THAU-UPM at EmoSPeech-IberLEF2024: Efficient Adaptation of Mono-modal and Multi-modal Large Language Models for Automatic Speech Emotion Recognition

Sergio Esteban-Romero[1], Jaime Bellver-Soler[1], Iván Martín-Fernández[1], Manuel Gil-Martín[1], Luis Fernando D'Haro[1] and Fernando Fernández-Martínez[1]

[1]*Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunication Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM)*

Abstract

Automatic Speech Emotion Recognition (SER) is a pivotal task across domains such as psychology, healthcare, or human-computer interaction. This study, conducted within the framework of the EmoSpeech challenge at IberLEF 2024, explores efficient adaptation techniques for mono-modal and multi-modal large language models (LLMs) for SER in Spanish. The challenge includes two tasks: one that relies solely on textual transcriptions and another that integrates audio signals. For the text-only task, we fine-tune the multilingual Gemma-2B model using the Low-Rank Adaptation (LoRA) method, optimizing low-rank adaptation parameters. For the multi-modal task, we employ Qwen-Audio-Chat, enhancing its performance using the LoRA technique, and we propose a novel Whisper-Gemma model that integrates Whisper-large-v3 audio encoder with the Gemma LLM, training only a projection layer. The metrics obtained as a result of the experimentation process demonstrate the potential of both approaches. The fine-tuned Gemma-2B model achieves an f1-macro score of 0.6094 on the text-only task, while the Qwen-Audio-Chat model reaches 0.8248, indicating significant improvements in emotional recognition capabilities when combining both modalities. Additionally, the Whisper-Gemma model achieves a competitive f1-macro score of 0.7904, underscoring the effectiveness of using pre-trained audio encoders and smaller LLMs in SER tasks. These findings highlight the value of parameter-efficient fine-tuning methods and the integration of robust audio encoders with LLMs to enhance SER performance.

Keywords

Speech Emotion Recognition, MultiModal Large Language Models, Low-Rank Adaptation, Parameter-efficient Fine-Tuning

1. Introduction

Understanding human emotions is crucial in various domains ranging from psychology and healthcare to human-computer interaction and marketing [1]. Accurately recognizing them can provide valuable information on human behavior, preferences, or mental health, among others. Although emotions manifest in multiple modalities, our study focuses specifically on spoken Spanish audio and its transcripts as sources of emotional expression. In addition, a relationship between the topic of conversation and the emotion expressed has been studied in the literature [2].

This work contributes to the EmoSPeech challenge at Iberlef 2024 [3]. It poses the problem of automatic Speech Emotion Recognition (SER) through two distinct tasks: one that uses only textual transcriptions and the other that integrates the audio signals corresponding to such transcriptions [4]. By focusing on both modalities, we aim to capture the nature of emotional expression while leveraging the strengths of combining both modalities for enhanced accuracy and robustness.

IberLEF 2024, September 2024, Valladolid, Spain

sergio.estebanro@upm.es (S. Esteban-Romero); jaime.bellver@upm.es (J. Bellver-Soler); ivan.martinf@upm.es (I. Martín-Fernández); manuel.gilmartin@upm.es (M. Gil-Martín); luisfernando.dharo@upm.es (L. F. D'Haro); fernando.fernandezm@upm.es (F. Fernández-Martínez)

0009-0008-6336-7877 (S. Esteban-Romero); 0009-0006-7973-4913 (J. Bellver-Soler); 0009-0004-2769-9752 (I. Martín-Fernández); 0000-0002-4285-6224 (M. Gil-Martín); 0000-0002-3411-7384 (L. F. D'Haro); 0000-0003-3877-0089 (F. Fernández-Martínez)

The human voice is a complex signal that condenses a wealth of information about the speaker, including their age, gender, health, and even emotional state [5]. Subtle variations in pitch, tone and rhythm capture these nuances. Analyzing these acoustic features along with textual transcripts, we can learn more about the emotional content of the spoken words, helping to recognize emotions more accurately and appropriately in different contexts. In detail, the challenge dataset consists of a set of 6 emotions: neutral, disgust, anger, joy, sadness and fear.

Our approach for the challenge is based on leveraging the strengths of a family of Language Models that, although smaller in size than the top performing LLMs [6, 7], have been trained on a vast amount of data spanning multiple disciplines and have therefore acquired a solid grounding of knowledge about our world. In particular, we focus on the Gemma [8] and Qwen [9, 10] models. More specifically, we explore parameter-efficient methods for adapting this pre-training sapiens to the context of SER, either using written language-only approaches or imbuing them with audio capabilities. When only textual transcripts are considered, we explore how adapting Gemma using Low-Rank Adaptation (LoRA) [11] and SER related data can capture semantic and syntactic information related to emotional patterns. Alternatively, when incorporating audio data, we fine-tune state-of-the-art MultiModal Large Language Models (MM-LLMs) so the LLM also has the ability to comprehend not only textual content but also the acoustic features extracted by the audio encoder, thus enriching the understanding of emotions across both modalities. In particular, we choose to fine-tune Qwen Audio [10] due to its demonstrated state-of-the-art performance across various benchmarks, leveraging its pre-trained capabilities. Additionally, to evaluate the trade-off between accuracy and computational efficiency, we compare Qwen-Audio's performance with a smaller model. This lightweight model integrates features extracted by an audio encoder into the Gemma LLM. Here, only the projection layer that maps the audio encoder output to the LLM semantic space is trained, allowing the model to effectively handle the audio data.

The contributions to emotion recognition described in this work can be summarized as:

- Leveraging large language models (LLMs): We fine-tune LLMs to directly predict emotions from textual transcripts. This approach allows us to obtain probability distributions for the emotions under analysis.
- Enhancing classification with audio features: We investigate the potential of incorporating audio features by fine-tuning multimodal LLMs (MM-LLMs). This aims to enhance the classification capabilities for emotion recognition.
- Exploring the efficiency of adapting and using small LLMs: We explore training a projection layer that allows pre-trained encoders (inspired by MM-LLMs) to be used independently with the LLM for emotion recognition. This approach investigates the possibility of achieving good results with computationally less expensive models.

