Multimedia Tools and Applications (2024) 83:13093–13112
https://doi.org/10.1007/s11042-023-15959-w

EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database

Esteban Garcia-Cuesta[1] · Antonio Barba Salvador[2] · Diego Gachet Pãez[3]

Received: 20 May 2022 / Revised: 1 April 2023 / Accepted: 29 May 2023 /
Published online: 4 July 2023

Abstract
In this paper we present a new speech emotion dataset on Spanish. The database is created using an elicited approach and is composed by fifty non-actors expressing the Ekman's six basic emotions of anger, disgust, fear, happiness, sadness, and surprise, plus neutral tone. This article describes how this database has been created from the recording step to the performed crowdsourcing perception test step. The crowdsourcing has facilitated to statistically validate the emotion of each collected audio sample and also to filter noisy data samples. Hence we obtained two datasets EmoSpanishDB and EmoMatchSpanishDB. The first includes those recorded audios that had consensus during the crowdsourcing process. The second selects from EmoSpanishDB only those audios whose emotion also matches with the originally elicited. Last, we present a baseline comparative study between different state of the art machine learning techniques in terms of accuracy, precision, and recall for both datasets. The results obtained for EmoMatchSpanishDB improves the ones obtained for EmoSpanishDB and thereof, we recommend to follow the methodology that was used for the creation of emotional databases.

Keywords Affective analysis · Speech emotion recognition · EmoMatchSpanishDB · Language resources · Machine learning

✉ Esteban Garcia-Cuesta
esteban.garcia@fi.upm.es

Antonio Barba Salvador
antonio.barba@universidadeuropea.es

Diego Gachet Pãez
diegogabriel.gachet@ufv.es

[1] Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Av. de Monteprincipe, S/N, Boadilla del Monte, 28660 Madrid, Spain

[2] Departamento de Ciencia Y Computación, Universidad Europea de Madrid, C/Tajo S/N, Villaviciosa de Odón, 28670 Madrid, Spain

[3] Departmento de Automãtica, Universidad Francisco de Vitoria, Av de Majadahonda S/N, Madrid, 28223 Madrid, Spain

1 Introduction

Affective computing research [27] to measure and recognize an individual's emotional state and model emotional interactions between humans and computer systems is increasing nowadays. Among others, one of the main important characteristic when interacting with other people is the ability to empathize (commonly express by the term Theory of Mind ToM [31]). This ability relies on the detection of emotions that others show through their facial and verbal expressions, physiological responses, and body gestures. Among them it is very important the vocal expressiveness that conveys for 38% of emotional information associated to a message [23]. This fact opens the possibility of adapting the verbal human–machine interaction to the emotional state of the user providing a better user experience. This capability has many applications such as chatbots (e.g. helping desk service), robotics (e.g. assisting elderly people), or Augmentative and Alternative Communication Systems (AACs, e.g. assistive emotional learning systems for people with disabilities). Despite there is no general agreement on how to define an emotion, there is some consensus in the use of a working definition that consist of a mere list of analogous terms such as'anger, disgust, fear, happiness, sadness, surprise' [14] which are a categorical description of a more complex human state that includes emotional experience and regulation processes according to [45].

During the last decade speech emotion recognition (SER) technology has matured enough to be used in noise-free and speaker dependent practical applications such as health care [44], education [5], or robotics [21]. Most of these studies use Support Vector Machines (SVMs) [35], or recurrent and deep neural networks [1]. These solutions rely on databases such as Emo-DB [3] that are usually generated using actors to simulate the emotions. A review and comprehensive analysis of datasets can be found at [36, 43]. Moreover, despite there are some practical applications, still SER technology is not mature enough to recognize the six speaker-independent Ekman's emotions and usually a subset of emotions (e.g. sentiment classification) are set to obtain good enough classification accuracies.

Most of the SER systems are developed for English and there is only a few for other languages because the lack of public datasets. For instance in Spanish there are only two simulated datasets recorded by a few actors according to [43]. Due to the worldwide Spanish language importance, the scarcity of speech emotionally annotated databases for this language, and the need of continue exploring some of the mentioned problems, in this article we present a new elicited (by non-actors) Spanish database and its application to the problem of human speech emotion recognition (SER). As mentioned, we didn't employ acted speech but elicited by combination of inductions productions similarly to DEMoS [26].

The creation of a dataset is costly because there is a need of actors to simulate the emotions or there is a need of a validation process for elicited or real scenarios. The creation of elicited non-actors datasets demand the detection of those audios that are not properly expressed and should be discarded. This can be done by a perception test that relies on human responses. However, multiple independently labeling actions are needed to statistically demonstrate that consensus exists for each audio sample. Crowdsourcing approaches are appealing for the accomplishment of this task as it has been already used successfully in other contexts [25][25]. We make use of this crowdsourcing approach as explained in the next Section 2 to create EmoSpanishDB and EmoMatchSpanishDB.

In Section 3 the experimental design is presented including the extracted audio features used to test the two databases. At Section 4 the results are discussed and finally, some conclusions are provided at Section 5.

2 Creation of Spanish emotional speech Corpus: EmoSpanishDB and EmoMatchSpanishDB

One of the major problems working in emotional detection studies is the limited number of speakers available in the current databases. This speaker specific information may play considerable role if speech utterances of the same speaker are used for training and testing in machine learning models. On the other hand, developed models may produce poor results due to the lack of generality if speech utterances of different speakers are used for training and testing the models [17]. The Fig. 1 shows the whole database workflow creation that we executed.

