Speech emotion recognition from voice messages recorded in the wild

Lucía Gómez-Zaragozá, Óscar Valls, Rocío del Amor, María José Castro-Bleda,
Valery Naranjo, Mariano Alcañiz Raya, Javier Marín-Morales

***Abstract*—Emotion datasets used for Speech Emotion Recognition (SER) often contain acted or elicited speech, limiting their applicability in real-world scenarios. In this work, we used the Emotional Voice Messages (EMOVOME) database, including spontaneous voice messages from conversations of 100 Spanish speakers on a messaging app, labeled in continuous and discrete emotions by expert and non-expert annotators. We created speaker-independent SER models using the eGeMAPS features, transformer-based models and their combination. We compared the results with reference databases and analyzed the influence of annotators and gender fairness. The pre-trained Unispeech-L model and its combination with eGeMAPS achieved the highest results, with 61.64% and 55.57% Unweighted Accuracy (UA) for 3-class valence and arousal prediction respectively, a 10% improvement over baseline models. For the emotion categories, 42.58% UA was obtained. EMOVOME performed lower than the acted RAVDESS database. The elicited IEMOCAP database also outperformed EMOVOME in the prediction of emotion categories, while similar results were obtained in valence and arousal. Additionally, EMOVOME outcomes varied with annotator labels, showing superior results and better fairness when combining expert and non-expert annotations. This study significantly contributes to the evaluation of SER models in real-life situations, advancing in the development of applications for analyzing spontaneous voice messages.**

***Index Terms*—Speech emotion recognition, natural database, valence, arousal, pre-trained model, speaker independent.**

I. Introduction

HUMAN communication allows individuals to express not only their ideas, but also their emotional state. In everyday conversations, people utilize both information to adjust their behavior, underscoring the importance of recognizing emotions. Speech Emotion Recognition (SER) is a research field that automatically identifies a person's emotional state from their voice. SER has potential applications for the study of human-to-human communications, such as detecting stress or depression in medical contexts. It is also relevant in human-computer interactions, for example, improving the naturalness of speech synthesis systems.

A key aspect of SER research is emotional databases containing labeled samples from which to learn patterns correlating with emotions. In the literature, emotions are modeled using continuous or discrete emotion models [1]–[3]. The discrete model is based on categories of basic emotions considered innate and universal, which can be combined to obtain other emotions. For example, Ekman [4] proposed the "big six" basic emotions: fear, surprise, happiness, sadness, anger and disgust. The dimensional model explains emotions using continuous dimensions that represent fundamental properties shared by all emotions. Various dimensional models exist, depending on the number and type of dimensions considered. For instance, Russell's circumplex model of affect [5] uses two dimensions —valence and arousal— on a Cartesian axis system. Valence represents the pleasantness (positive valence) or unpleasantness (negative valence) of emotion, while arousal measures its intensity, ranging from active/exciting (high arousal) to passive/dull (low arousal). Each emotion can be viewed as a combination of these two dimensions, illustrated in Fig. 1. Both emotional models have strengths and weaknesses, as described in [2]. SER literature predominantly focuses on the discrete model due to its intuitive labeling since individuals commonly use basic categories to express emotions in everyday life. However, it struggles to accurately capture some complex affective states. In contrast, the dimensional model is less intuitive, but it provides greater flexibility in categorizing a broader spectrum of emotions. It acknowledges the complexity of emotional experiences, where individuals may experience a mix of emotions rather than fitting into discrete categories, and captures the dynamic by showing how emotions relate to each other on a continuum.

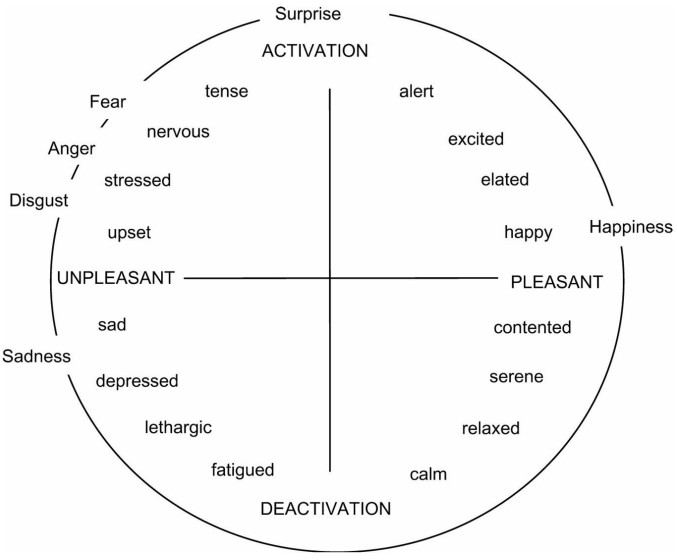

Fig. 1. Russell's circumplex model of affect. The outer circle illustrates where prototypical emotions are normally located. From [6].

Different databases exist depending on whether the samples are acted, elicited/induced and natural/spontaneous speech as described in several reviews [1]–[3]. Acted databases comprise speech samples of actors simulating emotions, typically while delivering predetermined sentences. Despite their prevalence in literature, acted emotions tend to be stereotypical and exaggerated, limiting their applicability in real-world scenarios. Elicited databases involve speech samples collected in artificially induced situations to elicit specific emotional states, for instance, listening to a story, watching a video or conducting a guided discussion. While these samples offer a closer approximation to genuine emotional expression, the induction process has limitations and ethical implications. Individuals may react differently to the same stimulus, requiring an extra subjective assessment to identify the emotion in the speech sample. Natural databases include speech samples extracted from diverse sources, including films, radio/TV talk-shows and occasionally therapy interviews. These databases encounter several challenges. They typically contain audios with overlapping voices and background noise, known as in-the-wild conditions. Furthermore, individuals who are aware of being recorded might unintentionally control their emotions or express them unnaturally. Additionally, determining the emotion in each sample necessitates an external subjective evaluation. Natural databases, though infrequent in literature and often private for ethical and legal reasons, are crucial for recognizing genuine, real-life emotions.

To create speech emotion recognition models, different approaches have been studied in the literature. Traditionally, SER systems have leaned on so-called hand-crafted features, i.e., acoustic properties such as prosodic or spectral features extracted at frame level or aggregated using high-level statistics functions [3]. With the advancements in deep learning, it became possible to directly use the raw signal to train the recognition models, capturing the temporal dynamics and avoiding feature engineering. Techniques such as convolutional neural networks and recurrent neural networks emerged as the primary approach in SER [2]. Nevertheless, they have been limited by the small size of the emotional datasets, which is a major challenge for deep learning methods to achieve optimal effectiveness [7]. In recent years, large pre-trained models have emerged as a powerful framework that is gaining significant attention in all speech-related domains, including automatic speech recognition (ASR), speaker verification and SER, among others [3], [8]. These transformer-based models (e.g., Wav2vec 2.0 [9] and HuBERT [10]) are trained using self-supervised learning approaches on large unlabeled datasets through pretext tasks. Then, they are used to extract speech representations, even from long sequences at once, overcoming the limitations of previous deep learning methods. The learned speech representations are fed into simple neural networks in downstream tasks, including SER, in which recent works have obtained promising results [7], [8], [11], [12].

SER is still an open-ended problem due to its complexity. Emotions are subjective internal states, which makes their theoretical conceptualization difficult [13]. Consequently, annotating audios is a challenging task, as emotions can be perceived differently between raters [14], as well as several emotions can also be combined in the same utterance. This is also reflected in the available emotional speech databases in the literature, often constrained by small sizes and limitations in recognizing real-life emotions. The availability of databases is even more limited for languages other than English, such as Spanish. A total of twelve Spanish databases are mentioned in the literature, consisting of six private [15]–[20] and six open for research, among which two are acted [21], [22], one elicited [23] and three natural [24]–[26]. Furthermore, the performance of the SER task is significantly influenced by the data separation strategy used. Speaker-dependent (SD) models are trained and tested using speech samples from the same speakers, reaching accuracies of over 70% regardless of the databases or emotion classes involved [3]. In contrast, speaker-independent (SI) models use different subjects for training and testing, so they are evaluated with unknown speakers. The majority of SI models yield accuracies ranging from 29% to 65%, indicating the inherent difficulty of SER in these cases [3]. Finally, another challenge for SER models is fairness, that is, the difference in model results with respect to individuals or population groups according to variables such as gender, age or ethnicity [27]. Given the increasing concerns in society about biases in ML systems [28], it is necessary to ensure fairness in SER models.

To fill the gap of existing literature on models trained in real-world settings, this study focuses on speech emotion recognition using a natural speech database we meticulously collected, called the Emotional Voice Messages (EMOVOME) database [29]. To the best of our knowledge, this is the first Spanish database of spontaneous emotions in real voice messages, and it is labeled by expert and non-expert annotators. We created speaker-independent SER models using acoustic features and state-of-the-art pre-trained models. We conducted a comprehensive analysis to assess how various EMOVOME properties influence the results of the SER models, comparing the results with the widely used Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [30] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [31]. We explored the following Research Questions (RQ):

- RQ1: *What is the performance of SER models in EMOVOME when utilizing classical acoustic features versus state-of-the-art speech embeddings?*
- RQ2: *To what extent does the performance of EMOVOME, a natural database, compare with elicited and acted reference databases in the literature, such as IEMOCAP and RAVDESS?*
- RQ3: *What impact do annotator labels exert on the performance of SER models when dealing with challenging natural databases like EMOVOME?*
- RQ4: *How does gender impact the fairness of SER model outcomes in EMOVOME compared to reference databases?*

The rest of the paper is organized as follows. Section 2 provides a review of related works on SER. Section 3 describes the databases used in this study. In Section 4, the methodology applied to create the SER models is detailed, and Section

5 reports the outcomes of these models. Section 6 discusses the results of the stated hypothesis. The paper concludes by summarizing the work and proposing future directions for SER with realistic data.