This paper is organized as follows: In Section 2, we provide a comprehensive review of related work, discussing traditional and deep learning approaches for SER and recent advancements in multimodal strategies and LLMs. Section 3 details our proposed methods for the EmoSpeech challenge, including the fine-tuning of Gemma-2B using LoRA for text-only classification and the adaptation of Qwen-Audio-Chat and Whisper-Gemma models for multimodal classification. Section 4 describes the materials and methods used in our experiments, including data sources, experimental setup, and hyperparameter configurations. In Section 5, we present our results and discussion, comparing in-house validation results and official challenge test results for both tasks. Finally, Section 6 concludes the paper, summarizing our findings and outlining future research directions to further improve automatic SER systems.

2. Related Work

Traditionally, SER approaches have relied on extracting meaningful descriptors of the audio signal. This process often involved determining the most effective features to identify emotions, including but not limited to Fast Fourier Transform (FFT) and Mel-Frequency Cepstral Coefficients (MFCC) [12]. This

research path travelled through approaches that exploit these and other features in order to train classical Machine Learning models such as Support Vector Machines (SVMs) [13], Linear Discriminant Analysis (LDA) [14], k-Nearest Neighbors [15] or even unsupervised clustering algorithms [16]. However, these types of approach have struggled to capture the specifics of emotion in audio [1]. With the advent of Deep Learning methods came systems that no longer rely on manually curated features but classified emotions directly from the audio signal or its spectral representations, using architectures such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) Networks [17, 18, 19].

Studies in Spanish SER lack abundance, mainly due to the sparcity of available data. However, recent efforts go beyond the audio signal alone and incorporate textual transcriptions in hybrid approaches using deep audio encoders and text models and late fusion strategies [20, 2].

The recent use of multimodal strategies for SER based on textual models motivates a shift of focus towards Large Language Models (LLMs), that have been revolutionizing the field of Natural Language Processing (NLP) for the past few years. Although the most capable LLMs contain hundreds of billions of parameters, after recent advancements, some models with significantly fewer parameters have shown great adaptability, enabling them to tackle complex tasks efficiently. Examples include Qwen-7B [9] with 7 billion parameters, developed by Alibaba and Google Gemma-2B [21] with a more compact size of 2 billion parameters. These models demonstrate impressive performance despite their relatively small size.

The Qwen [9] LLM family is based on the transformer-based LlaMa architecture [22]. However, Qwen models introduce several modifications such as untiding weights for input embedding and output projection, using a rotary positional embedding [23], removal of the bias input in all layers and an adjusted activation function among others. These refinements lead the Qwen-7B model to outperform Llama-7B on various benchmarks. However, a key limitation of Qwen is its training data, which is formed by English and Chinese limiting its usage in multilingual settings unless it is adapted.

The Gemma 2B model, despite its relatively smaller size with 2 billion parameters, showcased a competitive performance on different benchmarks when compared to larger models such as Llama-7B [8].

Another interesting example of a relatively small LLM is Microsoft Phi-2, which provides outstanding results while maintaining a relatively low number of parameters [24]. It follows the training methodology presented in "Textbooks are all you need" [24], which emphasizes the usage of high-quality data to enhance performance when using smaller models.

Recent advances have expanded the scope of LLMs by incorporating multimodal capabilities, so that they can now process and understand various data types at the same time beyond text, such as audio and images. By simultaneously processing different data modalities, MM-LLMs benefit from the semantic and syntactic understanding acquired by the LLM to extend its comprehension strengths over new domains such as image or audio processing. Typically, they are formed by encoders that convert each data type (e.g., image, audio) into a representation that is translated using a projection layer into a format the LLM can understand [25]. Finally, the output provided by the LLM combines the information from all modalities, capturing the relationships between them.

As an example, Qwen-Audio extends the capabilities of the Qwen-7B LLM with audio understanding by integrating a fine-tuned version of Whisper-large-v2 as an audio encoder with projection layer to serve as an interpreter of the audio features for the text encoder. Whisper [26] is a state-of-the-art audio encoder that uses 80-channel log-mel spectrograms for speech recognition and audio transcription using a transformer-based architecture. Despite the fact that Whisper models are trained for speech recognition and translation, they show good performance in a variety of tasks while using their audio representations. It obtains outstanding results for transcription tasks due to its powerful embedding representations, which are proven to contain relevant information about the sounds that form an audio file [27, 28]. As a result, it provides Qwen-Audio with great reasoning capabilities and understanding about the audios given as input.

Although Qwen-audio uses a fine-tuned version of Whisper-large-v2 as its audio encoder, recently an upgraded version of this speech recognition model has been developed. Whisper-large-v3 follows a similar architecture to that of its previous version but increases the size of the input by using 128 mel

frequency bins instead of 80 [26]. This architectural upgrade and higher performance motivate its use over its predecessors in building systems that solve a downstream task.

Although impressive in size and capabilities, Whisper V3 has not been trained explicitly for the SER problem. This issue is tackled by emo2vec [29], a model that although smaller in parameters is pre-trained with a specific focus on SER tasks, thus being able to generate audio representations that pose as more meaningful for the challenge at hand. Emo2vec uses a CNN as a feature extractor followed by a linear projection layer and a mask operation to obtain the input for the backbone network which is based on the data2vec architecture [30]. The backbone networks form the core component responsible for learning high-level features from the input data [31]. To the best of our knowledge, a comparative analysis between the strengths of a largely pre-trained, generalist audio model (Whisper) and smaller but more task-oriented approaches (emo2vec) on the SER task is yet to be carried out, which motivates further exploration of both.

While MM-LLMs have achieved outstanding results in a wide variety of tasks, training them is challenging due to their large number of parameters which turn into a huge computational demand. LoRA [11] is a technique designed to address these challenges by allowing to efficiently fine-tune large models. It is based on the idea that most of the information of the original matrix which stores the updates of the model is captured in matrices with a lower intrinsic rank. Therefore, it can be decomposed into a product of lower rank adaptation matrices that will significantly reduce the number of trainable parameters compared to traditional fine-tuning, leading to faster training and lower memory requirements.