2.1 Spanish sentences selection

The first step is the selection of the Spanish sentences to be recorded. According to the'Real Academia Española' RAE [33] there are 23 phonemes in the central area of Spain (see Table 1) and several phoneticians have studied their statistical distribution in a regular conversation. In [34] the authors reviewed and summarized the different studies obtaining a phonetic appearance global distribution (see Table 1). In order to replicate a regular conversation, we have used these phonetic percentages to create a total of 12 sentences that contains all these Spanish's phonemes within the minimum and maximum defined ranges. Moreover the sentences have been created without emotional semantic connotation to avoid any emotional influence in the speaker during the speaking and have similar length (approx. 2 s). We also analyzed the number of labeled audio samples per emotion and sentence to assure this independence (see Table 2) and performed for every pair of sentences a Mann Whitney U non-parametric test to verify that different sentences obtained similar results (H_0 considers that a pair of sentences follow the same distribution). The results for all cases had a p-value higher than 0.05 accepting the null hypothesis (the lower p-value is 0.12 between S7 and S8). The sentences used for the creation of the dataset can

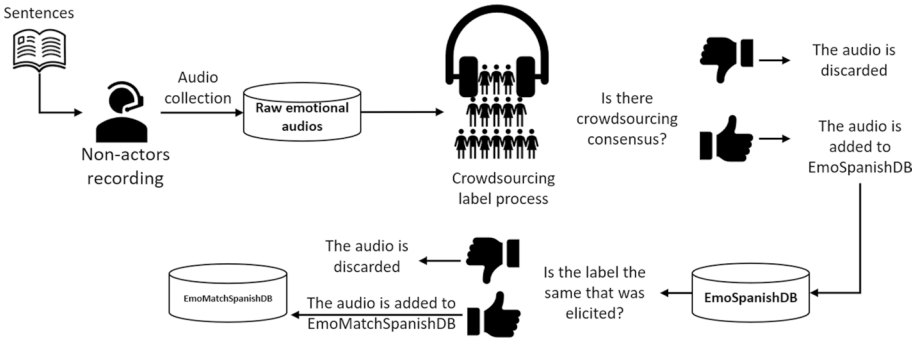

Fig. 1 EmoSpanish and EmoMatchSpanishDB workflow creation process description

Table 1 List of Spanish phonemes, its number of appearance within the sentences used in the study, and theoretical percentage ranges of appearance in Spanish language

Spelling	Sound	Appear	Freq (%)	Percentage (%)
a	/a/	50	13,09	From 11.45 to 15.21
b/v	/b/	11	2,88	From 1.92 to 3.6
c + a,o,u/k/q	/k/	15	3,92	From 3.33 to 4.64
c + e,i/z	/z/	7	1.83	From 1.42 to 2.49
ch	/ch/	1	0.26	From 0.15 to 0.57
d	/d/	18	4.71	From 3.81 to 5.42
e	/e/	53	13.87	From 9.73 to 14.99
f	/f/	2	0.52	From 0.51 to 1.46
g + a,o,u	/g/	4	1.04	From 0.71 to 1.46
i	/i/	25	6.54	From 4.2 to 8.6
j/g + e,i	/j/	3	0.78	From 0.37 to 1.02
l	/l/	20	5.23	From 2.96 to 5.46
ll/y	/y/	4	1.04	From 0.09 to 2.94
m	/m/	11	2.88	From 2.48 to 3.73
n	/n/	23	6.02	From 2.34 to 7.99
Ã ±	/Ã ±/	1	0.26	From 0.13 to 0.36
o	/o/	42	10.99	From 9.11 to 11.2
p	/p/	9	2.35	From 2.1 to 2.97
-r-	/r/	17	4.45	From 4.25 to 8.24
r-, -rr-	/rr/	3	0.78	From 0.39 to 1.17
s	/s/	33	8.63	From 4.26 to 10.24
t	/t/	19	4.97	From 4.29 to 5.32
u	/u/	11	2.88	From 1.76 to 3.33

be consulted in Appendix. It is worth to mention that people of the central area of Spain pronounce the phoneme /ll/ as /y/ and the letter "h" has no sound at all, and the "x" is the sum of the phonemes /k/ and /s/ (see Table 1).

Table 2 Number of labeled audio samples per emotion and sentence–EmoSpanishDB

Spelling	Happiness	Disgust	Anger	Fear	Neutral	Surprise	Sadness
S1	41	36	48	43	49	45	41
S2	44	37	41	40	46	41	30
S3	43	40	40	40	45	42	43
S4	50	42	46	39	45	41	37
S5	38	43	44	41	50	38	43
S6	42	35	44	38	47	40	44
S7	42	38	42	41	46	37	40
S8	44	38	42	41	48	45	42
S9	50	37	39	45	41	39	41
S10	43	33	44	38	48	46	40
S11	44	38	42	45	44	40	42
S12	46	37	39	38	47	46	45

Fig. 2 Noise-free professional radio studio used for record audio recording

2.2 Audio samples recording

A total of 50 individuals were recorded playing out the 12 selected sentences seven times (one for each Ekman's basic emotion [9],'anger, disgust, fear, happiness, sadness, surprise', plus neutral). The total number of audio samples collected were 4200 audio files (emotional raw audios). The participants' demographics are shown in Table 3. We can confirm that this database is the first in Spanish language that contains elicited emotional voices played out by non-actors. Moreover, it is also the largest publicly available dataset compared to previous ones [13, 24], and Berlin emotional speech database [3] that has 800 sentences. A professional radio studio was used to record these audios Fig. 2. The audio files were recorded noisy free in PCM format with a sampling rate of 48 kHz and a bit depth of 16 bits (no compressed audio). The audios were then incrusted within a waveform audio file format container (.wav). A dynamic mono channel cardioid microphone (Sennheiser MD421) and the AudioPlus (AEQ) software have been used to record the signal and the voice is obtained at a hand's distance from the microphone. At the beginning the speaker has a sheet with all the sentences and emotions that he/she had to simulate. Next, to induce emotion, he is shown a MIP (Mood Induction Procedure by watching pictures) image extracted from the Geneva Affective Picture Database (e.g., bugs or spiders) [7] along with MIP empathy in a manner similar to DEMoS [26]. This empathy MIP is based on the creation of an empathic reaction by reading text with an emotional content [12]. Hence, the speaker look at an image representing the emotion and listens a short text to induce the emotion before the recording.