II. Related work

There is a scarcity of research focused on SER using Spanish databases. Table I summarizes the related literature. For each of the twelve Spanish databases found, it indicates the type of speech (acted, elicited or natural), the number of samples (N), the number of speakers (Spk) with the gender distribution (males/females), the labeled emotions and the database access type (private, commercially available or free). It also includes the Unweighted Accuracy (UA) and/or the Weighted Accuracy (WA) of the monolingual SER models created using these databases (results of multilingual and cross-lingual models are not reported here since they represent more challenging tasks and monolingual outperforms the others for comparable number of training samples [32]). Four databases contain mixed languages: the Emotional Speech Synthesis database [22] is part of the INTERFACE project [33], in which data in English, Slovenian, French and Spanish was collected; RekEmozio [20] includes data in both Spanish and Basque; CMU-MOSEAS [25] includes YouTube videos in Spanish, Portuguese, German and French; and EmoFilm [26] includes clips from both original English films and their Spanish and Italian dubbed versions. LIRIS-ACCEDE [34] is also a multilingual database, but it is mostly in English, with only a small subset available in nine other languages (including Spanish). Due to a lack of specific details on these partitions, they have been excluded from the table. Additionally, EmoWisconsin [35] and IESC-child [36] are emotional databases in Spanish; however, they contain children's speech. As our focus here is on adult speech, they are not included in the table.

Most of the Spanish emotional databases contain speech acted by professional or non-professional actors [15]–[22], one is based on elicited responses [23] and three are natural databases [24]–[26]. Six of the databases are private [15]–[20], one is commercially available [22] and five are free [21], [23]–[26]. This paper focuses on the natural databases and, although the three existing ones are public, they have some limitations. MOUD database [24] comprises 105 video clips manually extracted from YouTube content in Spanish, covering various topics, and with an average duration of 30s each. Two annotators labeled these clips into three categories: positive, neutral, or negative. Notably, the neutral class is represented by only 4 samples, in contrast to 47 samples for the positive class and 54 for the negative class. Additionally, a large proportion of the speakers in the database are female. As for CMU-MOSEAS [25], this database comprises 10000 Spanish sentences extracted from monologue videos on YouTube, with an average sentence duration of 6.7s. Each sentence was annotated by 3 annotators using different labels, including sentiment and the six Ekman's emotions. The article asserts attempts to achieve gender balance in videos across languages and regions but lacks specific details on the final distribution. Furthermore, despite the public availability of the database, the original clips are not disclosed, and only high-level features are made accessible. Both the MOUD and CMU-MOSEAS databases share two common constraints. Firstly, their speakers record videos for YouTube, and awareness of being recorded might lead to unintentional emotional control or unnatural expression in response to the artificial situation [43]. Secondly, the nature of YouTube videos in both databases involves monologues, potentially lacking the naturalness found in conversational contexts, as highlighted in [44]. Finally, EmoFilm database [26] includes clips from original English films and their Spanish and Italian dubbed versions. The definition of natural databases includes speech extracted from films. However, films inherently involve acting, and as such, the emotions depicted are portrayed by actors. This holds true for the original films and their dubbed versions, potentially rendering the emotional expressions even more overacted in the latter. Therefore, this database may not be suitable for recognizing real emotions, as noted by their authors [26].

The performance of SER models in Spanish databases is conditioned by the type of data used. For acted databases, the accuracy is equal or greater than 90% (e.g. [18], [19], [21], [38]–[41]). An exception is the work in [37], in which they used the RekEmozio database to classify seven emotion categories and achieved 74.82% accuracy. However, they used a speaker-independent approach, which is known to yield comparatively lower results [3]. This trend is also reflected in [23], where they used the elicited database EmoSpanishDB to classify seven emotion categories and obtained 56.3% and 42.1% UA for speaker-dependent and speaker-independent, respectively. The difference between both approaches decreased when they used EmoMatchSpanishDB, a refined version of the database obtained by eliminating labeled samples that did not match the original elicited emotion, achieving an accuracy of 65.0% and 64.2% UA for SD and SI, respectively. Finally, natural databases have the lowest scores. Using only the audio modality, in [24] they achieved an accuracy of 46.75% in differentiating between positive and negative samples. In the study by [17], an accuracy of 55% was achieved for the classification of five emotions using a model trained on acted and tested on natural data. Higher scores were achieved with the EmoFilm database, which may be due to the inherent acted nature of the samples. In [32], they obtained 69.15% and 70.76% UA and WA, respectively, using a speaker-independent approach. In [42], the score increased to 85.0% WA, and 91.8% and 97.7% for female and male versions of the models.

Regarding the techniques used to create the SER models, a noteworthy paradigm shift involves leveraging large pre-trained models in emotion recognition tasks. The study in [32] used the pre-trained wav2vec2-large-robust model fine-tuned on MSP-Podcasts dataset from [11] to extract speech embeddings. These embeddings were then fed into a support vector machine to create the models, obtaining 69.15% unweighted accuracy for five emotion categories on the EmoFilm database. To our knowledge, this is the only previous study using pre-trained models for monolingual SER models in Spanish. Nevertheless, an increasing number of recent publications have focused on the use of pre-trained models with English databases (predominantly relying on acted datasets)

for categories of emotions and less frequently for dimensions.

As for studies on emotion category prediction, a very recent study [11] summarized the state-of-the-art results for 4-class emotion classification in the widely used IEMOCAP database, including studies using the pre-trained wav2vec 2.0 and HuBERT models. Comparing the cross-validation results, the unweighted accuracy values range from 60.0% to 74.3%, while the weighted average recall values vary from 62.6% to 79.6%, both top models. Overall, HuBERT surpassed the performance of the wav2vec 2.0 models. Other works have explored several pre-trained models and emotional databases. In [45], they used eight pre-trained models (including wav2vec 2.0) and four databases (CREMA-D, TESS, SAVEE, Emo-DB) to predict 6-7 emotions. The results showed that pre-trained models for speaker recognition (x-vector and ECAPA) achieved the highest scores, possibly due to their learned ability to identify unique features within an individual's speech. Subsequently, UniSpeech-SAT achieved the best results, a model pre-trained using multitask learning, including the speaker identity. In another study [46], the authors compared nineteen pre-trained models (including wav2vec 2.0, HuBERT, UniSpeech-SAT and wavLM) and five datasets (IEMOCAP, MSP-IMPROV, MSP-PODCAST, CMU-MOSEI, JTES) to predict 4-6 emotion categories using a speaker-independent approach. They found that the best scores were achieved with WavLM, UniSpeech-SAT, and HuBERT, all three in the large version.

As for studies on emotion dimension prediction, [7] used a multimodal model (audio + text) as a teacher to fine-tune HuBERT embeddings to predict valence, arousal and dominance on MSP-Podcast database [47]. They obtained state-of-the-art Concordance Correlation Coefficient (CCC) values of 0.757, 0.627 and 0.671 for arousal, valence and dominance, respectively. The previous state-of-the-art performance for valence was 0.377 [48], so 0.627 represents a substantial improvement. They also replicated the results for the IEMOCAP database, achieving again state-of-the-art CCC results for valence (0.667), arousal (0.582) and dominance (0.545). In [11], the authors conducted an exhaustive analysis using different variants of the pre-trained wav2vec 2.0 and HuBERT models for valence, arousal and dominance prediction on the MSP-Podcast database. They obtained the best result in the literature for valence prediction using only audio, with a wav2vec2-large-robust model that achieved a CCC of 0.638. Notably, their results showed that data used for pre-training the models and the fine-tuning of the transformer layers had a strong influence on valence prediction. Both factors played a role in shaping the models' ability to implicitly incorporate linguistic information embedded in the audio signal. This, in turn, accounts for their success in valence prediction, achieving similar performance to multimodal models that integrate explicit textual information.

Finally, research on evaluating model fairness is limited, particularly in the context of pre-trained models [27]. Some previous works have investigated gender-based fairness in Spanish databases, revealing differences in model performance between females and males. In [38], the authors reported a

TABLE I

SPEECH EMOTION RECOGNITION MODELS IN SPANISH. THE ASTERISK (*) INDICATES IF THE DATABASE IS MULTILINGUAL, BUT THE INFORMATION CORRESPONDS TO THE SPANISH PARTITION. THE "BIG SIX" BASIC EMOTIONS ARE: FEAR, SURPRISE, HAPPINESS, SADNESS, ANGER, AND DISGUST.

