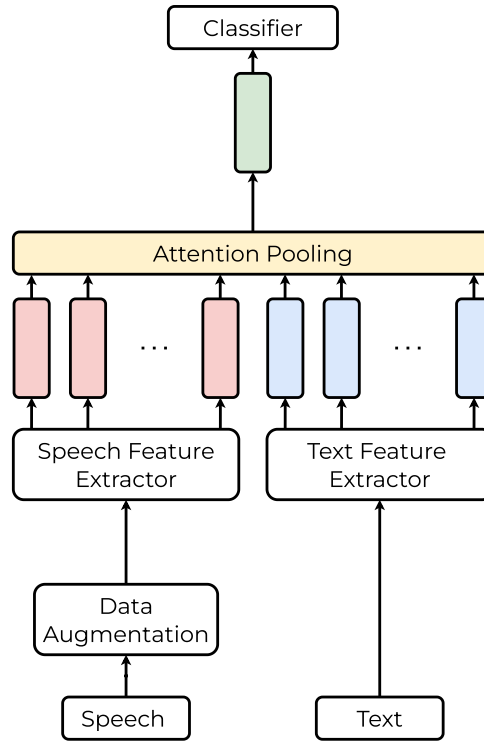
Figure 2: Diagram Attention Pooling for the Multimodal Emotion Recognition System. The speech utterances are represented in red and the text is represented in blue.

different samples in each epoch. This streaming data augmentation guarantees the model is exposed to a wide diversity of data without physically increasing the dataset size. As is common in the field, when doing inference is not desired to perturb the data, so in the validation and test sets, the probability of applying these transformations is zero.

Recent advancements in the field of Deep Learning have highlighted the importance of using Transformer-based pre-trained models. It has been proven that their ability to adapt to changes in the domain makes them suitable for extracting features from audio files in any dataset. In this work, the following models were experimented with:

- **WavLM** [24]: This cutting-edge model is trained with 80,000 hours and encompasses datasets such as Libri-Light [25], GigaSpeech [26] and VoxPopuli [27]. Two versions of this model were tried, the Large version and the Base. The output vector is 768 in the case of the base version and 1,024 in the case of the large.
- **XLSR-wav2vec 2.0** [19]: This model is based on wav2vec 2.0 [28] and it is trained with the datasets Common Voice [29], BABEL [30] and Multilingual LibriSpeech [31], which makes a total of 436,000 hours of audio in 128 languages. This model outputs a vector of dimension 1,024.
- **HuBERT** [32]: This model is trained with 60,000 hours of Libri-Light [25]. The output vector is of dimension 1,024.

3.2. Text

The text domain was the first to employ pre-trained large language models (LLMs) for different tasks. Since the creation of the BERT model [33], different approaches have emerged in the state of the art, following the same idea with some variations. In the approach of this work, the following SSL models were experimented with:

- **BERT** [33]: The large uncased version was used, with an an output dimension of 1,024.

- **XLM-RoBERTa Spanish** [34]: It is a pre-trained model based on XLM-RoBERTa [35] and trained with Spanish Unannotated Corpora [36]. This model outputs a vector with 1,024 dimensions.
- **BETO** [37]: BETO is one of the first pre-trained models produced with Spanish data. It follows the same structure as the BERT base. Consequently, it outputs a hidden vector of 768 dimensions. It is trained using Wikipedia data and all of the sources of the OPUS Project [38]. Additionally, a fine-tuned version of BETO for emotions was used in the system.

3.3. Classifier

Following the attention pooling process, the resultant vector undergoes further processing via a classifier module. This module is designed to include a layer that adjusts the vector's dimension to fit the desired hidden layer width. It also consists of a stack of hidden layers and an output layer that has a dimension equivalent to the number of classes being considered.

The model architecture consists of several linear layers, each followed by a dropout, layer normalization, and a Gaussian Error Linear Unit (GELU) activation function [39]. The Softmax activation function is used in the output layer to select the predicted class.