In this work, we focus on fine-tuning the Gemma-2B model and the Qwen-Audio model. Furthermore, we propose a novel architecture where we explore the potential of utilizing a combination of smaller, pre-trained LLMs, like Gemma or Phi, alongside the Whisper-large-v3 and emo2vec audio encoders.

3. Proposal

3.1. Task 1 - Fine-tune Gemma using LoRA

Our approach for the Speech Emotion Recognition through text transcription task works under the hypothesis that a fairly sized LLM that has been trained on a huge amount of multilingual data, such as Gemma [21], has the potential to understand and model the emotion of the speaker through the nuances embedded in the written language with few adjustments. For that reason, we fine-tune the Gemma model using the LoRA method on the transcriptions provided for the challenge dataset. In particular, we use the `google/gemma-1.1-2b-it` checkpoint from the HuggingFace Hub[1] as a starting point. As reported in the model card, the `1.1` family of Gemma models are instruction tuned using a novel Reinforcement Learning with Human Feedback (RLHF) method that enhances their truthfulness and ability to follow instructions.

We train our proposed system using the next token prediction task, i.e., we model the probability distribution of the next token x_n, $p(x_n|\{x_0, x_1, ...x_{n-1}\})$, given the previous sequence of tokens, or *prompt*, $\{x_0, x_1, ...x_{n-1}\}$. The prompt used for the training process can be found in Table 1, where the tag <transcription> is replaced on each case with the corresponding text transcription. Since the Gemma model has been trained as a conversational agent, we pre-process the prompt by applying a chat template that instructs the model to predict the next conversational turn:

`<start_of_turn>user\n[prompt]<end_of_turn>\n<start_of_turn>model`

where `[prompt]` is the original instruction prompt as shown in Table 1. During training, the label is appended as the first word of the `model` turn, so that our system learns to generate that word given the preceding context (in our case, x_n would be the emotion assigned to a given sample and $\{x_0, x_1, ...x_{n-1}\}$ represents the prompt that includes a transcription of the audio belonging to that

[1] https://huggingface.co/google/gemma-1.1-2b-it

Table 1
Prompt used to both perform zero-shot evaluation and to fine-tune the model using LoRA

Model	Prompt
Gemma 2B	Instruction: Given the following audio transcription, predict the emotion of the speaker. The emotion can be one of the following: [neutral, sad, angry, happy, surprise, fear, disgust, contempt] The transcription is:<transcription>
Qwen-Audio-Chat	Select the predominant emotion from neutral, disgust, anger, joy, sadness and fear. Answer only with one of the emotion words and nothing else. Consider also its transcription: <transcription>

sample). By modeling next token prediction, the LLM is able to elicit the nuances involved in the language behind each subject and how they translate into the perceived emotion.

In the spirit of leveraging the immense capabilities of state-of-the-art Language Models and adapting them to the downstream task in the most efficient manner, we use the LoRA technique in order to build our systems for this subtask. This way, instead of adapting the whole set of parameters that constitute the LLM, a low rank adaptation of each weight matrix is appended and trained whilst keeping the original model untouched. This “addenda”, which requires significantly less computational budget to be trained, can be used in combination with the pre-trained LLM in order to modify its original behaviour, in this case with the aim of turning it into a SER system. Mathematically speaking, a forward pass over any given layer of the adapted model can be formulated as:

$$h = W_0 x + \frac{\alpha}{r} \Delta W x \tag{1}$$

where W_0 is the original weight matrix of the Transformer layer that is not subject to backpropagation, ΔW is the LoRA matrix of rank r that is trained at this stage, and α is a design hyperparameter that controls the prevalence given to the transformations carried out by the LoRA surrogate with respect to the original LLM. This process enables learning a low-rank matrix representation of the original model that is easily trainable with the available resources and data, thus effectively combining large-scale pre-training knowledge with task-aware capabilities. The two design parameters associated with this technique, α (LoRA alpha) and r (LoRA rank), may be definitive on the overall performance of the final system. To evaluate this, we employ a 5-fold cross-validation (CV) scheme to compare different combinations of LoRa parameters, allowing us to identify the hyperparameter combination that yields the best average f1-macro score.

3.2. Task 2 - Fine-tuning Qwen-Audio using LoRA

Our proposal for the SER task that uses the audio and corresponding text transcriptions is based on fine-tuning the Qwen-Audio-Chat [10] model. Figure 1 shows the architecture of Qwen-Audio-Chat model used for speech emotion classification. Our interest comes from its strengths, which are that it has been trained using huge amounts of data from diverse datasets across various tasks and it has shown remarkable performance on various benchmarks. In addition, SER is one of its pre-training objectives, making it a strong foundation for our study. We specifically utilized the `Qwen-Audio-Chat`[2] checkpoint, which has improved reasoning and understanding capabilities due to an additional supervised training stage.

However, the output of the model often includes the predicted emotion within a sentence, requiring a post-processing step to extract the actual predicted label. In case no emotion is returned or it falls outside from those expected, the predominant emotion ’neutral’ will be assigned. The prompt used to evaluate model performance is specified in Table 1, where the tag <transcription> is replaced on each case with the transcription corresponding to the analyzed audio sample.