Database	Type	N	Spk (M/F)	Emotions	Access	UA (%)	WA (%)
Iriondo et al. [15]	Acted	336	8 (4/4)	Big six + desire	Private	-	-
Spanish Emotional Speech [16]	Acted	1288	1 (1/0)	Happy, sad, cold anger, surprise	Private	-	-
Martínez & Cruz [17]	Acted + natural (films)	300 + 80	15 (-/-) + N/A (-/-)	Happy, sad, anger, neutral, fear	Private	-	55 (tested on natural partition) [17]
Emotional Mexican Spanish speech [18]	Acted	240	6 (3/3)	Happy, sad, anger, neutral	Private	$\geq$ 95 [18]	-
Spanish Expressive Voices [19]	Acted	3890	2 (1/1)	Big six + neutral (cold/hot anger)	Private	-	95 [19]
RekEmozio * [20]	Acted	2618	10 (5/5)	Big six + neutral	Private	-	SI: 74.82 [37]
Mexican Emotional Speech Database [21]	Acted	3456	8 (4/4) + 8 (3/5) children	Happy, sad, anger, neutral, fear, disgust	Free	89.49 female [21] 93.90 male [21] 83.30 children [21]	-
Emotional speech synthesis [22] (from INTERFACE * [33])	Acted	5520	2 (1/1)	Happy, sad, anger, neutral, fear, boredom, disgust	Comm. avail.	90.9 female [38] 89.4 male [38]	91.16 [39] 91.51 [40] 94.01 [41]
EmoSpanishDB, EmoMatchSpanishDB [23]	Elicited	3550, 2020	50 (30/20)	Big six + neutral	Free	SD: 56.3, SI: 42.1 [23] SD: 65.0, SI: 64.2 [23]	-
MOUD [24]	Natural (YouTube)	550	105 (21/84)	Positive, neutral, negative	Free	-	46.75 (audio, positive vs negative) [24]
CMU-MOSEAS * [25]	Natural (YouTube)	10000	341 (-/-)	Big six, sentiment, subjectivity, others	Free	-	-
EmoFilm * [26]	Natural (films)	342	57 (33/24)	Happy, sad, angry, fear, contempt	Free	SI: 69.15 [32]	SI: 70.76 [32] 85.0 [42] 91.8 female [42] 97.7 male [42]
Ours (EMOVOME) [29]	Natural (voice messages)	999	100 (50/50)	Valence, arousal Big six + neutral	Free	SI: 49.27% (audio, 3-class valence) [29] SI: 44.71% (audio, 3-class arousal) [29]	-

female accuracy of 90.9% and a male accuracy of 89.4%, yet the database only included one male and one female speaker. Conversely, [21] and [42] showed a different pattern, with males achieving higher accuracy than females. Specifically, [21] presented 89.49% for females and 93.90% for males (with 4 female and 4 male speakers), while [42] obtained 91.8% for females and 97.7% for males using pre-trained models (with 33 male and 24 female speakers). The same trend has been found in some works on SER for English data. In [27], they studied the fairness of SER systems using pre-trained models and observed a reduction of 0.234 in CCC for arousal among females as opposed to males on MSP-Podcast. In [11], they found that pre-trained models tend to exhibit greater fairness in predicting arousal and dominance than in valence. Notably, for valence, the majority of models showed higher CCC for females than for males. Nevertheless, overall the speech representations obtained with pre-trained models seem to be invariant to domain, speaker, and gender. Additionally, the authors in [11] also explored fairness across individual speakers and found that different pre-trained models show overall consensus on categorizing speakers as 'good' or 'bad', obtaining lower CCC values for some individuals in the latter group. These findings underline the importance of incorporating fairness assessment in future research.

III. Speech emotion databases

This research used three databases: EMOVOME, IEMOCAP and RAVDESS, detailed below. A comparison of the main features is presented in Table II, indicating the type of speech (acted, elicited or natural), the language, the labeled emotions categories or dimensions, the number of samples (N), the number of speakers (Spk) with the gender distribution (males/females), the mean number of samples per speaker (N/Spk) and the sample duration (mean and range).

A. Emotional Voice Messages database (EMOVOME)

The Emotional Voice Messages (EMOVOME) database [29] was created from scratch for this work to obtain emotional speeches in real-world conditions. It contains 999 audio messages collected from real WhatsApp conversations of 100 Spanish speakers (50 female, 50 male). Voice messages were produced in-the-wild conditions before participants were recruited, avoiding any conscious bias due to the laboratory environment. Samples were labeled by two clinical psychologists (considered experts in the task of recognizing emotions) and three other annotators (deemed non-experts) in terms of valence and arousal using a 5-point scale. The experts labeled half of the audios each, so henceforth it will be considered a single expert label. The labels were processed to obtain three categories for arousal (high, neutral and low) and valence (positive, neutral and negative). They were finally aggregated by majority voting to obtain a single label per audio, giving priority to the expert's label in case of a tie. The expert also labeled the audios in the following categories: happy, angry, fearful, sad, surprise, disgust and neutral. In this work, only the four most frequent emotions were used for classification, i.e., happy, angry, neutral and surprise. Details of the experimental procedure used for data collection and labeling are described in [29]. To the best of our knowledge, EMOVOME is the first collection of spontaneous emotions from real voice messages.

B. Interactive Emotional Dyadic Motion Capture database (IEMOCAP)

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [30] stands out as a benchmark dataset for studying emotional expression and communication. It contains dyadic interactions of 10 English-speaking actors engaged in scripted and improvised dialogues designed to elicit different emotions. In this work, only the audio files corresponding to the segmentation of the conversations into utterances were used to create the SER models. This set includes 10039 utterances annotated into 9 emotion categories and 3 dimensions.

To align the classification results with the EMOVOME database, valence and arousal scores were stratified into three categories. Samples with a score below 2.5 were considered negative/low valence/arousal. Those samples with scores between 2.5 and 3.5 were considered neutral valence/arousal. Finally, we considered samples as positive/high valence/arousal if the score was greater than 3.5. Moreover, to facilitate comparison with the literature, only the four most frequent emotions were used for the classification into categories, i.e., angry, neutral, sad, happy and excited (the latter two were merged, following previous studies).

C. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [31] is a multimodal database that contains speech and song recordings expressing different emotional states. In this investigation, the voice-only data from RAVDESS was used. In this set, 24 actors pronounce two phrases ("Kids are talking by the door" and "Dogs are sitting by the door") with different emotional intentions in a recording studio. Actors vocalized each statement in normal and strong emotional intensity for each of the eight emotions (except for neutral), resulting in 1440 samples.

TABLE II
Comparison of the emotion databases used in this work: EMOVOME, RAVDESS and IEMOCAP.

Database	Type	Language	Dimensions	Categories	N	Spk (M/F)	N/Spk	Sample duration	
								Mean	Range
EMOVOME	Natural	Spanish	Valence, arousal	Big six + neutral	999	100 (50/50)	10	17.59s	1-60s
IEMOCAP	Elicited, acted	English	Valence, arousal, dominance	Big six + neutral, frustration, excited	10039	10 (5/5)	1004	4.59s	1-35s
RAVDESS	Acted	English	-	Big six + neutral, calm	1440	24 (12/12)	60	3.70s	3-5s

To compare the classification results with the EMOVOME database, the emotion labels were transformed from the discrete emotional model to the dimensional emotional model, as also studied in [49]. For this purpose, the eight emotion categories were converted into valence and arousal dimensions following their distribution in Russell's circumplex model of affect [5] (see Fig. 1). The label conversion in the valence dimension was the following: negative (sad, angry, fearful, disgust), neutral (neutral, surprised) and positive (happy, calm). Likewise, the label conversion for the arousal dimension was: low (calm, sad), neutral (neutral, disgust) and high (happy, angry, fearful, surprised).

IV. Methods

Three approaches were implemented to create the SER models. First, a standard feature set with classical machine learning algorithms was used as a baseline (Section IV-A). Then, pre-trained models were used as feature extractors, followed by a linear layer for classification (Section IV-B). Finally, a combination of the pre-trained models and the standard features was analyzed (Section IV-C).

To evaluate the three methods, the speech databases were divided into 80% for development and 20% for testing using a speaker-independent approach. Details of the label distribution in each partition are provided in Table III. To facilitate reproducibility and comparison of results, for test, we used: the test set in [29] for EMOVOME; session 5 for IEMOCAP; and "fold 0" proposed in [50] for RAVDESS. Additionally, parameter tuning also employed a SI cross-validation scheme, utilizing the StratifiedGroupKFold from Scikit-learn [51], with 4 folds for IEMOCAP and 5 folds for EMOVOME and RAVDESS. Folds are formed by grouping all audio samples from the same subject into a single fold, ensuring a roughly equal distribution of labels within each fold. All methods were assessed using two evaluation metrics. Accuracy, also called weighted accuracy (WA), evaluates the ratio of correctly predicted class samples to the total samples. WA is suitable for balanced datasets but less so for imbalanced ones, as it gives more weight to classes with more samples. In SER literature, unweighted accuracy (UA) is often preferred, representing the average of individual class accuracies. The Python libraries LibROSA, Scikit-learn, TensorFlow and Keras were used to implement the models.