4. Attention Pooling

Different approaches have emerged for integrating information extracted from pre-trained models in the field of multimodal learning. Attention mechanisms, particularly Multi-Head Attention (MHA), have been popular in recent years for combining text and speech utterances. This study used an alternative Attention Pooling mechanism used in works such as [40, 41, 42] to reduce the dimensionality of the hidden state vector created by concatenating the outputs of the two pre-trained models.
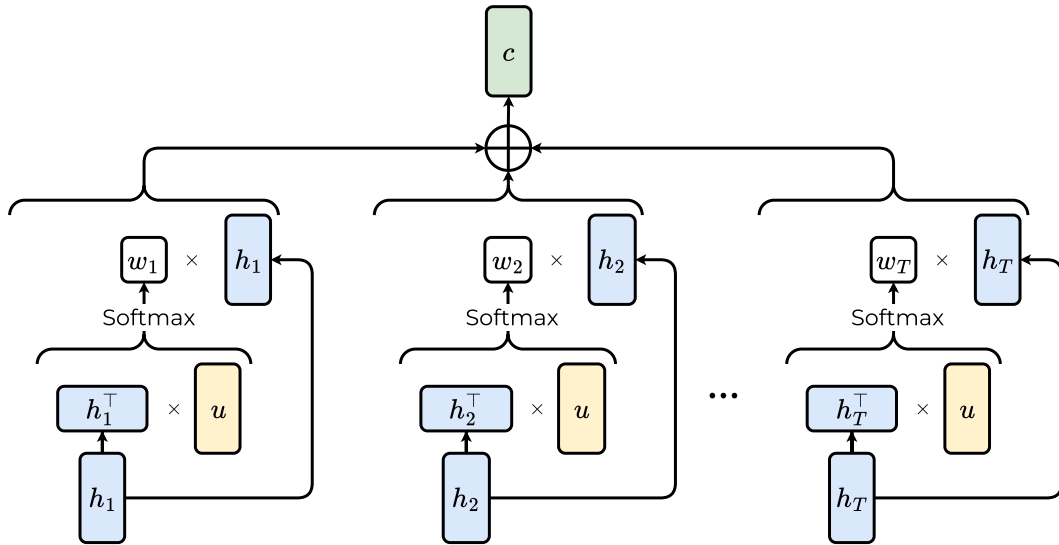

Figure 3: Diagram of the Attention Pooling operation. The hidden vector $h = h_1, ..., h_T$ is pooled into a lower dimension c, using a learnable parameter u.

Considering the embedding dimension E and a batch size of one, we define the hidden states as the sequences of the extracted features $\{h_t \in \mathbb{R}^E | t = 1, ..., T\}$. Then, for each hidden state h_t we calculate its weight as described in Equation (1):

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

where $u \in \mathbb{R}^E$ is a trainable parameter initialized with the Xavier initialization [43] and w_t is the weight associated at the hidden vector h_t. Then, the pooled representation of the hidden vector is calculated using Equation (2).

$$c = \sum_{t=1}^{T} w_t h_t \tag{2}$$

The vector c encapsulates the relevant information of the features extracted in the text and speech systems. This approach is computationally more efficient compared to the general Attention mechanism, where the key, query, and values are calculated. This characteristic is especially convenient for this system due to the scarce data provided. Figure 3 demonstrates graphically the functioning of the attention pooling mechanism.

5. Experimental Setup

PyTorch requires all samples in the batch to have the same dimensions. Therefore, during training, the audio files were cropped using a window of 5.5 seconds, which was the optimal value found. In inference, the whole audio waveform was used. As detailed in Section 3, the audio waveforms were normalized using the mean and standard deviation of the training set. The values extracted were -33.62 and 56.15, respectively. These same values are applied to the other sets when doing inference. Data augmentation techniques were applied by varying the probability based on the capacity of the model. The optimal value was found to be 0.3.