To tailor the model for SER and address the output format, we fine-tuned the LLM in the Qwen-Audio-Chat architecture employing the LoRA technique [11]. The audio encoder remains frozen during

---

[2] https://huggingface.co/Qwen/Qwen-Audio-Chat

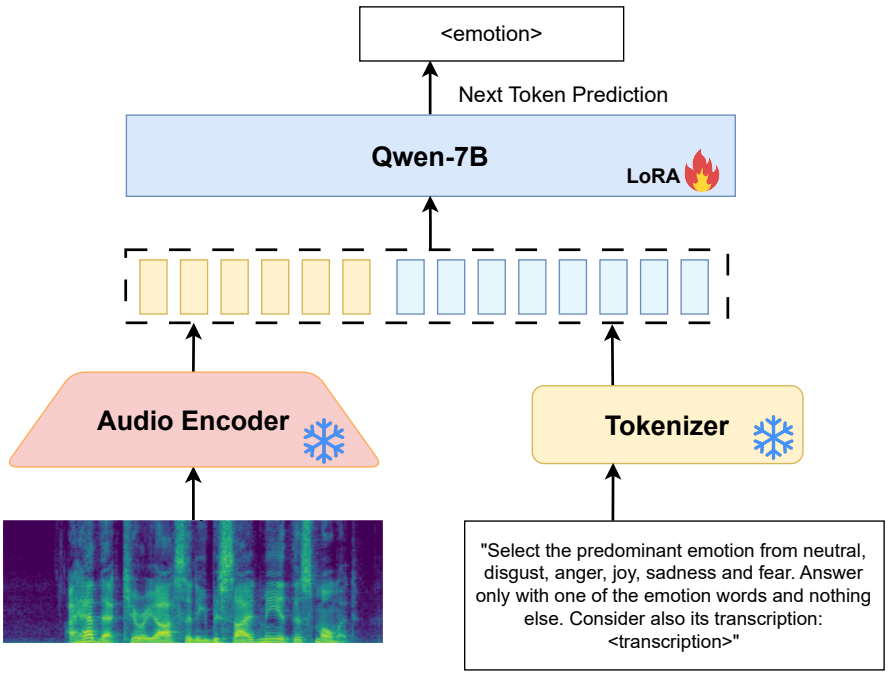

Figure 1: The diagram illustrates the architecture of Qwen-Audio [10] model used as an speech emotion classifier. The Audio Encoder and the Tokenizer provide a sequence of tokens that represent the audio signals and prompt with the transcriptions given as input, respectively. These sequences are concatenated and processed by the Qwen-7B LLM which has been fine-tuned using LoRA to provide the most likely emotion based on the combined information. Snowflakes represent the components that remain frozen throughout the training process, while the flame indicates those adapted via fine-tuning. The <emotion> is expected to be one among those specified in the input prompt.

the entire process. LoRA facilitates efficient training of large models, greatly reducing computational demands as described in Section 3.1. Here, we have also conducted a 5-Fold CV process to identify the LoRA rank and alpha configuration that achieves the best performance based on the average f1-macro. The chosen hyperparameters will be applied to the final model evaluated in the challenge test dataset. For this fine-tuning process, we use the Swift framework [32] due to its support for training MM-LLMs.

3.3. Task 2 - MM-LLM: Gemma + Whisper

We propose a model architecture, see Figure 2, that integrates both audio and textual information inspired on Qwen-Audio model's structure. However, the core difference with Qwen-Audio [10] or similar approaches [9, 33] is that only the projection layer is trained, as we hypothesize that it will allow the language model to comprehend the audio features extracted by the audio encoder. Another relevant modification in our proposal regarding Qwen-Audio [10] is that we employ Whisper-large-v3 [26] as the audio encoder, instead of Whisper-large-v2, since it processes spectrograms with higher resolution as input, 128 mels filters instead of 80 [26]. In addition, we prioritize efficiency by employing smaller LLMs, reducing in more than half the parameters of the original Qwen-Audio model [21, 24].

Our model employs a frozen LLM as the textual understanding backbone and a frozen audio encoder tailored for the extraction of acoustic features. We employ a linear layer as a projector that leverages the communication between the audio encoder and the LLM. To address the imbalanced emotion distribution present in the dataset, we applied a weighted cross-entropy loss function during training, assigning higher weights to underrepresented classes to enhance model performance and robustness. Then, we extract the final logits of the LLM that correspond to each emotion token to ensure the LLM to categorize with just the provided emotion labels. Finally, for inference, we apply a softmax layer to

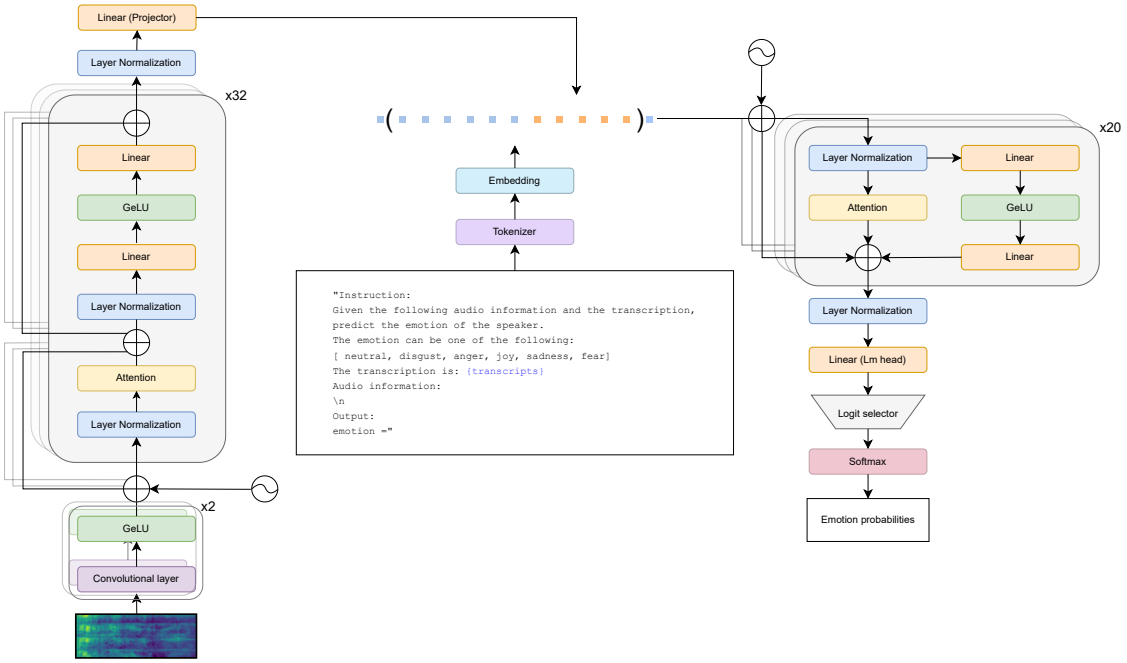

Figure 2: Architecture for processing audio and textual modalities with an LLM. On the left, the audio encoder processes the audio signal and a projection layer maps the extracted audio features into the LLM domain. This Linear (Projector) module is trained from scratch. Then, information from both modalities is combined and used as input to the LLM, represented on the right, which remains frozen the whole process. Finally, the logits corresponding to all the emotions considered are gathered to obtain the probability distribution of all emotions.

obtain the probability distribution of all the possible emotions.