A. Baseline: eGeMAPS and machine learning

As a first approach, we repeated the baseline method proposed in [29] using long-term acoustic features and machine learning algorithms. First, audio samples in EMOVOME were resampled to 44.1 kHz, while for RAVDESS and IEMOCAP, the sample rate was kept at 48 kHz and 16 kHz, respectively. Next, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [52] was extracted using the openSMILE toolkit [53]. The 88 features per audio were normalized by subtracting the mean and dividing by the standard deviation of the development samples. For feature selection, high-correlated features ($p > 0.95$) were first eliminated using Pearson's correlation matrix, and a filter method was then used to select 25%, 50% or 75% of the features based on the highest ANOVA F-values. Finally, Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) models were fitted on the development set according to the cross-validation scheme indicated above, tuning the combination of hyperparameters detailed in [29]. The chosen combination of features and hyperparameters was used to train a model on the entire development set, followed by an evaluation of the test set.

B. Pre-trained model embeddings

We evaluated several pre-trained models with different architectures, pre-training methodologies and pre-training audio data, indicated in Table IV. First, several variations of the widely used Wav2vec 2.0 model were selected. Considering that EMOVOME is in Spanish, and we wanted to compare it with other two databases in English, we selected models pre-trained using a multilingual approach: facebook/wav2vec2-xls-r-300m (w2v2-xlsr-128) [54] pre-trained on 436k hours of audios in 128 languages, and facebook/wav2vec2-large-xlsr-53 (w2v2-xlsr-53) [55], pre-trained on 56k hours of audios in 53 languages. Both models included Spanish in the pre-training data. Despite prior research suggesting that fine-tuning models for automatic speech recognition do not help with speech emotion recognition [12], [56], we decided to include a fine-tuned model for Spanish ASR. This decision aimed to explore whether this approach outperforms the utilization of a pre-trained model in a different language since [56] and [12] used pre-trained models in the same language as the database tested (English). Therefore, we also used the fine-tuned ver-

TABLE III
Data distribution in training and test partitions of EMOVOME, IEMOCAP and RAVDESS.

Database	Spk (M/F)	Valence			Arousal			Categories							
		Negative	Neutral	Positive	Low	Neutral	High	hap	ang	neu	sur	sad	fea	dis	cal
Training															
EMOVOME - E	80 (40/40)	251	247	309	265	347	195	264	157	184	105	-	-	-	-
EMOVOME - N	80 (40/40)	219	363	219	76	244	411	-	-	-	-	-	-	-	-
EMOVOME - C	80 (40/40)	241	305	261	148	328	331	-	-	-	-	-	-	-	-
IEMOCAP	8 (4/4)	2622	3725	1522	1023	5405	1441	1194	933	1324	-	839	-	-	-
RAVDESS	19 (10/9)	608	228	304	304	228	608	152	152	76	152	152	152	152	152
Test															
EMOVOME - E	20 (10/10)	59	46	87	34	76	82	78	42	41	13	-	-	-	-
EMOVOME - N	20 (10/10)	56	70	66	13	51	128	-	-	-	-	-	-	-	-
EMOVOME - C	20 (10/10)	57	62	73	13	70	109	-	-	-	-	-	-	-	-
IEMOCAP	2 (1/1)	799	883	488	258	1503	409	442	170	384	-	245	-	-	-
RAVDESS	5 (2/3)	160	60	80	80	60	160	40	40	20	40	40	40	40	40

sion facebook/wav2vec2-large-xlsr-53-spanish (w2v2-xlsr-53-spa) [55] to obtain the embeddings. We also included a model pre-trained using noisy audios, the model facebook/wav2vec2-large-robust (w2v2-L-robust) [57], as it may be useful for the EMOVOME database despite being in English. As a widely-used alternative to Wav2Vec2, we used a HuBERT model, particularly the large version facebook/hubert-large-ll60k (hubert-L) [10]. Additionally, recent studies [45], [58], [59] have indicated that including information about the speaker is helpful for speech emotion recognition. Therefore, following these investigations, two more models were selected: Microsoft's UniSpeech-SAT-Large (unispeech-L) [60] and a Statistics Pooling Time Delay Neural Network to obtain x-vector embeddings [61]. The former is a model pre-trained using multitask learning, including also the speaker identity during training. The latter provides a speaker embedding learning during a speaker verification task. All the pre-trained models are available in HuggingFace. The x-vector also requires the Speechbrain toolkit [62]. All of them require the input audio to be resampled to 16 kHz.

TABLE IV
PRE-TRAINED MODELS USED IN THIS STUDY.

Model name	HuggingFace name
w2v2-xlsr-128	facebook/wav2vec2-xls-r-300m
w2v2-xlsr-53	facebook/wav2vec2-large-xlsr-53
w2v2-xlsr-53-spa	facebook/wav2vec2-large-xlsr-53-spanish
w2v2-L-robust	facebook/wav2vec2-large-robust
hubert-L	facebook/hubert-large-ll60k
unispeech-L	microsoft/unispeech-sat-large
x-vector	speechbrain/spkrec-xvect-voxceleb

Previous investigations have adapted the architecture of the pre-trained models by adding a classification head and fully or partially fine-tuning the model. Nevertheless, this process requires high computational resources due to the large size of some models and the audio lengths (particularly in EMOVOME). For this reason, the embedding extraction process was implemented offline, saving the audio embeddings in independent files. For all pre-trained models except for the x-vectors, we extracted the last hidden state of the last transformer layer. As a result, we obtained a vector of dimensions $(X, 1024)$, where X varies depending on the audio length. Following previous research [11], [56], we applied the average over the time dimension to obtain a 1024-dimension vector per audio sample. For x-vectors, the pre-trained model has a built-in statistics pooling layer that computes the mean and standard deviation of information from the last frame-level layer. These statistics are combined and input into a 512-dimensional hidden layer that is finally used to obtain the embeddings. As a result, the pre-trained model consistently outputs a 512-dimensional vector, independent of audio length, eliminating the need for extra aggregation strategies. Subsequently, we trained a neural network comprised solely of a linear layer, which took the embeddings as input and produced the output corresponding to the number of labels. This architecture, proven effective in previous literature [56], simplifies the model while still capturing essential features in the merged embeddings. The model was trained during a maximum of 3000 epochs during cross-validation using Adam optimization, a learning rate of 0.001 and a batch size of 128. The early stopping callback was applied, set to monitor the validation loss with a patience of 50. For the selected hyperparameters, a final model was trained on the development set for a fixed number of epochs selected based on the cross-validation results and evaluated on the test.

C. Pre-trained model embeddings and eGeMAPS

We integrated previous methods by combining speech embeddings with the eGeMAPS feature set. The 1024-dimensional embeddings (except for the x-vector, which is 512-dimensional) were concatenated with the 88 eGeMAPS features. The combined set underwent normalization before being fed into the neural network, which consists of a linear layer, following the procedure detailed in the previous section.

V. RESULTS

The cross-validation results, depicted in Fig. 2, provide a comprehensive overview of the performance across the different databases and emotion labels. In this visualization, the unweighted accuracy of the optimal baseline model using acoustic features and machine learning (called eGeMAPS) is indicated in gray. Adjacent to this baseline, the results for all models employing pre-trained models (named Embeddings) are displayed, along with a third column with the method integrating the speech embeddings extracted from pre-trained models and the eGeMAPS features together (called Emb+eGeMAPS). The figure indicates the mean and standard deviation of the UA across the five folds (four folds in the case of IEMOCAP).

Regarding the baseline eGeMAPS models, the lowest UA values are from the EMOVOME database, with 41-44% for valence, 41-51% for arousal and 31.42% for four categories. IEMOCAP achieves 50.20%, 53.21% and 59.50% for valence, arousal and categories, respectively. RAVDESS achieves the highest UA results for valence and arousal, with 57.19% and 62.48%, respectively. For the eight categories, it obtains 45.52%. Overall, SVC was preferred to KNN in most cases.

The embedding approach demonstrates a notable enhancement in the models' performance compared to the baseline method. Again, EMOVOME yields the lowest scores, with UA values in the range 54-60% and 49-59% for valence and arousal, respectively, and 43.30% for the 4-way classification. IEMOCAP models have UA results of 61.73%, 57.68% and 68.79% for valence, arousal and categories, respectively. RAVDESS again reaches top UA values, with 71.91%, 74.09% and 67.81% for valence, arousal and categories, respectively. In general, unispeech-L consistently demonstrated superior performance compared to alternative options. Using pre-trained embeddings results in an enhancement of approximately 10% in the UA across all combinations.