A batch size of 16 samples was selected, as it provided an optimal trade-off between minimizing the duration of each epoch and avoiding GPU memory exhaustion. To further accelerate computation, data parallelization across two GPUs was utilized. In particular, the GPUs employed were two NVIDIA GeForce RTX 2080Ti. The optimizer selected was the AdamW [44], with a learning rate of 0.00005, which decayed by 10% after five epochs without improvements in the validation F1-score. The dropout rate, set at 0.1, was adjusted according to the network's complexity. The number of epochs utilized also depended on the model's capacity. Although model parameters were only stored when the F1-score improved, early stopping was necessary due to the noisy and variable learning curves. This variability could lead to an overfitting model being saved based on local improvements in the F1-Score during validation. Each experiment lasted one to two days, depending on the configurations.

To improve our position on the leaderboard, we made Macro F1-Score our top priority since it was the metric used to evaluate the participants' submissions. This metric combines the Precision and the Recall into one single number by applying their harmonic mean. Specifically, the Macro F1-Score is the average of the F1-Score of each class. This metric treats all classes equally, regardless of their amount of data, making it a fair measure of overall performance. Given that the metric of interest is the macro F1-Score, it is imperative to mitigate the class imbalance present in this dataset. A wide variety of losses try to palliate this disparity. Beyond these possibilities, the loss criterion finally chosen is the weighted cross-entropy loss.

After doing the hyperparameter search, two classical machine learning techniques were employed with the aim of improving the results. The first approach involved applying thresholds to modify the final decision over the logits. However, this did not enhance the results. The second strategy leveraged the variability of different models by using a 3-model ensemble. Hard voting was chosen as the ensemble technique, where the most voted prediction among the three models was selected. In the event of a tie,

the prediction from the model with the highest F1-Score on the validation set was chosen. The code of the project is available here: https://github.com/marccasals98/BSC-UPC_EmoSPeech

6. Results

In the initial stages of the competition, it was necessary to evaluate various pre-trained self-supervised models to determine the most suitable one for the data. At this moment, only the training corpus was available; therefore, it was necessary to make a validation partition to evaluate the performances of the different self-supervised models. Table 1 presents the best results obtained with the diverse text and feature extractors. The best configuration used RoBERTa for text and XLSR-wav2vec 2.0 for audio, achieving an F1-score of 89.73% on the validation set. This superior performance is likely because RoBERTa was trained on Spanish data, making it more effective for this domain than other models trained in English. In addition, XLS-wav2vec 2.0 was trained by using 436,000 hours of audio in 128 languages, including Spanish, which possibly contributed to the improvement of this metric score. It is worth noting that, despite BETO being a Spanish version of BERT, the results with this encoder were poor. This fact could be attributed to WavLM not being trained with as much Spanish data as XLSR-wav2vec 2.0 or that BETO's vector dimension is 768 instead of 1,024, resulting in fewer features captured by the model.

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%

Table 1
Different results were obtained during the validation test with the different feature extractor models. All of these configurations had their corresponding hyperparameter tuning, and the best of each one was selected.

It is remarkable that, in this initial stage, some experiments were conducted with other architectures that involved more parameters. For instance, attempts were made to combine features extracted from the encoder models using multi-head attention (MHA) with one and two heads. These experiments yielded unsatisfactory results, with F1-scores of 84.4% and 82.2%, and the models exhibited significant overfitting. Consequently, it was decided to discontinue these lines of experimentation and concentrate all efforts on the hyperparameter tuning using RoBERTa and XLSR-wav2vec 2.0 and the attention pooling.

The top three different models that obtained the best score achieved 86.20%, 85.96%, and 82.43%. Their confusion matrices over the test set are displayed in Figure 4. As can be seen, anger is the most difficult emotion to classify, normally getting confused with disgust. It is remarkable that, despite being very scarce in the dataset, fear is very separable from the rest of the emotional spectrum.