In the end, audio content and temporal dynamics are captured and processed through the Mel spectrograms received by Whisper-large-v3 as input providing a set of features that are merged with the hidden representations provided by the LLM. The purpose of the introduced projection layer is to obtain a unified representation by mapping audio features into the semantic space of the LLM. This enables the LLM to effectively interpret the audio data and leverage any latent emotional patterns present.

4. Materials and Methods

4.1. Data Source

The official data for the challenge come from the Multimodal speech–text corpus for Emotion Analysis in Spanish from natural environments (MEACorpus) 2023 Dataset [2]. A pioneer source of multimodal data annotated in terms of emotion in Spanish, it consists of a collection of audio segments extracted from YouTube videos uploaded to Spanish channels. The audios, lasting 9 seconds approximately on average, are annotated in terms of Ekman's six basic emotions: anger, disgust, sadness, joy, fear, and surprise, as well as a neutral class. A subset of the Spanish MEACorpus 2023 Dataset is used for the task, comprising a grand total of 3,750 audios split into a train (3,000 samples) and test (750 samples) subset. The corpus includes both processed audio segments and automatically generated transcriptions, as well as the golden labels for the train split.

The distribution of labels for the training data is shown in Figure 3. Firstly, as originally reported by the authors, it can be seen that there are no examples of the "surprised" class in the dataset. Secondly, there are clear signs of class imbalance in the data, with "fear" being heavily misrepresented.

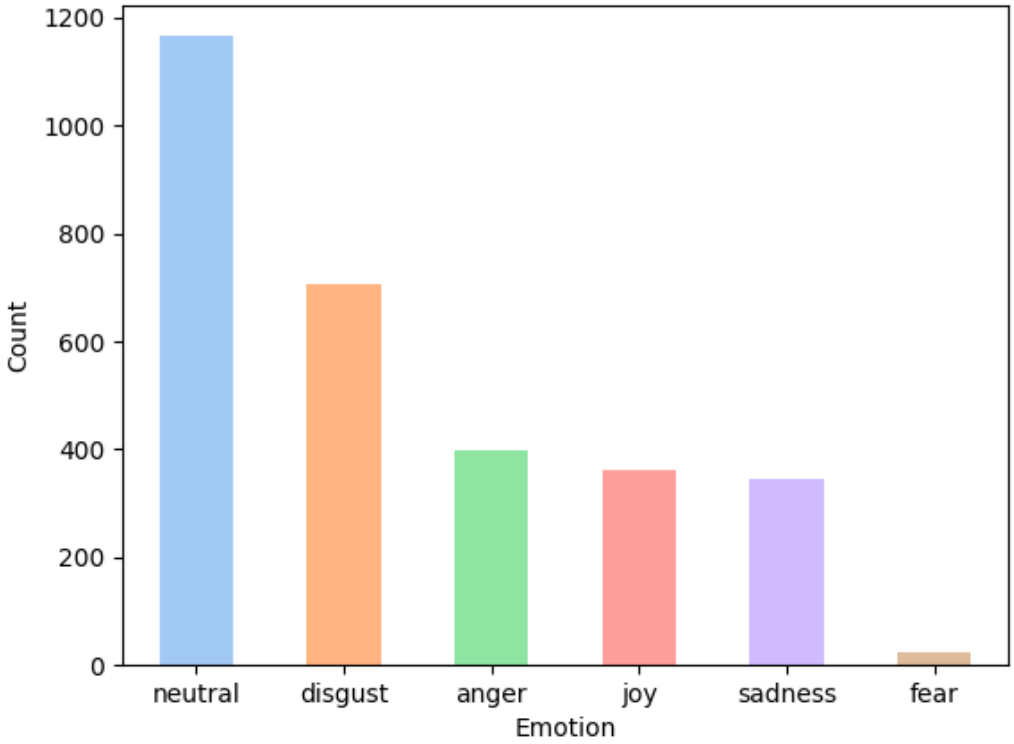

Figure 3: Label distribution for the train partition of the MEACorpus 2023.

4.2. Experimental Setup

In order to validate our approaches in-house before submitting the challenge runs, a Stratified 5-Fold Cross-Validation schema is used in all our experiments, which benefits robustness and statistical relevance. In this way, the training data are split 5 times into training and validation, with the validation splits of each fold being disjoint. Furthermore, the data splits are stratified, i.e. they contain approximately the same distribution of labels as the entire training corpus, thus enforcing similar experimental conditions across the folds. As validation metric, we use the f1-macro score, the official figure of merit for the task, which does not take into account class imbalance.

For the challenge runs, we utilized the entire train dataset to obtain our final models. For Task 1, all models are trained for 6 epochs using the Adam optimizer, a batch size of 4 and a learning rate of 10^{-4} with a decreasing linear scheduler to a final value of 0. For task 2, Qwen-Audio-Chat is fine-tuned for 6 epochs using Adam optimizer with weight decay of 0.1, a batch size of 3 and a learning rate of 10^{-4} with a decreasing linear scheduler. All experiments were carried out on a single NVIDIA A100 40GB GPU.