Finally, the Emb+eGeMAPS method obtains similar results to the previous approach. EMOVOME achieves similar results in both valence and arousal (53-59% and 52-60% respectively), while the UA in 4-emotions classification decreases (40.22%). A possible reason is that EMOVOME was recorded

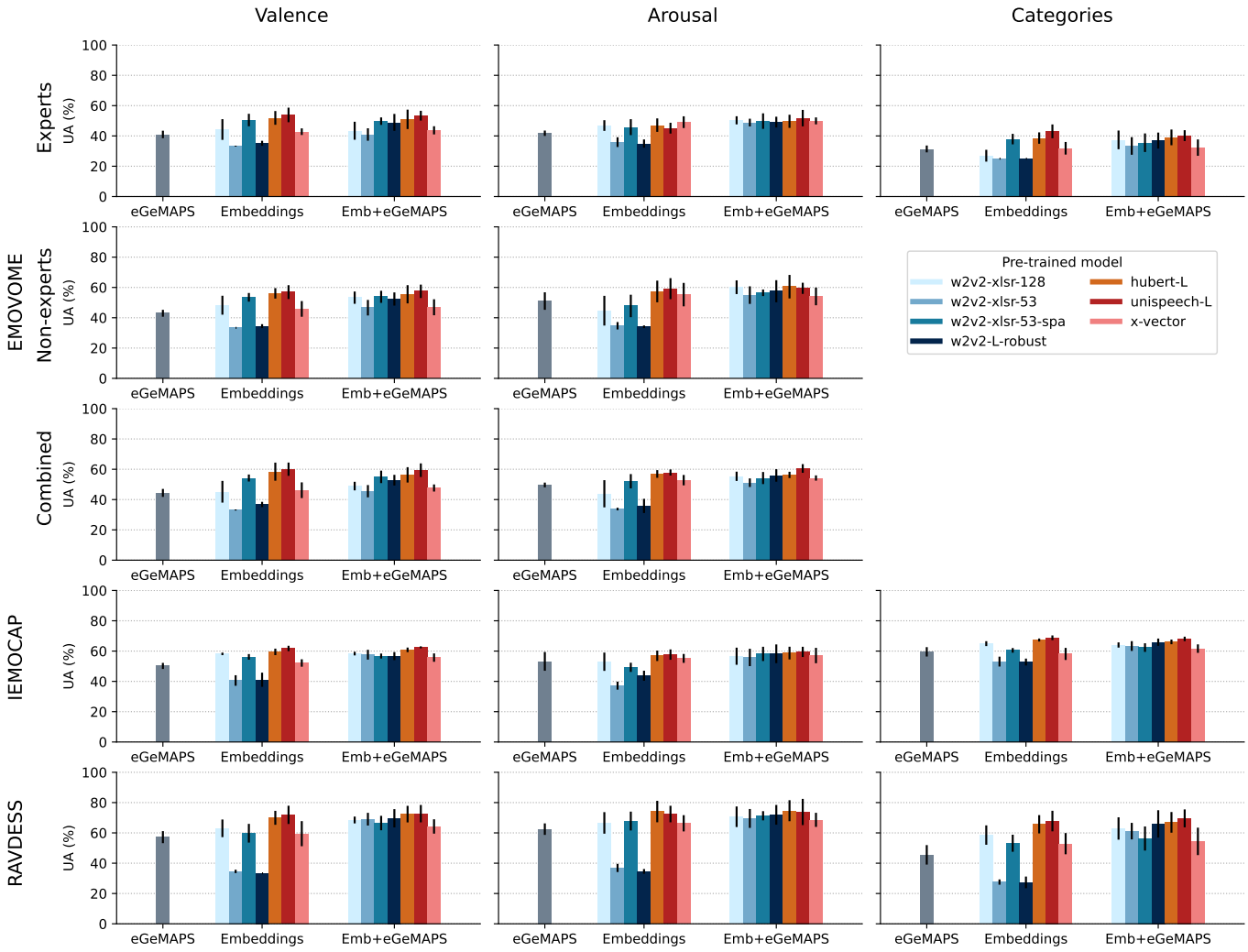

Fig. 2. Cross-validation results for the three methods implemented (eGeMAPS, Embeddings and Emb+eGeMAPS) across the different databases (EMOVOME, IEMOCAP and RAVDESS) and emotion labels (valence, arousal and categories of emotions).

in-the-wild conditions. Therefore, eGeMAPS features may be affected by microphone quality and background noise [52]. Conversely, IEMOCAP and RAVDESS were recorded in a controlled environment, and the UA values increased around 1-3% compared to the previous approach (except for the emotion categories with IEMOCAP, which slightly decreased). IEMOCAP achieves 62.48%, 60.41% and 68.04% for valence, arousal and categories, respectively. RAVDESS obtains 72.70%, 75.27% and 69.58% for valence, arousal and categories, respectively. Again, unispeech-L is the best option in the majority of cases, but now the other pre-trained models, which exhibited lower results in the previous approach, demonstrated significant improvement when combined with the eGeMAPS features, particularly in the cases of w2v2-xlsr-53 and w2v2-L-robust.

Using the hyperparameters of the models obtaining the highest UA in cross-validation, we trained new models using the development set and evaluated them on the test set. The results are shown in Table V and VI for valence and arousal prediction, respectively. The tables include the model achieving the highest cross-validation results among the three implemented methods: eGeMAPS, Embeddings and Emb+eGeMAPS. The test results follow the trends found in cross-validation, with RAVDESS outperforming IEMOCAP and EMOVOME across the two dimensions (73.54% in valence and 71.94% in arousal). IEMOCAP UA scores are higher than EMOVOME scores in arousal prediction (61.20% vs. 43.57-58.73% respectively), whereas, for valence prediction, IEMOCAP and EMOVOME achieve similar results (60.40% for the former and 57.53-61.64% for the latter). In the case of EMOVOME, we also explored the difference between expert (E) and non-expert (N) annotations, as well as their combination (C). Table V and VI show a comparison of results among them in the EMOVOME database, and their confusion matrix is represented in Fig. 3.

An important limitation in IEMOCAP and RAVDESS is the label transformation applied to obtain the valence and arousal categories for their comparison with EMOVOME. For IEMOCAP, we selected two thresholds to categorize the continuous valence and arousal scores into the three categories

TABLE V
TEST RESULTS FOR 3-CLASS VALENCE PREDICTION. "PTM" SPECIFIES THE PRE-TRAINED MODEL APPLIED. FOR EMOVOME, THE TYPE OF RATER IS INDICATED: EXPERT (E), NON-EXPERT (N), OR COMBINED (C).

Database	Method	PTM	WA	UA
EMOVOME - E	Embeddings	Unispeech-L	59,38	57,53
- N	Emb+eGeMAPS	Unispeech-L	61,46	61,36
- C	Embeddings	Unispeech-L	62,50	61,64
IEMOCAP	Emb+eGeMAPS	Unispeech-L	62,35	60,04
RAVDESS	Emb+eGeMAPS	Unispeech-L	77,33	73,54

TABLE VI
TEST RESULTS FOR 3-CLASS AROUSAL PREDICTION. "PTM" SPECIFIES THE PRE-TRAINED MODEL APPLIED. FOR EMOVOME, THE TYPE OF RATER IS INDICATED: EXPERT (E), NON-EXPERT (N) OR COMBINED (C).

Database	Method	PTM	WA	UA
EMOVOME - E	Emb+eGeMAPS	Unispeech-L	42,71	43,57
- N	Emb+eGeMAPS	Hubert-L	70,83	58,73
- C	Emb+eGeMAPS	Unispeech-L	65,62	55,57
IEMOCAP	Emb+eGeMAPS	Unispeech-L	74,38	61,20
RAVDESS	Emb+eGeMAPS	Hubert-L	77,00	71,94

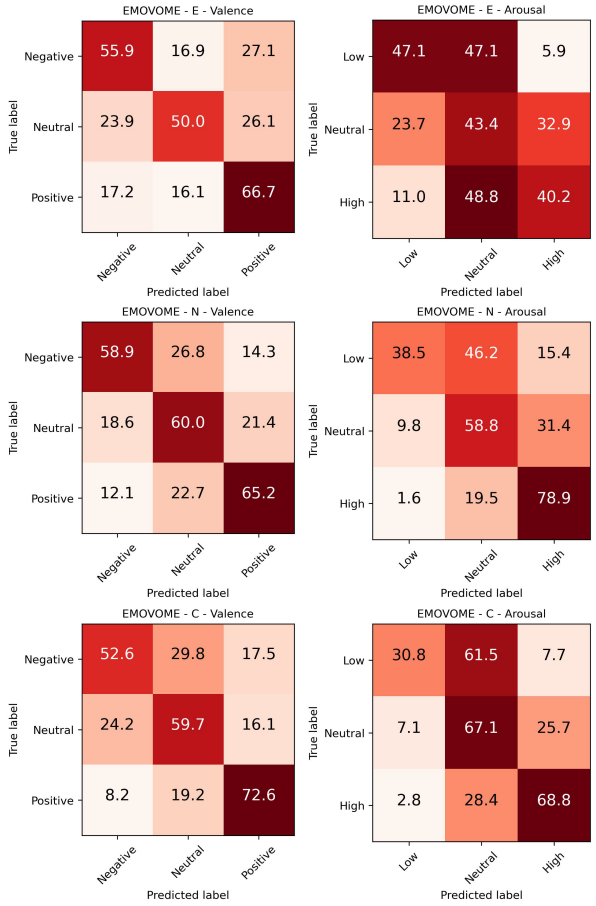

Fig. 3. Confusion matrix for the test samples for valence and arousal prediction on EMOVOME. The rater is indicated: expert (E), non-expert (N) or combined (C).

studied. The model misclassifies those samples close to the thresholds, as shown in Fig. 4, where the bars indicate the distribution of the original valence and arousal values in IEMOCAP for the test samples, and the legend refers to the categories formed by converting labels. In the case of RAVDESS, we used Russell's circumplex model of affect to derive the valence and arousal categories. Fig. 5 shows the distribution of the original emotion categories in RAVDESS for the test samples, the darker colors indicate the misclassified samples, and the legend refers to the categories resulting from the transformation of labels.