Table 3 Dataset Demographics

Participants	Men	Women	Average Age	Age Standard Deviation
50	30 (60%)	20 (40%)	33.9	10.08

2.3 Crowdsourcing corpus label process

To ensure that the proposed creation of elicited audios doesn't contain noisy samples, a perception test was conducted by a set of independent individuals using a crowdsourcing approach. The goal of this crowdsourcing process is to label with an emotion all the recorded audio samples. One of the major constraints in using crowdsourcing for human computation purposes is the lack of guarantees on the expertise of the participants. Thereof the tasks that the participants have to resolve need to be small, simple, and well-formed (usually called micro-task). Moreover, due to human error and bias, it is mandatory to ensure that collected responses are adequately reliable and a quality control mechanism is needed [25]. In our case, a straightforward multiple-choice emotion type questionnaire is used. The recorded raw audios are played by participants and they have to select the emotion that they perceive among the seven proposed. They can play the audio many times until they decide to label it or skip it (the participant can quit the process any time). The selection is then stored in a remote database for further analysis. We refer with the term crowd-labeling [28] to the set of micro-tasks performed by people to label with an emotion the recorded audios.

It has been illustrated that the quality of the majority vote schema for multiple responses collected by a crowdsourcing tool is at least as good as that of answers provided by individual experts when a large enough number of labels are obtained for a sample and the majority voting schema is used [41]. If a large number of samples needs to be labeled there is also a need of a large number of people to perform the micro-tasks. To minimize the number of micro-tasks performed by participants (the total number of micro-tasks needed to label all emotional raw audios) a binomial p-value significance'two-sided test' is proposed as a metric to find consensus and assign an emotion to an audio. The goal is to set a label to an audio sample as soon as any of the possible emotions obtains a significant p -value< 0.05. The'two-sided test' tests the null hypothesis that the probability of success in a Bernoulli experiment is p. This value p has been established $p = 1/7 \approx 0.143$ assuming that all emotions are equally probable, i.e. 1 out of 7 is the expected probability for each possible label given that we have 6 emotions plus neutral. Following this approach, in a majority voting scenario and assuming that participants are non-expert, it is known that an average of 4 labels are needed in order to emulate expert-level label quality [41]. According

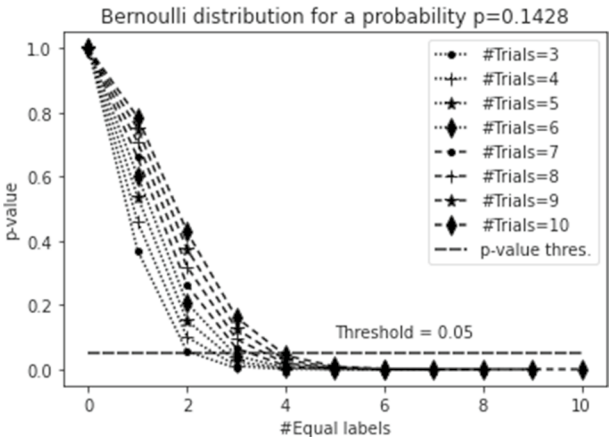

Fig. 3 Bernoulli probabilities distribution for different number of trials with p = 0.1428 used to evaluate consensus during the crowdsourcing process

Table 4 Percentage of matches between the EmoSpanishDB labels and the original elicited emotion categorized per emotion

Emotions	No. of samples in EmoSpanishDB database	Nº and percentage of labels that matches with the original elicited emotion (%)
Happiness	528	258 / 48,9%
Disgust	453	128 / 28,3%
Anger	519	318 / 61,3%
Fear	489	262 / 53,6%
Surprise	501	298 / 59,5%
Sadness	501	275 / 54,9%
Neutral	559	481 / 86,0%

to that value we limited the maximum number of trials to 10 to alleviate the crowdsourcing effort. Hence, the stopping criteria for an audio label depends on the number of equal labels needed to reach consensus. For instance, given that the maximum number of labels allowed during the crowd-labeling process is 10, otherwise the audio is discarded, at least 3 equal labels are needed for 3,4,5 and 6 trials, or 4 for 7,8,9 and 10 trials (the most relaxed cases need 3 out of 6 or 4 out of 10 equal labels). Figure 3 shows the Bernoulli probability distribution curves for the just mentioned number of possible trials from 3 to 10. This criteria has been applied to the whole set of raw audio samples. The audios are shown randomly to the participants until they are labeled or discarded. For completion of the crowd-labeling process a total of 21,490 micro-tasks were needed with an average of 5 labels per audio sample that is just 1 above the average obtained in [41], and 194 independent native Spanish speakers were involved.

A total of 3550 audios were labeled with an emotion and those audios compose the EmoSpanishDB database[1] (the other 650 were discarded because didn't reach consensus. See Table 2). Among the labeled samples at EmoSpanishDB, 2020 also matched the original elicited emotion and those audios are collected at EmoMatchSpanishDB[2] (see the number and percentage of matches per emotion at Table 4).

3 Experimental design and methodology

Figure 4 shows the experimental flow. The speech feature extraction plays a key role in SER systems to reflect the most important emotional characteristics. The most common categorization of emotional acoustic features include two categories, spectral and prosodic [6]. Following these categories we have selected the following set of features that are adequate for the emotion classification task as proposed in other works [39]. Indicate that the spectral features (frequency-based features) are obtained by converting the time based signal into the frequency domain using the Fourier Transform. Since the speech signal is constantly changing, the features that represent the spectrum can't be extracted.

[1] The resulting database can be accessed at EmoSpanishDB folder https://doi.org/10.6084/m9.figshare.14215850.v1 under disclosure.

[2] The resulting database can be accessed at EmoMatchSpanishDB folder https://doi.org/10.6084/m9.figshare.14215850.v1 under disclosure.