5. Results and Discussion

5.1. In-house results

5.1.1. Task 1 - Text only classification

When adapting LLMs to a downstream task using LoRA, the alpha and rank parameters become crucial in the model design. These parameters effectively control the complexity of the "surrogate model" learned by LoRA and how much it overrides the original LLM. To find the optimal settings for our task, we performed a 5-fold cross-validation (CV) on the challenge's training data using various alpha and rank combinations. The results are visualized in Figure 4 as a heatmap of the mean f1-macro scores for the 5 splits in our CV setup for each configuration. The figure reveals a clear sensitivity to the alpha parameter. The best performance occurs when alpha is close to the chosen rank, but it drops

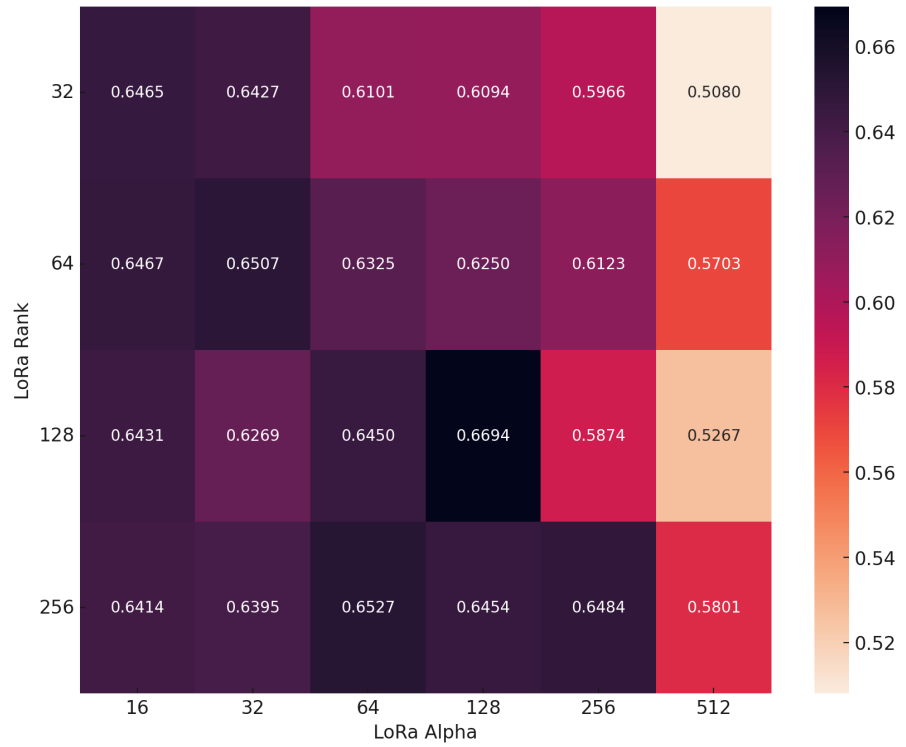

Figure 4: Average f1-macro score after 5-Fold CV for different LoRA parameters configuration fine-tuning Gemma 2B model.

significantly at a value of 512. This suggests that overly influential LoRA adaptations can diminish the value of the pre-trained LLM knowledge. Conversely, very low alpha values lead to suboptimal results, as the task-specific knowledge is underutilized.

Regarding the rank parameter, for the optimal alpha, higher values generally improve performance. A peak f1-macro score of **0.6694** is achieved with both rank and alpha set to 128. However, this benefit decreases at rank 256. Therefore, it can be empirically induced that, for this specific problem and model configuration, the original weight matrices are best described using a rank 128 decomposition.

5.1.2. Task 2 - Multimodal classification

To evaluate Qwen-Audio-Chat [10] performance, we performed a zero-shot evaluation on the entire train dataset achieving an f1-macro result of **0.27** indicating a misalignment with the task requirements. As shown in Table 1, our prompts specified that only the emotion should be provided. However, the model struggled to follow these instructions accurately, often adding unexpected information to its responses. Consequently, a post-processing step was required to retrieve the actual emotion provided. This suggests difficulty following instructions and potential language mismatch, as Qwen-Audio (based on Qwen-7B [9]) was trained primarily on Chinese and English data. Additionally, pre-training data for SER tasks might use different emotion taxonomies than this study.

To address these issues, we fine-tuned Qwen-Audio-Chat using LoRA. Figure 5 presents the results of a 5-Fold CV for different LoRA configurations. The F1-macro score generally increased with the LoRA alpha parameter, indicating the benefit of task-specific adaptation. However, excessively high alpha values (above 4,096) yielded lower f1-macro scores, suggesting that the importance given to the newly learned weights is so high that the original model knowledge is being sidelined. Regarding rank, values above 64 showed minimal performance improvement. In fact, the best configuration achieved a f1-macro score of **0.8164** with a LoRA rank and alpha of 64 and 1,024, respectively.

We further explored multimodal approaches by combining audio encoders with the Gemma LLM. First, we evaluated both Whisper-large-v3 and Emo2vec encoders with Gemma-2B using a batch size