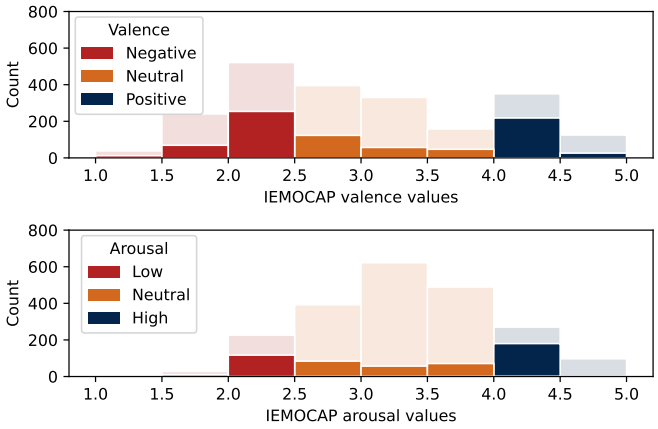

Fig. 4. Prediction errors in the test for valence and arousal in IEMOCAP. The darker colors indicate the misclassified samples.

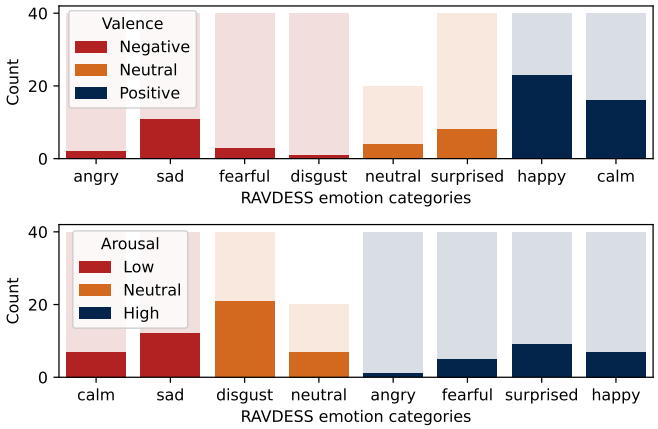

Fig. 5. Prediction errors in test for valence and arousal in RAVDESS. The darker colors indicate the misclassified samples.

Considering the emotion categories, test results for the three databases are presented in Table VII. It includes the model achieving the highest CV results among the three implemented methods. RAVDESS obtains the highest UA score (75.00%), followed by IEMOCAP (69.58%) and EMOVOME (42.58%). To evaluate which emotions are misclassified, Fig. 6 shows the confusion matrix for each database.

TABLE VII
TEST RESULTS FOR EMOTION CATEGORIES PREDICTION. "PTM" SPECIFIES THE PRE-TRAINED MODEL EMPLOYED.

Database	Method	PTM	WA	UA
EMOVOME - E	Embeddings	Unispeech-L	54,60	42,58
IEMOCAP	Embeddings	Unispeech-L	70,59	69,58
RAVDESS	Emb+eGeMAPS	Unispeech-L	74,67	75,00

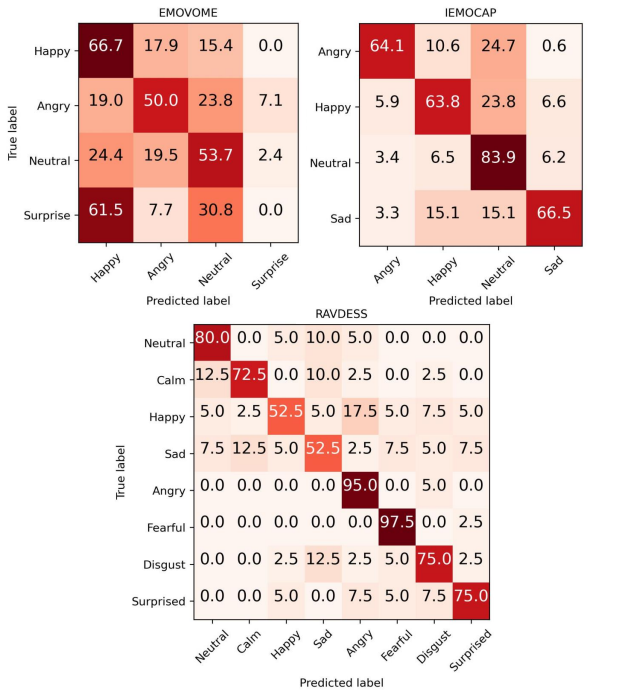

Fig. 6. Confusion matrix for the test samples for emotion category prediction on EMOVOME, IEMOCAP and RAVDESS.

Finally, we evaluated model fairness in terms of gender by calculating the difference between the UA for male speakers (UA_M) and the UA for female speakers (UA_F) on the test set, which is presented in Fig. 7. A positive difference means that the model had a better performance for male speakers.

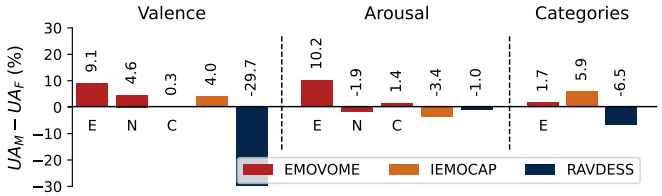

Fig. 7. Evaluation of gender fairness for the three labels and databases.

VI. Discussion

In this work, we have developed speech emotion recognition models for EMOVOME, a natural database comprising spontaneous emotions from real voice messages collected in the wild. We have compared different methodologies to create the SER models, and we have explored the influence of annotators' labels and the speaker's gender on their performance. Additionally, we have compared the results with two other reference databases in the literature, IEMOCAP and RAVDESS. The UA is used for comparison in all cases since it is more suitable for assessing unbalanced data. The subsequent subsections individually address each of the research questions, and we finally discuss the limitations and future research directions.

A. RQ1: What is the performance of SER models in EMOVOME when utilizing classical acoustic features versus state-of-the-art speech embeddings?

We compared three methods used to create the SER models: the baseline using acoustic features and machine learning (eGeMAPS), the pre-trained models (Embeddings) and the integration of pre-trained models and eGeMAPS features (Emb+eGeMAPS). As shown in Fig. 2, overall, the embedding approach significantly improves model performance compared to the non-transformer baseline, leading to an approximately 10% improvement in unweighted accuracy across all combinations. Unispeech-L consistently showcased better results than other pre-trained models (as shown in [45]), closely followed by hubert-L (which also outperforms wav2vec2 models in previous studies [11]). Interestingly, for the EMOVOME database, these two pre-trained models are followed by w2v2-xlsr-53-spa for valence and categories prediction and by x-vectors for arousal prediction. For the English databases, w2v2-xlsr-53-spa was surpassed by w2v2-xlsr-128 and x-vectors, and in general, the models pre-trained on multiple languages performed worse than those trained on English-only data (as in [11]). The lowest scores for all databases correspond to w2v2-xlsr-53 and w2v2-L-robust. The former also obtained the worst results in [63] for a multilingual model with Spanish and English among several pre-trained multilingual models. As for w2v2-L-robust, it obtained the best results on IEMOCAP in [11], but the model was trained on the MSP-Podcast corpus and was only tested on IEMOCAP. As for the Emb+eGeMAPS method, it shows similar CV results to the embeddings for the EMOVOME database, and UA values slightly increase (max. 3%) for IEMOCAP and RAVDESS in some cases. One potential explanation is that EMOVOME was recorded in natural, uncontrolled conditions, which might impact the reliability of eGeMAPS features, unlike the other two databases that were conducted in a controlled environment. Unispeech-L is again the top choice in most cases, but notably, other pre-trained models that initially performed lower (e.g. w2v2-xlsr-53 and w2v2-L-robust) show significant improvement when combined with eGeMAPS features. Test results mirror cross-validation trends, with RAVDESS consistently outperforming IEMOCAP and EMOVOME across the three predictions. In arousal prediction, IEMOCAP UA scores are higher than EMOVOME, while similar results are observed for valence prediction between IEMOCAP and EMOVOME.

Considering the test results (see Tables V, VI, VII), they follow the trends found in cross-validation, obtaining similar performance scores. In summary, we found that Embeddings and Emb+eGeMAPS give similar results for the top-performing pre-trained model, i.e., Unispeech-L. Both approaches improve the speech baseline models presented in [29] in approximately 10% UA for the combined label in valence (61.64% vs 49.27%) and arousal (55.57% vs 44.71%). These results also improve the test UA score obtained in [29] using the transcriptions for arousal (47.43%). Nevertheless, our speech-based models for valence do not surpass the transcription-based model presented in [29] (61.15%). This suggests that both

speech and text modalities offer complementary information across various emotion dimensions, underscoring the benefit of adopting a dimensional model approach for emotions.

B. RQ2: To what extent does the performance of EMOVOME, a natural database, compare with elicited and acted reference databases in the literature, such as IEMOCAP and RAVDESS?

We compared the results of our natural database, EMOVOME, with other two databases of different natures. We selected the well-known IEMOCAP database, including elicited speech, and RAVDESS, comprising acted recordings. We compared them in valence and arousal dimensions (see Tables V and VI), and emotion categories (see Table VII).