Section 5 remarked that thresholding techniques failed to outperform the models and were, therefore, discarded. Consequently, the only non-trainable approach used to improve the model's F1-Score was model ensembling. Initially, models with different feature extractors were employed to leverage the diversity of features and create a robust system. However, this approach proved to not be effective. Instead, the three best-performing models on the validation set were ensembled, improving the F1-score to 86.69%. Table 2 compares the results of these approaches with the challenge baseline.
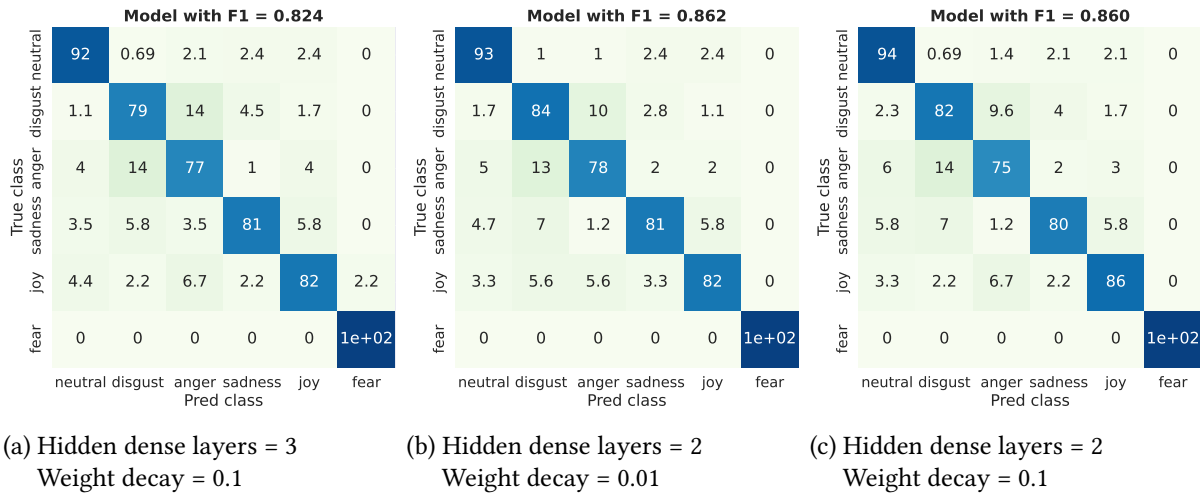
Figure 4: Confusion matrices in the test set. The drop-out was set to 0.1 and the data augmentation probability to 0.3.

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%

Table 2
Results of the top 3 models, the model ensemble (hard voting), and the baseline of the challenge over the test set.

7. Conclusions and future work

This study presented a multimodal model for the emotion recognition challenge EmoSPeech 2024 within the IberLEF 2024 framework, aimed at recognizing emotions from speech and text inputs. The architecture comprised two pre-trained models, one dedicated to speech and the other to text. They extracted feature vectors that the model concatenates into a unified hidden representation vector. On the one hand, for the audio side, different experiments were conducted with WavLM, XLSR-wav2vec 2.0, and HuBERT. On the other hand, the optimization of the textual component of the architecture involved exploration with the models BERT, XLM-RoBERTa for Spanish, BETO, and its finetuned version for emotion. The best performance is achieved by jointly combining RoBERTa and XLSR-wav2vec 2.0.

After the model concatenates the text and speech feature vectors, it employs a dimensionality reduction via reduced attention pooling. This mechanism, with fewer parameters than its standard counterpart, facilitates the seamless integration of text and audio while mitigating the risk of overfitting to the training set. Subsequently, a stack of dense layers processed the output vector, using its compressed information to extract the class prediction. Additionally, to optimize performance and maximize the F1-Score in the competition, model ensembling techniques were adopted, employing hard voting on the top three models. In summary, the system was capable of achieving an F1-Score of 86.69%, an absolute increment of 33.61% compared to the baseline, and securing the first position in the challenge.

After the conclusion of this competition, continuing the research of new paradigms for Speech Emotion Recognition (SER) could be a captivating line of research. One of the lines is improving the efficiency of speech feature extractors. In the paper of WavLM, the authors claim that, in general, most self-supervised learning models (SSL) models focus primarily on Automatic Speech Recognition (ASR) tasks. However, by training an SSL model to jointly learn masked speech prediction and denoising in the pretraining stage, the model's capabilities extend beyond ASR, outperforming other SSL in fields such as SER.