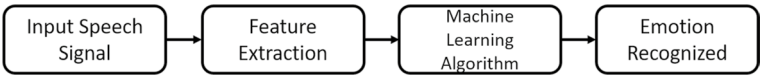

Fig. 4 Speech emotion recognition experimental flow

Thus, the signal is framed into 20 ms windows to analyze its frequency content in a short time segment of a longer signal (this is a typical time window size but other sizes may be also a valid option). The spectral features extracted were: first 13 Mel Frequency Cepstral Coefficients (MFCCs) and their mean, standard deviation, kurtosis and skewness, the first and second derivatives of MFCC (Δ MFCC and ΔΔ MFCC), spectral centroid, spectral flatness, spectral contrast, and Linear Predictive Coding (LPC). The prosodic features represent those supra-segmental elements of oral expression which are elements that affect more than one phoneme and can't be segmented into smaller units, such as accent, tones, rhythm and intonation. Among the prosodic features we used the fundamental frequency (F_0), intensity, and tempo. To obtain a common number of input features, the mean and some statistics are calculated for all frames, resulting in a total of 140 features to be used as input for the machine learning models (Table 4 contains a full description of these features). The spectral features were extracted using Praat [15] and the prosodic using librosa [22]. Moreover, we also tested EmoMatchSpanishDB with other two commonly used features sets: eGeMaps [10] (that contains 88 features) and ComparE [38] (that contains 6373 features). These features were extracted using OpenSmile library [11].

The machine learning algorithms were validated using Cross-Validation (CV). In standard CV, instances partitioning is based on random sampling of file from a pool wherein all speakers are mixed (meaning that is not speaker-independent) into CV partitions. We have also adapted the CV process to validate the models for Leave-One-Speaker-Out (LOSO) scenario. For this second approach data has been split into 10 folders and complete individuals' audios are split. Therefore, fold 1 contains the 1/10 individuals, fold 2, other 1/10 different individuals, and so on. This guarantees that, at least, training and testing partitions will never contain instances belonging to the same individual. We also tested two different set of features to compare EmoSpanishDB and EmoMatchSpanishDB datasets. The first only contain the 13 Mel frequencies and the second the extended 140 features described in Table 5. Moreover, we tested EmoMatchSpanishDB with eGeMAPS and ComparE feature sets that are commonly used in other SER studies.

Given a ground-truth and a prediction, the machine learning models were optimized using the unweighted F1-score metric because its robustness to the imbalance in the number of samples for each category. The F1-$score = \frac{2*Precision*Recall}{Precision+Recall}$ describes the harmonic mean between precision $precision = \frac{TP}{TP+FP}$ and recall $recall = \frac{TP}{TP+FN}$. The unweighted accuracy is also measured for interpretation purposes being $accuracy = \frac{TP+TN}{Total\ number\ of\ cases}$. We applied a min–max normalization to each feature before starting the learning process. EXtreme Gradient Boosting (XGBOOST), Support Vector Machines (SVM), and Feed-Forward Deep Neural Network (FFNN) machine learning methods have been selected to measure the SER performance on the presented datasets (EmoSpanishDB and EmoMatchSpanishDB).

These models are defined by some hyper-parameters that require to be set. XGBOOST has three main hyper-parameters that were fitted: minimum number of samples per leaf, number of estimators, and tree depth. The following range of values for each hyper-parameter were tested: minimum number of samples per leaf: 3, 5, 10; number of estimators (GBC): 5, 10 50, 100; tree depth (GBC): from 1 to 10 in steps of 2; sampling of

Table 5 Description of the extracted audio features

Category	Name	Related Statistics	Number	Description
SPECTRAL	4*MFCC	Mean	13*	MFCC coefficients for the representation of speech based on human auditory perception. *The 13 mean MFCC coefficients have been chosen as basic features for comparative purposes in the experiments
		Standard Deviation	13	
		Kurtosis	13	
		Skewness	13	
	Δ MFCC	Mean	13	First derivative of MFCC coefficients. It contains relevant information related to the speed of speech
	ΔΔ MFCC	Mean	13	Second derivative of MFCC. Valuable information related to the acceleration of speech
	Spectral Centroid	Mean	1	Denotes the weighted average of the frequencies of a speech power spectrum generated from a speech frame. These coefficients present valuable information that can be related to the prosodic features of the voice, such as the speed of speech and acceleration during speech [30]
		Standard Deviation	1	
		Kurtosis	1	
		Skewness	1	
		Max Value	1	
		Min Value	1	
	Spectral Flatness	Mean	1	Is the ratio of the geometric mean to the arithmetic mean of the magnitude spectrum of the signal [20].A high flatness indicates a similar energy in all spectral bands while a low flatness indicates the energy is concentrated in a small area of the spectrum
		Standard Deviation	1	
		Kurtosis	1	
		Skewness	1	
		Max Value	1	
		Min Value	1	
	Spectral Contrast	Mean	7	Is the difference between the highest and the lowest values in all audio frames and it is useful to determine the variance in the values extracted for a frame [16]
		Standard Deviation	7	
		Kurtosis	7	
		Skewness	7	

Table 5 (continued)

Category	Name	Related Statistics	Number	Description
		Max Value	1	
		Min Value	1	
	LPC		17	The linear predictor models the macro-shape or envelope of the spectrum. Is a method for characterizing speech from a series of predictive coefficients and it is based on the principle of autocorrelation [47]. The number of coefficients used were 17
PROSODIC	F_0	Mean	1	Average number of oscillations per second. Is a very relevant feature in speech emotion recognition since it represents the vibration rate of vocal cords. It has information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure [40]
	Intensity	Mean	1	Is the average amount of energy passing through a unit area per unit of time. Sounds with higher intensities are perceived to be louder. Emotions like anger have more intensity than others like sadness
	Tempo	Mean	1	It determines the rhythm or the speaking speed of a person [4] and is measured in beats per minute (bpm). A tempo higher than normal speech value suggest stress or excitement, while a lower tempo suggest sadness