Considering the prediction of emotion categories, the test results follow the trend in cross-validation, where EMOVOME obtains the lowest evaluation metric (45.58% UA), followed by IEMOCAP (69.58% UA) and finally RAVDESS (75.00% UA) (even though the first two classify 4 emotions and the later 8 emotions). Although IEMOCAP contains elicited speech, part of the data consists of actors performing scripted dialogues, which could lead us to expect higher classification results. However, there is a notable difference between IEMOCAP and RAVDESS. The former contains utterances whose text content is different throughout the database, i.e., it is text-independent (same as EMOVOME). Conversely, RAVDESS is text-dependent, as the actors portrayed emotions using two fixed sentences. This could have caused the models for RAVDESS to prioritize variations related to emotion rather than differences in semantic content. Consequently, there is a performance gap in IEMOCAP compared to the other acted databases, as highlighted in [64]. Nevertheless, our classification results (70.59% WA, 69.58% UA) are comparable to the state of the art for IEMOCAP, which is in the range from 60.0% to 74.3% UA [11], especially considering that we implemented a speaker-independent approach, unlike other previous studies. Additionally, we examined the misclassified emotions for each database in Fig. 6. For EMOVOME, surprise is not correctly predicted in any case and is mainly confused with happy and neutral emotion, but this category is underrepresented in the data (see Table III). In the case of IEMOCAP, all emotions are mainly mistaken for the neutral category. For RAVDESS, the majority of emotions are accurately classified ($>$72%), except for happy, which is sometimes misclassified as angry, and sad, which is mistaken for calm.

Regarding the test results for arousal and valence dimensions, the SER models for the EMOVOME database achieve UA values of 61.64% for valence and 55.57% for arousal, considering the combined label between expert and non-experts. Surprisingly, the IEMOCAP database obtains similar values to EMOVOME, particularly 60.04% for valence and 61.20% for arousal. As for RAVDESS, the UA values are 73.54% and 71.94% for valence and arousal, respectively. Initially, one might have anticipated better outcomes in valence, given that earlier studies [11], [65] found that pre-trained models inherently capture linguistic information in the audio signal, aiding valence prediction. However, in our approach, we utilized pre-trained embeddings solely as feature extractors to obtain speech embeddings without fine-tuning the transformer layers for SER. This step has proven to be fundamental for the models to effectively learn the semantic content [11]. Furthermore, the data used for pre-training significantly impacts the models' capacity to capture linguistic information, with the inclusion of multi-lingual data adding complexity to the task [11]. These considerations might explain the relatively minor differences observed between valence and arousal predictions. In the case of RAVDESS, the use of fixed semantic content during recording prevented pre-trained models from leveraging text information for valence prediction. Furthermore, an important limitation in both IEMOCAP and RAVDESS is the label transformation applied to obtain the valence and arousal categories, as well as the unbalanced distribution of samples (see Table III). In the case of IEMOCAP, errors in the test occur mostly around the thresholds selected to obtain the categories (see Fig. 4). Moreover, the arousal data is highly imbalanced, primarily dominated by samples in the neutral category. However, for valence, there is a relatively more balanced distribution across the three categories. For RAVDESS, in valence prediction, the majority of errors are related to the happy and calm category (see Fig. 5). Calm was categorized as positive, but it is close to the neutral valence. This ambiguity may have influenced the model's positive valence predictions, especially considering these samples constitute half of the training data of this class. In arousal prediction, the model tends to make numerous errors in predicting disgust and sad. While disgust was labeled as neutral arousal, its proximity to neutral activation allows for variations, contributing to prediction inaccuracies. The same applies to sadness, as it is also close to neutral activation, and we consider it to be in the low arousal category. These findings suggest that predicting dimensions is more challenging than predicting categories.

In summary, the EMOVOME database achieves lower results than other reference databases in the literature. Both RAVDESS and IEMOCAP outperform EMOVOME in emotion categories, and the former also obtains higher valence and arousal prediction results. However, EMOVOME and IEMOCAP exhibit more comparable results in valence and arousal, possibly owing to their text-independent nature (unlike RAVDESS) and the approach used to transform the original labels in IEMOCAP into valence and arousal categories.

C. RQ3: What impact do annotator labels exert on the performance of SER models when dealing with challenging natural databases like EMOVOME?

Research on evaluating the influence of annotators' demographics on SER is limited, but evidence suggests that biases can arise based on their gender, age or educational level [14]. We hypothesize that this bias may be more pronounced when evaluating natural databases (such as EMOVOME) since they do not contain stereotypical emotion (unlike acted databases), and thus can be more challenging to label. In this work, we explore the difference between expert (E) and non-expert (N) annotations, as well as their combination (C) (see Table V and VI). Clinical psychologists, given their professional training and expertise, possess the necessary skills to recognize and interpret emotions, thus can be considered experts in the task.

Surprisingly, the SER models using the expert's labels achieved the lowest results for both valence and arousal. Non-experts got higher UA values in arousal prediction (58.73%), compared to the combined labels (55.57%) and the expert's labels (43.57%). For valence, the combined label obtained the highest UA score (61.64%), closely followed by the non-experts (61.36%) and lastly the expert (57.53%). It is worth highlighting that despite the expertise of clinical psychologists in the task, emotions remain highly subjective, and they can be influenced by individual experiences [14]. In fact, an examination of the confusion matrix (see Fig. 3) reveals discernible biases towards specific emotion categories in both valence and arousal results, potentially shaping what the SER models learn. For valence, the model trained on expert labels shows a bias towards positive valence. Conversely, the model trained on non-expert labels tends to misclassify neutral samples as either positive or negative. The combined label model mitigates the bias towards positive expert categories but increases misclassification for negative and neutral categories. In arousal, unlike valence, there's an unbalanced data distribution (see Table III). Expert labeled 43% of training data as neutral arousal, leading the model to often assign this category to test samples. Non-experts show a bias toward high arousal. Combining labels mitigates expert bias toward the neutral class and reduces non-expert bias toward the positive category. Consequently, the model is trained with fewer low arousal samples, leading to lower accuracy in this category.

Overall, the annotators' biases may cause differences in UA scores to up to 4% for valence and 15% for arousal. The better performance of the models based on the non-experts and the combined labels may be due to the higher number of annotators included, which may reduce individual biases in their interpretation of emotions.

D. RQ4: How does gender impact the fairness of SER model outcomes in EMOVOME compared to reference databases?

There is limited research on evaluating model fairness, particularly in the context of pre-trained models [27]. Our focus here is on gender, given the insufficient information for other attributes considered in the reference databases. We measure fairness by calculating the difference between the UA for the male speakers and the UA for the female speakers (see Fig. 7). For the EMOVOME database, models trained using expert labels exhibit a notable bias towards males, as the UA is around 10% higher for males in valence and arousal prediction and 1.7% in emotion categories prediction. The use of non-expert labels resulted in an increase in UA for males of 4.6% valence, but it was 1.9% higher for females on arousal. Interestingly, the combined label yielded the most similar results for both genders, with 0.3% for valence and 1.4% for arousal. In the case of IEMOCAP, again, UA was higher for male speakers in valence (4%) but lower in arousal (-3.4%). For categories, the UA for males was +5.9% compared to females. Finally, RAVDESS presents the highest difference between both genders, with the UA for females being 29.7% higher for females compared to males. For arousal and valence, the results are also higher for females (1% and 6.5%, respectively).

Overall, SER models obtain better test results for male speakers in EMOVOME, following previous studies in Spanish databases [21], [42]. In the reference databases, IEMOCAP aligns with the observed trend, while RAVDESS shows the opposite results. However, it's important to note that both databases have a limited number of speakers in the test set (two for IEMOCAP and five for RAVDESS), so no significant conclusions can be drawn from these results.

E. Limitations and future work

This section addresses the identified limitations, offering insights into areas for potential improvement in future research. Firstly, enhancing the annotation process of EMOVOME samples in emotion categories may involve introducing new raters to alleviate potential individual biases that could impact SER models. Similarly, expanding the pool of non-expert labels with more raters may help mitigate bias and achieve a more balanced data distribution regarding arousal. Additionally, other databases, such as EmoSpanishDB or MOUD, could be explored to increase the number of training samples and assess potential improvements in model performance. Furthermore, in the creation of SER models based on pre-trained models, the current use of average time pooling as an aggregation approach may exhibit suboptimal performance, particularly noticeable in the EMOVOME database, where audio duration exhibits significant variability. Future research endeavors will focus on refining time aggregation methods and exploring alternative techniques to address these challenges effectively.

VII. Conclusions and future work

A comprehensive study was conducted to assess the influence of different properties of EMOVOME, a natural speech database representative of real-world settings created for this work, on the performance of speaker-independent SER models. Superior results were achieved with state-of-the-art pre-trained transformer-based models compared to baseline models based on acoustic features. However, these results demonstrated lower performance compared to the SER models trained on the acted RAVDESS database. For the elicited IEMOCAP database, the prediction of emotion categories outperformed EMOVOME, but similar results were obtained for predicting valence and arousal. Notably, we found variations in EMOVOME results depending on the labels provided by different annotators, with superior outcomes observed when utilizing combined labels from both expert and non-experts. Interestingly, this combined label also yielded the most equitable results when assessing gender fairness, even though SER models generally performed better for male speakers.

Acknowledgments

This work was supported by the European Union's Horizon 2020 funded project "HELIOS: A Context-aware Distributed Social Networking Framework" (No 825585), by the Universitat Politècnica de València (PAID-10-20), by the Generalitat Valenciana (ACIF/2021/187 and PROMETEO/2020/024), by the Spanish Government (BEWORD PID2021-126061OB-C41) and by the Spanish Ministry of Science and Innovation for the DIPSY project (TED2021-131401B-C21).