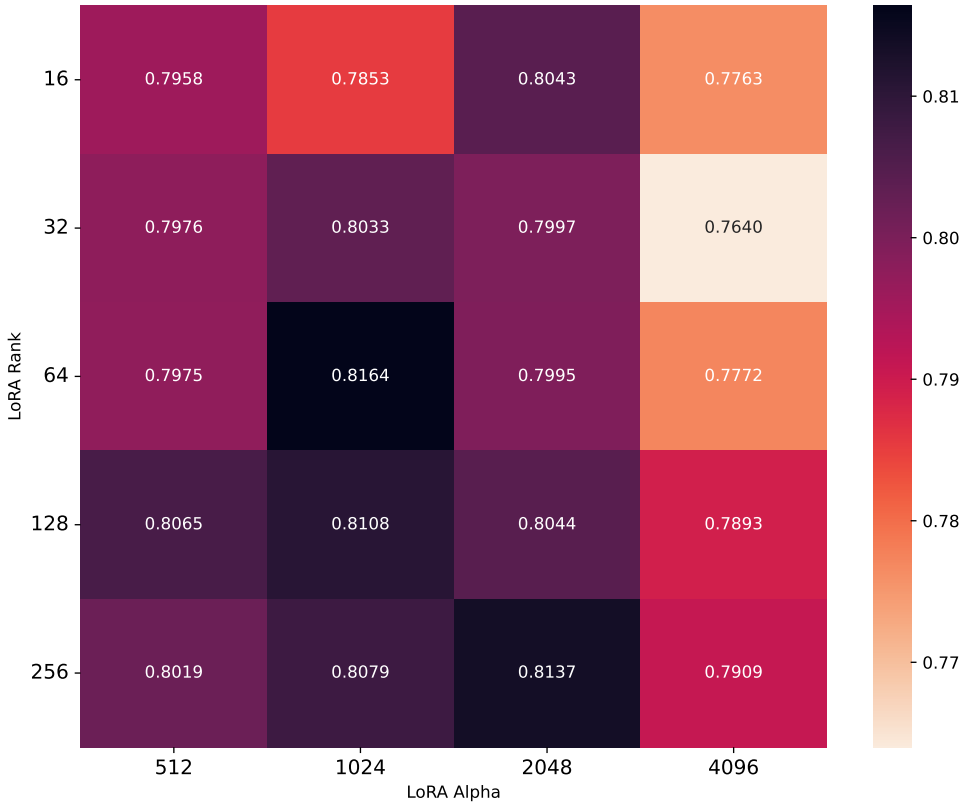

Figure 5: Average f1-macro score after 5-Fold CV for different LoRA parameters configuration fine-tuning Qwen-Audio-Chat model.

Table 2
Average f1-macro results after 5-Fold CV for different hyperparameters configuration for Whisper-Gemma model.

Whisper + Gemma		
Batch size	**Learning rate**	**Validation f1-macro**
4	10^{-3}	**0.8125**
4	10^{-4}	0.7660
6	10^{-3}	0.7515
6	10^{-4}	0.7404

of 4 and learning rate 10^{-4}. Our results on the official test set showed that the Whisper-based model significantly outperformed Emo2vec (0.6636 F1-macro vs. 0.5134).

To optimize the Whisper-Gemma model, we conducted a hyperparameter search on batch size and learning rate. Table 2 shows the validation f1-macro scores obtained from the different combinations of hyperparameters. Despite the potential benefits of larger batch sizes in terms of convergence speed and computational efficiency, hardware limitations restricted us to batch sizes of 4 and 6. Interestingly, our experimentation reveals that the model trained with a batch size of 4 achieved a higher validation f1-macro score compared to the one trained with a batch size of 6. Our best model scored an average f1-macro of 0.8125 after 5-fold CV.

Table 3
Official challenge f1-Macro results over the 750 samples of the test set for various models and tasks, along with 95% Confidence Intervals.

Task 1	
Model	**Test f1-macro**
Gemma 2B (Alpha = 32 | Rank = 32)	0.5814 ± 0.0352
Gemma 2B (Alpha = 128 | Rank = 128)*	0.6094 ± 0.0349
Task 2	
Model	**Test f1-macro**
Qwen-Audio-Chat (Alpha = 2,048 | Rank = 256)	0.8248 ± 0.0272
Qwen-Audio-Chat (Alpha = 1,024 | Rank = 64)	0.8072 ± 0.0282
Whisper - Gemma	0.7904 ± 0.0291
Emo2Vec - Gemma	0.5134 ± 0.0359

* *This result was obtained at the post-evaluation stage.*

5.2. Challenge runs

Table 3 summarizes the performance metrics achieved on the official test set containing 750 examples. For Task 1, the best fine-tuned Gemma-2B model achieved an f1-macro score of 0.6094. This configuration used LoRA rank and alpha values of 128. A configuration with a lower rank and alpha (32) yielded a slightly lower score (0.5814). These results are aligned with the trends observed during validation.

For Task 2, the test results highlight the significance of audio encoder choice in multimodal models [34]. Specifically, the Whisper-Gemma model achieved a strong f1-macro score of 0.7904, demonstrating consistent performance across different emotion classes. In contrast, the Emo2vec-Gemma model scored a lower f1-macro of 0.5134.

Furthermore, we tested for Task 2 the Qwen-Audio model which achieved a higher f1-macro score of 0.8248 on the official test split with a LoRA rank and alpha values of 256 and 2048, respectively. Notice that the value is higher than the one corresponding to the best configuration obtained in our in-house experimentation, which is LoRa rank of 64 and alpha of 1024, which achieved a 0.8072 f1-macro score. However, the reader may note that the difference between both configurations was not statistically significant (see Figure 5). We hypothesize that it may be due to the use of higher rank decomposition matrices results in a better adaptation of the target distribution to be modeled.

6. Conclusions and Future Work

This work explored the potential of MM-LLMs to capture the intrinsic patterns of emotions in spoken language, specifically in the context of the EmoSpeech IberLEF 2024 challenge.

For task 1, we explored parameter-efficient fine-tuning of LLMs using audio transcripts, working under the hypothesis that the vast generic knowledge held by this kind of architecture can be leveraged in order to build a strong emotion prediction with little adjustments. We obtained a top result of **0.6094** with a Gemma-2B model fine-tuned using the LoRA technique with a rank and alpha parameters of 128 (0.6694 according to our in-house cross-validation-based experimentation setup). This promising result opens the door to further exploration of text-only models for automatic SER by efficiently tuning large models.

For task 2, we fine-tuned the Qwen-Audio model to efficiently adapt the LLM using LoRA to improve the emotional recognition capabilities of the model. It achieved a **0.8248** f1-macro in the challenge test dataset, being the second best result, showing that the audio features extracted from the pre-trained audio encoder contain relevant emotional information that can be adapted so that the LLM can accurately process it.