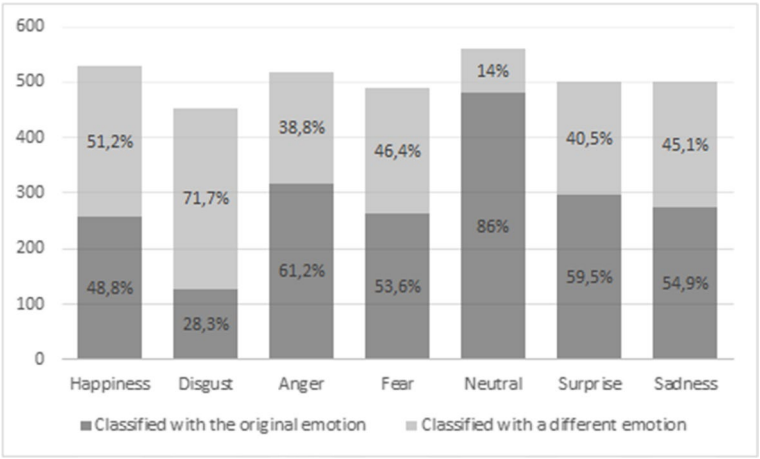

Fig. 5 Percentages of labeled emotions vs. original ones after crowdsourcing process led emotions vs. original ones after crowdsourcing process

features: 0.5, 1; sampling of samples: 0.5, 1. SVM has three main hyper-parameters: kernel type, gamma value that defines the influence of each point, and C value that establishes how large is the margin separation among classes. The following range of values for each hype-parameter were tested: kernel type: linear, polynomial, and radial; gamma values: $1e^{-4}, 1e^{-3}, 1e^{-2}$; polynomial degree: 2, 3, 4, 510, and 20; C values: $1e^{-4}, 1e^{-2}, 1e^{-1}, 1, 10, 100$. FFNN has four main hyper-parameters: the number and size of hidden layers, learning rate, and the activation function. The following range of values were tested: layer size depth: 2, 4; number of neurons: 2, 5, 10, 20; learning rate: 0.001, 0.01; activation function: RELU.

In order to tune the hyper-parameters, a systematic procedure known as grid-search was used. This method tries all possible combinations of hyper-parameter values. Models for each hyper-parameter combination are trained using the above exposed cross-validation procedure. The best combination on the validation set is selected. Two experiments were done. The first is speaker dependent that allows that audios of the same speaker may be in training, validation, and test sets, and the second is speaker independent that assures that all the samples of an individual are only in one of the three sets. In both cases we use a k-fold = 10 and to avoid bias on the test we computed the average of the results repeating the same experiment 10 times (each one using different randomly selected samples). The experiments were done using EmoSpanishDB that contains 3550 audios and EmoMatchSpanishDB that contains 2020 audios.

4 Results and discussion

Figure 5 shows the percentages of labeled audio samples per emotion after the crowdsourcing validation process was applied and consensus was achieved. It also shows the percentage of those that also match with the original elicited emotion. A total of 3550 labeled audios were obtained via crowd-label consensus and their distribution by emotion is quite homogeneous but in the case of disgust it is a bit lower and in the case

Table 6 Distribution of elicited emotion vs. crowdlabeling results (columns represents the elicited emotions and rows the crowdlabeled emotion) during the creation of the EmoMatchSpanishDB procedure

	Happiness	Disgust	Anger	Fear	Neutral	Surprise	Sadness
Happiness	**258**	24	52	13	1	91	5
Disgust	1	**128**	15	2	4	4	6
Anger	17	51	**318**	4	0	27	1
Fear	5	27	7	**262**	2	7	41
Neutral	118	143	83	67	**481**	73	165
Surprise	127	36	42	40	6	**298**	8
Sadness	2	44	2	101	65	1	**275**
Total	528	453	519	489	559	501	501
Percentage	48.86%	28.26%	61.27%	53.58%	86.05%	59.48%	54.89%

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

of neutral a bit higher. An explication to this fact after observation is related with the tendency of crowd-labeling an audio as neutral whenever the crowd-labeler is not sure about the emotion that the audio contains. This is specially relevant for emotions that are harder to detect as disgust as was shown in [19]. It can also be observed that the percentage of coincidence between the originally elicited emotion and the crowdsourcing results varies significantly for the different emotions. The percentage of match ranges from 86% for neutral to 28.3% for disgust. Ordered from lower to higher: disgust, happiness, fear, sadness, surprise, anger, and neutral. It is observed that it is more difficult for humans, including both the expression and recognition of an elicited emotion, to find consensus for disgust and happiness, and easier for anger or neutral. In [19] the authors obtained similar conclusions. They found that human emotion recognition (crowd-labelers) have higher accuracy and confidence ratings labeling anger and neutral emotions and in contrast lower for disgust. They also found an interesting pattern saying that the categorization of surprise has more confident than disgust and fear that also occurs in our case.

There are different reasons why a person has difficulty to express or recognize emotions. It can be due to Social Anxiety Disorders [2] that leads a person towards interpreting ambiguous cues as negatives or a threat [42]. In prosodic emotion recognition has been also demonstrated a bias towards correct identification of fearful voices and a decrease identification of happy voices [32] and, more recently the same behavior

Table 7 Unweighted F1-Score and accuracy results obtained for different machine learning techniques, EmoMatchSpanishDB all samples and LOSO

	All samples						LOSO					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
NÂº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
	0.481	0.447	0.478	0.515	0.341	0.483	0.301	0.345	0.309	0.370	0.301	0.391
	0.611	**0.643**	0.512	0.565	0.480	0.571	**0.424**	**0.452**	0.380	0.439	0.392	0.463

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

Table 8 Unweighted F1-Score and accuracy results obtained for different machine learning techniques, EmoMatchSpanishDB women and men

	Men						Women					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
Nº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
	0.478	0.552	0.422	0.488	0.432	0.521	0.467	0.478	0.439	0.512	0.374	0.522
	0.615	**0.655**	0.513	0.592	0.501	0.584	**0.593**	**0.632**	0.546	0.594	0.502	0.554

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

has been observed in a study over 31 SAD patients [46]. This can also be observed in Table 6 where anger has the highest match percentage 61.27% (exempting neutral that may not be considered as an emotion itself) and happiness has the second lowest 48.84% after disgust.

From an application perspective, all the above explanations suggest that it is hard to have just one homogeneous dataset for all applications but there is always some uncertainty that must be considered as part of human subjectivity and not part of the error in the model itself. This reasoning is not applicable to other datasets and studies [40] that just analyze audios simulated by actors and hence a'perfect' emotion is assumed.

Observe that the number of audio files that reached consensus on a label different from the elicited one is quite large (it is the difference between the audio samples of EmoSpanishDB and EmoMatchSpanishDB 3550–2020 = 1530 samples). As introduced above, the main reason is that there is a tendency to label an audio sample as neutral when there is not enough confidence about the emotion it contains (note that there are 1130 audios labeled as neutral whereas only 600 where expected according to the elicited process). Thereof, we created and compared two alternative datasets EmoSpanishDB and EmoMatchSpanishDB showing that the latter solves (at least partially) the problem and it is more accurate, as it was expected.