REFERENCES

[1] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[3] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning—a systematic review," *Intelligent systems with applications*, p. 200266, 2023.

[4] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[5] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[6] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[7] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6442–6446.

[8] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[13] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[14] Y. Ding, J. You, T.-K. Machulla, J. Jacobs, P. Sen, and T. Höllerer, "Impact of annotator demographics on sentiment dataset labeling," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–22, 2022.

[15] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernadas, J. M. Oliver, D. Tena, and L. Longhi, "Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[16] J. M. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enriquez, and J. M. Pardo, "Analysis and modelling of emotional speech in spanish," in *Proceedings of international conference on phonetic sciences*, vol. 2, 1999, pp. 957–960.

[17] C. Á. Martínez and A. B. Cruz, "Emotion recognition in non-structured utterances for human-robot interaction," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 2005, pp. 19–23.

[18] S.-O. Caballero-Morales, "Recognition of emotions in mexican spanish speech: An approach based on acoustic modelling of emotion-specific vowels," *The Scientific World Journal*, vol. 2013, 2013.

[19] R. Barra-Chicote, J. Montero, J. Macias-Guarasa, S. Lufti, J. Lucas, F. Fernandez, L. D'haro, R. San-Segundo, J. Ferreiros, R. Cordoba *et al.*, "Spanish expressive voices: Corpus for emotion research in spanish," in *Proc. of LREC*. Citeseer, 2008.

[20] J. M. López, I. Cearreta, I. Fajardo, and N. Garay, "Validating a multilingual and multimodal affective database," in *International Conference on Usability and Internationalization*. Springer, 2007, pp. 422–431.

[21] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The mexican emotional speech database (mesd): elaboration and assessment based on machine learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1644–1647.

[22] "Emotional speech synthesis database elra-s0329," https://catalog.elra.info/en-us/repository/browse/ELRA-S0329/, 2011, accessed: 2023-12-30.

[23] E. Garcia-Cuesta, A. B. Salvador, and D. G. Pãez, "Emomatchspanishdb: study of speech emotion recognition machine learning models in a new spanish elicited database," *Multimedia Tools and Applications*, pp. 1–20, 2023.

[24] V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.

[25] A. Zadeh, Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2020. NIH Public Access, 2020, p. 1801.

[26] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs dimensional perception of italian emotional speech," in *Proc. Interspeech 2018*, 2018, pp. 3638–3642.

[27] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender De-Biasing in Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2823–2827.

[28] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy ai: A computational perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 1, 2022.

[29] L. Gómez-Zaragozá, R. del Amor, E. P. Vargas, V. Naranjo, M. A. Raya, and J. Marín-Morales, "Emotional voice messages (emovome) database: emotion recognition in spontaneous voice messages," 2024.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[31] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[32] B. T. Atmaja and A. Sasou, "Multilingual, cross-lingual, and monolingual speech emotion recognition on emofilm dataset," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1019–1025.

[33] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, and A. Nogueiras, "Interface databases: Design and collection of a multilingual emotional speech database." in *Proceedings of the 3rd international conference on language (LREC'02*, 2002.

[34] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[35] H. Pérez-Espinosa, C. A. Reyes-García, and L. Villaseñor-Pineda, "Emowisconsin: an emotional children speech database in mexican spanish," in *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer, 2011, pp. 62–71.

[36] H. Pérez-Espinosa, J. Martínez-Miranda, I. Espinosa-Curiel, J. Rodríguez-Jacobo, L. Villaseñor-Pineda, and H. Avila-George, "Iesc-child: an interactive emotional children's speech corpus," *Computer Speech & Language*, vol. 59, pp. 55–74, 2020.

[37] A. Arruti, I. Cearreta, A. Álvarez, E. Lazkano, and B. Sierra, "Feature selection for speech emotion recognition in spanish and basque: on the use of machine learning to improve human-computer interaction," *PloS one*, vol. 9, no. 10, p. e108975, 2014.

[38] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "Mexican emotional speech database based on semantic, frequency, familiarity, concreteness, and cultural shaping of affective prosody," *Data*, vol. 6, no. 12, p. 130, 2021.

[39] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo," *Speech Communication*, vol. 114, pp. 22–35, 2019.

[40] G. Assunção and P. Menezes, "Intermediary fuzzification in speech emotion recognition," in *2020 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–6.

[41] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," 2019.

[42] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, "The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning," *Sensors*, vol. 22, no. 7, p. 2461, 2022.

[43] G. McIntyre and R. Göcke, "The composite sensing of affect," in *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*. Springer, 2008, pp. 104–115.

[44] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[45] O. C. Phukan, A. B. Buduru, and R. Sharma, "A comparative study of pre-trained speech and audio embeddings for speech emotion recognition," *arXiv preprint arXiv:2304.11472*, 2023.

[46] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[47] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[48] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6329–6333.

[49] K. Kovacek, S. Livingstone, G. Singh, and F. A. Russo, "Ryerson audio-visual database of emotional speech and song (ravdess): Arousal and valence validation," submitted to the 16th Annual McMaster NeuroMusic Virtual Conference.

[50] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, vol. 12, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/1/327

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[52] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[53] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[54] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[55] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[56] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[57] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[58] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7169–7173.

[59] C. L. Moine, N. Obin, and A. Roebel, "Speaker Attentive Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2866–2870.

[60] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[61] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, vol. 2018, 2018, pp. 105–111.

[62] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[63] B. T. Atmaja and A. Sasou, "Ensembling multilingual pre-trained models for predicting multi-label regression emotion share from speech," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 1026–1029.

[64] M. S. Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital signal processing*, vol. 110, p. 102951, 2021.

[65] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing speech emotion recognition transformers for linguistic knowledge," *arXiv preprint arXiv:2204.00400*, 2022.

Lucía Gómez-Zaragozá graduated in biomedical engineering and obtained a master's degree in artificial intelligence, pattern recognition and digital imaging from the Polytechnic University of Valencia (UPV) in 2019 and 2021 respectively. She worked from 2019 to 2021 at the Human-Tech Institute, where she is currently pursuing a PhD. Her research interests are focused on signal processing and artificial intelligence for healthcare applications. Specifically, she is working on speech analysis and natural language processing for psychological assessment.

Óscar Valls graduated in telecommunications engineering from the Polytechnic University of Valencia (UPV) in 2021. He started working at the CVBLab research group and is currently pursuing an MSc in Artificial Intelligence at the Valencia International University (VIU). His research interests center around artificial intelligence for audio processing and speech analysis, with a particular emphasis on human sound perception and communications. Currently, he is involved in research projects on affective computing and the development of new methodologies in speech emotion recognition.

Rocío del Amor graduated and master's degree in biomedical engineering from the Polytechnic University of Valencia (UPV) in 2019 and 2020, respectively. Her research interests center around artificial intelligence for different data modalities, highlighting histological image and signal analysis. Currently, she is involved in different research projects on affective computing and is leading a project to evaluate emotions in psychology. She is also the founder of a spin-off related to her field of research.

María José Castro-Bleda holds a position of Full Professor of Computer Science at the Universitat Politècnica de València (UPV), in Spain. She is a professor of computer science engineering and also teaches courses in the master's degree program in artificial intelligence, pattern recognition, and digital imaging. María José is a senior researcher at the Valencian Research Institute for Artificial Intelligence (VRAIN). Her primary research interests lie in machine learning and deep learning, with a particular emphasis on speech and handwritten text recognition, as well as natural language processing. Her work seeks to advance these areas, exploring innovative approaches to artificial intelligence to transform technology and information systems in these fields.

Valery Naranjo is Full Professor (tenure position) at Universitat Politècnica de València (Spain) and founding director of the Computer Vision & Behavior Analysis Lab (CVBLab). Her area of expertise is signal, video and image analysis to develop artificial intelligence algorithms applied to different fields such as diagnostic aid, human behavior analysis and specific applications in the industrial sector. During her research career she has participated in more than 80 competitive research projects, leading more than 30 as principal investigator (including European projects), as well as in more than 20 research contracts with companies. Dr. Naranjo has disseminated the results of her research activity in about 200 scientific articles in journals and conferences. It is also worth mentioning the doctoral theses she has supervised and her five six-year research periods recognized. She is also the founder of a spin-off related to her field of research.

Mariano Alcañiz Raya is a full professor at the Polytechnic University of Valencia, Spain, and Director of the European Laboratory of Immersive Neurotechnologies. He is a professor of biomedical engineering and has courtesy appointments in virtual reality. His research interest is focused on a better understanding and enhancement of human cognition, combining insights and methods from computer science, psychology, and neuroscience. From a technological point of view, his objectives are to improve interactive technology in virtual environments used in different formats and the development of algorithms, methods, and techniques for ubiquitous and non-obtrusive measurement of human activity. His research is being applied in different fields like health, psychology, marketing, human resources, education, and training.

Javier Marín-Morales is an assistant professor of Statistics at the Universitat Politècnica de València, Spain. His research focuses on developing intelligent systems that enable computers to analyze, recognize, and simulate human communicative dynamics. Additionally, he investigates the creation of multimodal machine-learning models for the recognition of emotions, psychological traits, mental illnesses, and neurodevelopmental disorders, through the use of behavioral and physiological biomarkers.