Interestingly, a comparable performance with our Qwen-Audio best result was achieved using the Whisper-Gemma model, which only fine-tuned a projection layer while keeping both models frozen. This results suggests that smaller models (8.4 billion parameters for Qwen-Audio vs. 3.2 billion parameters for Whisper-Gemma) can be effective with appropriate audio encoders (Whisper achieved **0.7904** f1-macro compared to **0.5134** with Emo2Vec).

These findings suggest several promising directions for future work. First, we plan to investigate the effect of fine-tuning the Qwen-Audio audio encoder specifically for the target language, potentially leading to further performance improvements. Additionally, this could be extended to using models fine-tuned for specific tasks, followed by the joint fine-tuning of the projection layer to create even more capable SER systems.

Overall, this work demonstrates the effectiveness of MM-LLMs for emotion recognition in spoken language, contributing valuable insights to the EmoSpeech Iberlef 2024 challenge. Our findings pave the way for further research into efficient and accurate models for this task.

7. Acknowledgments

Sergio Esteban-Romero research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). Iván Martín-Fernández's research was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". We also thank the equipment provided by the INFRARED program of the JCyL under grant IR2020-1-ULE01.

References

[1] Y. Ulgen Sonmez, A. Varol, In-depth investigation of speech emotion recognition studies from past to present – The importance of emotion recognition from speech signal for AI, Intelligent Systems with Applications 22 (2024) 200351. URL: https://www.sciencedirect.com/science/article/pii/S2667305324000279. doi:https://doi.org/10.1016/j.iswa.2024.200351.

[2] 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.

[3] 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.

[4] R. Pan, J. A. García-Díaz, M. A. 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).

[5] Y. U. Sonmez, A. Varol, In-Depth Analysis of Speech Production, Auditory System, Emotion Theories and Emotion Recognition, in: 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1–8. doi:10.1109/ISDFS49300.2020.9116231.

[6] OpenAI, GPT-4 Technical Report, 2024. arXiv:2303.08774.

[7] Anthropic AI, The claude 3 model family: Opus, sonnet, haiku, Claude-3 Model Card (2024).

[8] Gemma Team, Google Deepmind, Gemma: Open Models Based on Gemini Research and Technology, 2024. arXiv:2403.08295.

[9] Qwen Team, Qwen Technical Report, 2023. arXiv:2309.16609.

[10] Y. Chu, J. Xu, X. Zhou, et al., Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models, 2023. arXiv:2311.07919.

[11] E. J. Hu, Y. Shen, P. Wallis, et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021. arXiv:2106.09685.
[12] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Transactions on Speech and Audio Processing 8 (2000) 429–442. doi:10.1109/89.848224.
[13] B. Schuller, R. Müller, M. Lang, G. Rigoll, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles (2005).
[14] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech communication 53 (2011) 768–785.
[15] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, K. R. Anne, A comparative analysis of classifiers in emotion recognition through acoustic features, International Journal of Speech Technology 17 (2014) 401–408.
[16] C. Yogesh, M. Hariharan, R. Yuvaraj, et al., Bispectral features and mean shift clustering for stress and emotion recognition from natural speech, Computers & Electrical Engineering 62 (2017) 676–691.
[17] Mustaqeem, M. Sajjad, S. Kwon, Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, IEEE Access 8 (2020) 79861–79875. doi:10.1109/ACCESS.2020.2990405.
[18] R. Jahangir, Y. W. Teh, G. Mujtaba, et al., Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion, Machine Vision and Applications 33 (2022) 41. URL: https://doi.org/10.1007/s00138-022-01294-x. doi:10.1007/s00138-022-01294-x.
[19] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control 47 (2019) 312–323. URL: https://www.sciencedirect.com/science/article/pii/S1746809418302337. doi:https://doi.org/10.1016/j.bspc.2018.08.035.
[20] I. Zubiaga Amar, R. Justo Blanco, M. De Velasco Vázquez, M. I. Torres Barañano, Speech emotion recognition in Spanish TV Debates, in: roc. IberSPEECH 2022, ISCA, 2022, pp. 186–190.
[21] Gemini Team, Google Deepmind, Gemini: A Family of Highly Capable Multimodal Models, 2024. arXiv:2312.11805.
[22] H. Touvron, T. Lavril, G. Izacard, et al., LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971.
[23] J. Su, Y. Lu, S. Pan, et al., RoFormer: Enhanced Transformer with Rotary Position Embedding, 2023. arXiv:2104.09864.
[24] S. Gunasekar, Y. Zhang, J. Aneja, et al., Textbooks Are All You Need, 2023. arXiv:2306.11644.
[25] R. Zhang, J. Han, C. Liu, et al., LLAMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention, 2023. arXiv:2303.16199.
[26] A. Radford, J. W. Kim, T. Xu, et al., Robust Speech Recognition via Large-Scale Weak Supervision, 2022. arXiv:2212.04356.
[27] Y. Gong, S. Khurana, L. Karlinsky, J. Glass, Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers, in: INTERSPEECH 2023, interspeech_2023, ISCA, 2023. URL: http://dx.doi.org/10.21437/Interspeech.2023-2193. doi:10.21437/interspeech.2023-2193.
[28] D. Zhang, S. Li, X. Zhang, et al., SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities, 2023. arXiv:2305.11000.
[29] Z. Ma, Z. Zheng, J. Ye, et al., emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation, 2023. arXiv:2312.15185.
[30] A. Baevski, W.-N. Hsu, Q. Xu, et al., data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, 2022. arXiv:2202.03555.
[31] O. Elharrouss, Y. Akbari, N. Almaadeed, S. Al-Maadeed, Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches, 2022. arXiv:2206.08016.

[32] The ModelScope Team, SWIFT: Scalable lightWeight Infrastructure for Fine-Tuning, https://github.com/modelscope/swift, 2024.
[33] C. Tang, W. Yu, G. Sun, et al., SALMONN: Towards Generic Hearing Abilities for Large Language Models, 2024. arXiv:2310.13289.
[34] B. McKinzie, Z. Gan, J.-P. Fauconnier, et al., MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training, 2024. arXiv:2403.09611.