Table 7 shows the results of SVM, XGBOOST, and FFNN for the different experiments with 13 and 140 features. In all cases it is observed that the use of EmoMatchSpanishDB improves the results over the EmoSpanishDB (improvement of 19% and 10% for all samples and LOSO respectively). This means that when the audio is expressed and recognize with the same emotion the ML model is able to learn better that using all the sample audios labeled by the human crowdsourcing recognition process. This reinforces the previous reasoning and explanations were we affirmed

Table 9 Unweighted F1-Score and accuracy results obtained for different machine learning techniques, EmoSpanishDB all samples and LOSO

	All samples						LOSO					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
Nº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
	0.389	0.442	0.389	0.465	0.281	0.432	0.253	0.305	0.298	0.357	0.283	0.380
	0.512	**0.563**	0.411	0.524	0.453	0.541	**0.386**	**0.421**	0.348	0.429	0.357	0.429

Table 10 Unweighted F1-Score and accuracy results obtained for different machine learning techniques, EmoSpanishDB women and men

	Men						Women					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
Nº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
	0.435	0.517	0.414	0.498	0.292	0.460	0.410	0.489	0.365	0.420	0.230	0.351
	0.460	**0.555**	0.442	0.531	0.480	0.571	**0.465**	**0.510**	0.520	0.545	0.411	0.494

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

that there is some uncertainty associated to different human expressiveness and recognition capabilities and were solved (at least partially) using the EmoMatchSpanishDB alternative.

The best results, using f1-score as metric to optimize the ML models, are always obtained for SVM method (see Table 7 for all samples and LOSO, and Table 8 for men and women comparatives). Note that in a multi-class problem it is desirable to get a unique score to get an global overview of the performance. For this purpose Cohenâ€™s Kappa Coefficient was used and values of 0.573 and 0.394 were obtained for all samples at EmoMatchSpanishDB. This difference is reasonable because LOSO use independent speakers to test the model. Table 8 shows the results for EmoMatchSpanishDB separated by gender and no significant differences were observed in the results for the three models (p-value > 0.99). For EmoSpanishDB, not only the split of the dataset doesn't improve the results but worsen them as shown in Table 8. These facts reject the need of separating the audios by gender to obtain better results as other studies have also defended [48].

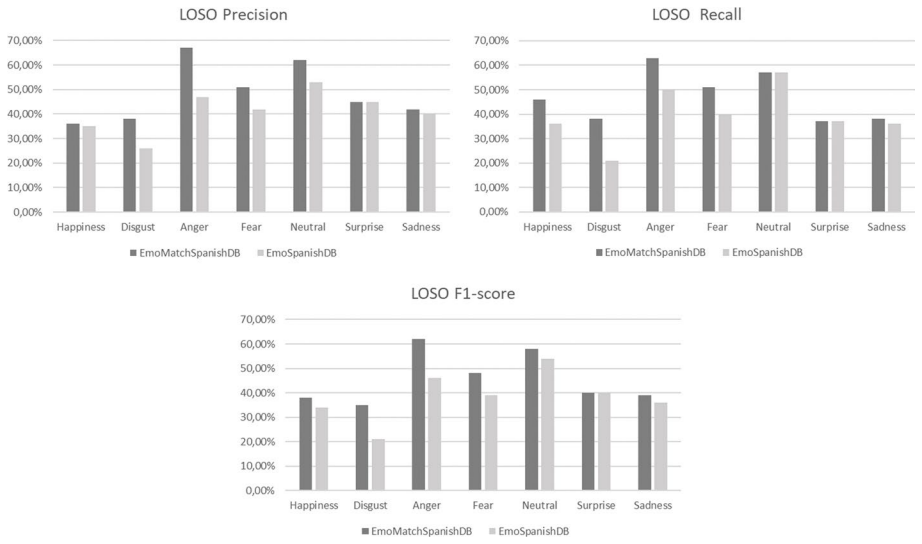

Fig. 6 Precision, Recall, and F1-score LOSO results by emotion for the best model (SVM and 140 features)

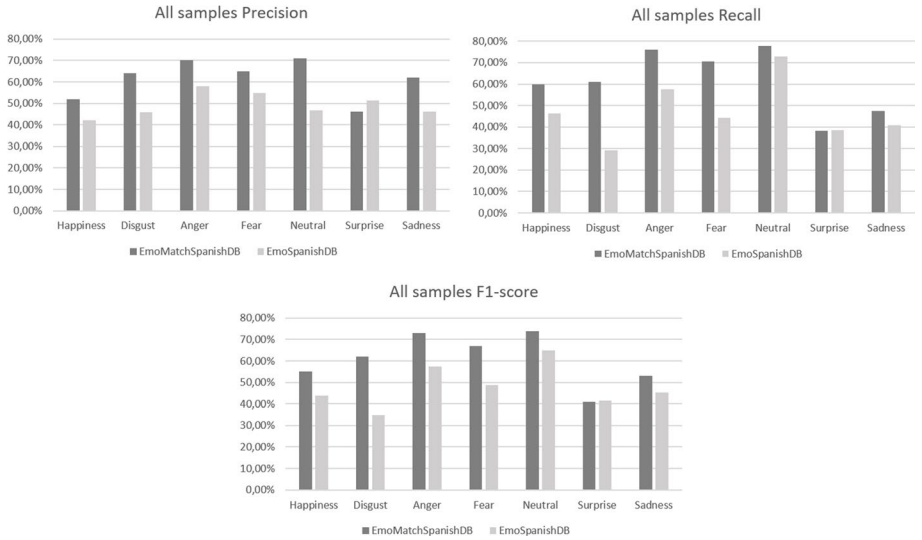

Fig. 7 Precision, Recall, and F1-score All samples results by emotion for the best model (SVM and 140 features)

Analyzing the results on the number of features (13 vs. 140), it is observed that the use of a large number of features results in better F1-score and accuracy and the best results are always obtained with the larger number of features for all cases (see Tables 7, 8, 9 and 10). In order to quantify this improvement we calculated the average percentage for both datasets (all samples and LOSO) being $\approx 37.5\%$.