This statement could seem contradictory because, in Section 6, it is proven that XLSR-wav2vec 2.0 outperforms WavLM Large. Nevertheless, this outcome is likely due to XLSR-wav2vec2.0 being trained with a very extensive multilingual dataset. If WavLM was trained with a comparable volume of data, it could potentially outperform XLSR-wav2vec 2.0, leveraging its joint learning of masked speech prediction and denoising to achieve superior performance in various tasks, including Speech Emotion Recognition.

Acknowledgments

References

[1] C. E. Izard, Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues, Annual review of psychology 60 (2009) 1–25. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723854/. doi:10.1146/annurev.psych.60.110707.163539.

[2] N. S. Consedine, J. T. Moskowitz, The role of discrete emotions in health outcomes: A critical review, Applied and Preventive Psychology 12 (2007) 59–75. URL: https://www.sciencedirect.com/science/article/pii/S0962184907000315. doi:10.1016/j.appsy.2007.09.001.

[3] N. Jaques, O. O. Rudovic, S. Taylor, A. Sano, R. Picard, Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation, in: Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing, PMLR, 2017, pp. 17–33. URL: https://proceedings.mlr.press/v66/jaques17a.html, iSSN: 2640-3498.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2023. URL: http://arxiv.org/abs/1706.03762, arXiv:1706.03762 [cs].

[5] A. F. Adoma, N.-M. Henry, W. Chen, Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition, in: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2020, pp. 117–121. URL: https://ieeexplore.ieee.org/document/9317379. doi:10.1109/ICCWAMTIP51612.2020.9317379, iSSN: 2576-8964.

[6] P. Kumar, B. Raman, A BERT based dual-channel explainable text emotion recognition system, Neural Networks 150 (2022) 392–407. URL: https://www.sciencedirect.com/science/article/pii/S0893608022000958. doi:10.1016/j.neunet.2022.03.017.

[7] X. Qin, Z. Wu, T. Zhang, Y. Li, J. Luan, B. Wang, L. Wang, J. Cui, BERT-ERC: Fine-Tuning BERT Is Enough for Emotion Recognition in Conversation, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 13492–13500. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26582. doi:10.1609/aaai.v37i11.26582.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805 [cs].

[9] M. D. Pell, L. Monetta, S. Paulmann, S. A. Kotz, Recognizing Emotions in a Foreign Language, Journal of Nonverbal Behavior 33 (2009) 107–120. URL: https://doi.org/10.1007/s10919-008-0065-7. doi:10.1007/s10919-008-0065-7.

[10] A. B. Ingale, D. S. Chaudhari, Speech Emotion Recognition 2 (2012).

[11] M. Xu, F. Zhang, W. Zhang, Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset, IEEE Access 9 (2021) 74539–74549. URL: https://ieeexplore.ieee.org/document/9381872/. doi:10.1109/ACCESS.2021.3067460.

[12] V. Singh, S. Prasad, Speech emotion recognition system using gender dependent convolution

neural network, Procedia Computer Science 218 (2023) 2533–2540. URL: https://www.sciencedirect.com/science/article/pii/S1877050923002272. doi:10.1016/j.procs.2023.01.227.

[13] M. Mohan, P. Dhanalakshmi, R. S. Kumar, Speech Emotion Classification using Ensemble Models with MFCC, Procedia Computer Science 218 (2023) 1857–1868. URL: https://www.sciencedirect.com/science/article/pii/S1877050923001631. doi:10.1016/j.procs.2023.01.163.

[14] M. Macary, M. Tahon, Y. EstÃ¨ve, A. Rousseau, On the Use of Self-Supervised Pre-Trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 373–380. URL: https://ieeexplore.ieee.org/abstract/document/9383456. doi:10.1109/SLT48900.2021.9383456.

[15] L. Pepino, P. Riera, L. Ferrer, Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, 2021. URL: http://arxiv.org/abs/2104.03502. doi:10.48550/arXiv.2104.03502, arXiv:2104.03502 [cs, eess].