Overall, the best technique is SVM that obtains the best f1-scores for all the experiments with an average improvement of $\approx$ 16.75% over XGBOOST, and $\approx$ 14% over FFNN. Figures 6 and 7 shows the precision, recall, and F1-score by category for this model and 140 features set. Moreover, the Figs. 8 and 9 show the confusion matrices for all samples and LOSO using the SVM + 140 features as well. The results obtained using

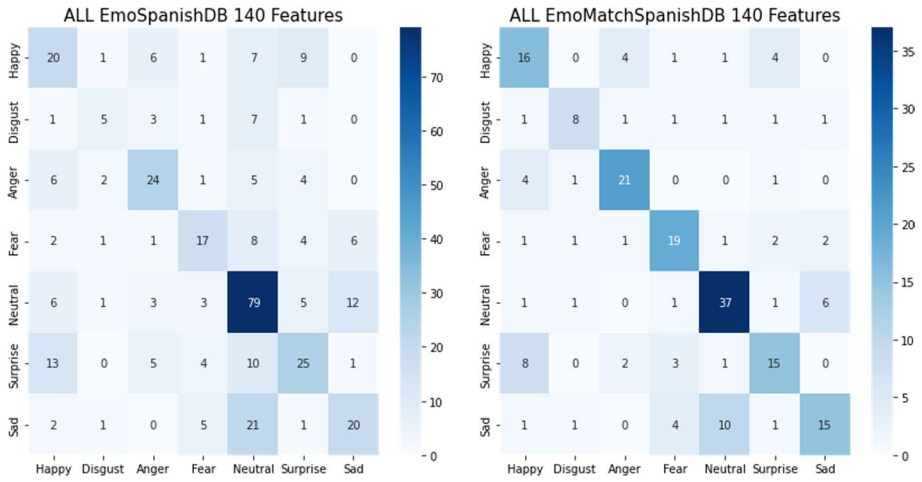

Fig. 8 Confusion matrix for all samples and best model (SVM using 140 features)

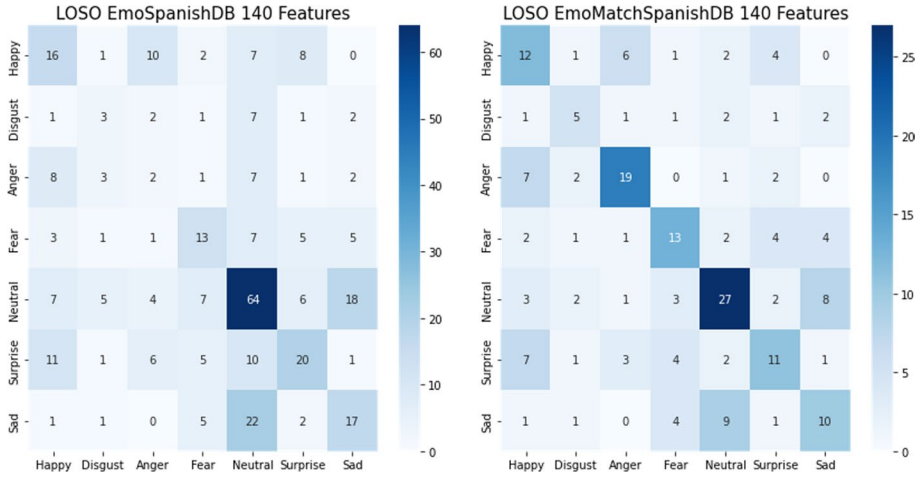

Fig. 9 Confusion matrix for LOSO and best model (SVM using 140 features)

EmoMatchSpanishDB improves the ones of EmoSpanishDB for all the emotions as expected. In all samples experiment there are four emotions (disgust, anger, fear, and neutral) that have a F1-score higher than 60% that is considered to be good from a learning perspective. In the contrary, surprise has the lowest value (41% aprox.) and is mainly confused with happy (notice that happy is also confused with surprise meaning that both emotions share common characteristics). This also was highlighted by the authors at [19]. The same behavior can also be observed in LOSO experiment (Fig. 9). The results of LOSO are significantly worsen than using all samples, meaning that emotions are individual dependent and some pre-processing to remove that personification is needed in case a non individual dependent SER systems wants to be developed. The emotion that most suffers of this dependency is'disgust' that worsens $\approx 44\%$, followed by'happy' with a decrease of $\approx 31\%$ and'fear' with $\approx 28\%$'.

Finally, Tables 11 and 12 show the results obtained using EgeMaps and ComparE features for EmoMatchSpanishDB. It can be observed that the best results are always for SVM model and the use of ComparE features always improves the results over the other set of features (eGeMAPS and, 13 or 140 features presented above. See 7 and 8). The improvement in F1-score is of $\approx 2\%$, $\approx 39\%$, $\approx 11\%$, and $\approx 14\%$ for all samples, LOSO, men, and women experiments respectively. It is also important to note that the best recall for LOSO

Table 11 Unweighted F1-Score and accuracy results obtained for SVM, XGBOOST, and FFNN at EmoMatchSpanishDB all samples and LOSO (eGeMAPS and ComparE features sets)

	All samples						LOSO					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
Nº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
eGeMAPS	0.632	0.654	0.521	0.570	0.460	0.522	0.542	0.573	0.503	0.562	0.511	0.554
ComparE	**0.621**	**0.650**	0.640	0.671	0.612	0.638	**0.589**	**0.642**	0.588	0.642	0.567	0.596

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

Table 12 Unweighted F1-Score and accuracy results obtained for SVM, XGBOOST, and FFNN at EmoMatchSpanishDB women and men (eGeMAPS and ComparE features sets)

	Men						Women					
	SVM		XGBOOST		FFNN		SVM		XGBOOST		FFNN	
Nº Features	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
eGeMAPS	0.598	0.622	0.549	0.621	0.593	0.628	0.573	0.594	0.596	0.632	0.559	0.578
ComparE	**0.683**	**0.702**	0.572	0.608	0.592	0.637	**0.674**	**0.692**	0.581	0.608	0.611	0.661

Bold indicates the number of samples where the elicited emotions match the crowdlabel obtained. The sum of all values in bold adds to the total number of samples of EmoMatchSpanishDB

is 0.59; this value is similar to the state-of-the-art results obtained in similar datasets for other languages (see Tables 11 at [26]).