[16] Z. Zhao, Y. Wang, Y. Wang, Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition, 2022. URL: http://arxiv.org/abs/2207.04697. doi:10.48550/arXiv.2207.04697, arXiv:2207.04697 [cs, eess].

[17] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, R. Valencia-García, Spanish MEACorpus 2023: A multimodal speech-text corpus for emotion analysis in spanish from natural environments, Computer Standards & Interfaces (2024) 103856.

[18] P. Ekman, H. Oster, Facial Expressions of Emotion, Annual Review of Psychology 30 (1979) 527–554. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.ps.30.020179.002523. doi:10.1146/annurev.ps.30.020179.002523, publisher: Annual Reviews.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised Cross-lingual Representation Learning for Speech Recognition, 2020. URL: http://arxiv.org/abs/2006.13979, arXiv:2006.13979 [cs, eess].

[20] L. Lange, H. Adel, J. Strötgen, Boosting transformers for job expression extraction and classification in a low-resource setting, CoRR abs/2109.08597 (2021). URL: https://arxiv.org/abs/2109.08597. arXiv:2109.08597.

[21] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, F. García-Sánchez, R. Valencia-García, Overview of EmoSPeech 2024@IberLEF: Multimodal Speech-text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[22] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. arXiv:2212.04356.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, F. Wei, WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518. URL: http://arxiv.org/abs/2110.13900. doi:10.1109/JSTSP.2022.3188113, arXiv:2110.13900 [cs, eess].

[25] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, E. Dupoux, Librilight: A benchmark for asr with limited or no supervision, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020. URL: http://dx.doi.org/10.1109/ICASSP40776.2020.9052942. doi:10.1109/icassp40776.2020.9052942.

[26] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, Z. Yan, Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio, 2021. arXiv:2106.06909.

[27] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, E. Dupoux, Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised

learning and interpretation, 2021. `arXiv:2101.00390`.

[28] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020. URL: http://arxiv.org/abs/2006.11477, arXiv:2006.11477 [cs, eess].

[29] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, 2020. `arXiv:1912.06670`.

[30] M. J. F. Gales, K. Knill, A. Ragni, S. P. Rath, Speech recognition and keyword spotting for low-resource languages: Babel project research at cued, in: Workshop on Spoken Language Technologies for Under-resourced Languages, 2014. URL: https://api.semanticscholar.org/CorpusID:7439227.

[31] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, MLS: A Large-Scale Multilingual Dataset for Speech Research, in: Interspeech 2020, 2020, pp. 2757–2761. URL: http://arxiv.org/abs/2012.03411. doi:`10.21437/Interspeech.2020-2826`, arXiv:2012.03411 [cs, eess].

[32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, 2021. URL: http://arxiv.org/abs/2106.07447, arXiv:2106.07447 [cs, eess].

[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. `arXiv:1810.04805`.

[34] L. Lange, H. Adel, J. Strötgen, Boosting transformers for job expression extraction and classification in a low-resource setting, in: Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings, 2021. URL: http://ceur-ws.org/Vol-2943/meddoprof_paper1.pdf.

[35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. `arXiv:1911.02116`.

[36] J. Cañete, Compilation of large spanish unannotated corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731. doi:`10.5281/zenodo.3247731`.

[37] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[38] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

[39] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, CoRR abs/1606.08415 (2016). URL: http://arxiv.org/abs/1606.08415. `arXiv:1606.08415`.

[40] M. India, P. Safari, J. Hernando, Self multi-head attention for speaker recognition, CoRR abs/1906.09890 (2019). URL: http://arxiv.org/abs/1906.09890. `arXiv:1906.09890`.

[41] F. Costa, M. India, J. Hernando, Double multi-head attention multimodal system for odyssey 2024 speech emotion recognition challenge, 2024. `arXiv:2406.10598`.

[42] M. India, P. Safari, J. Hernando, Double multi-head attention for speaker verification, 2021. `arXiv:2007.13199`.

[43] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Y. W. Teh, M. Titterington (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of *Proceedings of Machine Learning Research*, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[44] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. `arXiv:1711.05101`.