5 Conclusions

One of the main problems in Speech Emotion Recognition is the absence of public databases. This is very relevant for all languages except English. We created and made public a Spanish Elicited Emotion Dataset consisting of fifty subjects. The generated audio samples were curated using a crowdsourcing approach to avoid discrepancies between the elicited emotions and the emotion recognized by humans. Consensus was obtained using an a priori probability of $1/7$ in a Bernoulli distribution. An average of six labels were needed to complete the process and $\approx 84\%$ of the dataset was classified successfully with an emotion, and 48% were classified according to the original elicited emotion. This dataset is the largest public Spanish dataset as far as the authors know. The EmoMatchSpanishDB dataset has been tested using some of the most successful machine learning models and different set of audio characteristics were also tested in a comparative study. The results show that, using SVM model and the ComparE feature set, up to 65% accuracy can be obtained for six Ekman's emotions plus neutral. Finally, Leave One Speaker Out test was performed and the results show an accuracy of 64.2% and a 0.573 Cohen's Kappa Coefficient. These results are similar to the state-of-the-art of other recently created elicited databases. We envision that other machine learning methods would benefit differently from the release of this dataset and the comparative study presented here provides a good baseline for future improvements and advances in this area for the Spanish Language.

Appendix

Spanish Sentences used for the Creation of the Dataset

The following sentences were used during the creation of the Spanish Emotional Dataset (note: English translation is shown for text legibility):

- S1: El mãdico dijo a tu padre que no tome mãs vino (The doctor told your father not to drink any more wine).

- S2: Donde crece hierba siempre pueden crecer las setas (Where grass grows, mushrooms can always grow).
- S3: La gata gris estaba en la casa (The grey cat was in the house).
- S4: Es mucho mejor si tienen hielo de sobra que no que falte (It is much better to have too much ice than too little).
- S5: Si haces el bestia, es fãcil que te lesiones (If you play hard, it's easy to get injured.).
- S6: El caballo pesado aguantarã subir a los cinco picos del valle, pero un toro no (The heavy horse will endure climbing the five peaks of the valley, but a bull will not.).
- S7: Es muy raro que pise todo el rato (It is very weird that he steps all the time).
- S8: La sala nunca ha estado tan lisa en todo el aã±o (The room has never been so smooth throughout the year).
- S9: Hoy he visto cinco veces a la misma monja (Today I have seen the same nun five times).
- S10: La deuda que tenãan costã unos pisos (The debt they had cost some apartments).
- S11: Cuidado, perro con rabia (Beware, the dog has rabies).
- S12: Mi cordero ha ganado (My lamb has won).

Acknowledgements The authors would like to thank to Emma Rodero for its insightful comments and guidance. We are also specially grateful to all the people that has voluntary participate with this project.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The anonymised data features samples that support the findings of this study are available via a public data repository https://doi.org/10.6084/m9.figshare.14215850.

Declarations

Conflict of interest Partial financial support was received from Universidad Europea de Madrid under the research project APRENDE-R (#2019/UEM60).

References

1. Amer MR, Siddiquie B, Richey C, Divakaran A (2014) Emotion recognition in speech using deep networks. In: ICASSP. Florence, Italy, pp 3752–3756
2. Attwood AS, Easey KE, Dalili MN, Skinner AL, Woods A, Crick L, Ilett E, Penton-Voak IS, Munafó MR (2017) State anxiety and emotional face recognition in healthy volunteers. R Soc Open Sci. 4(5:160855
3. Burkhardt F, Paeschke, Rolfes M, Sendlmeier W, Weiss B (2005) 1129 *A database of German emotional speech*. In: Proc. Interspeech, pp. 1517–1520
4. Byun S, Lee S (2016) Emotion Recognition Using Tone and Tempo Based on Voice for IoT. Trans Korean Inst Electr Eng 65:116–121
5. Calvo RA, D'Mello S (2012) Frontiers of Affect-Aware Learning Technologies. Intell. Syst. IEEE. 27(27):86–89

6. Cao H, Verma R, Nenkova A (2014) Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. Comput Speech Lang.
7. Cavanagh SR, Urry HL, Shin LM (2011) Mood-induced shifts in attentional bias to emotional information predict ill-and well-being. Emotion 11(2):241–248
8. Chang-Hong L, Liao WK, Hsieh WC, Liao WJ, Wang JC (2014) Emotion *identification using extremely low frequency components of speech feature contours*. Hindawi Publishing Corporation. Sci World J. Volume 2014
9. Ekman P (1984) Expression and the nature of emotion. In: Scherer K, Ekman P (eds) Approaches to Emotion. Erlbaum, Hillsdale, NJ, pp 319–344
10. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Truong KP (2015) The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7(2):190–202
11. Florian E, Wöllmer M, Schuller B (2010) *openSMILE - The Munich Versatile and Fast* Open-*Source Audio Feature Extractor*, Proc. ACM Multimedia (MM), ACM, Florence, Italy, pp. 1459–1462
12. Grichkovtsova I, Morel M, Lacheret A (2012) The role of voice quality and prosodic contour in affective speech perception. Speech Comm. 54(3):414–429
13. Iriondo I, Guaus R, Rodriguez A, Lázaro P, Montoya N, Blanco JM, Bernadas D, Oliver JM, Tena D, and Longhi L (2000) *Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques*. In ITRW on speechand emotion, New Castle, Northern Ireland, UK Sept. 2000
14. Izard CE (2010) The many meanings/aspects of emotion: Emotion definitions, functions, activation, and regulation. Emot Rev 2(4):363–370
15. Jadoul Y, Thompson B, de Boer B (2018) Introducing Parselmouth: A Python interface to Praat. J Phon 71:1–15. https://doi.org/10.1016/j.wocn.2018.07.001
16. Jiang D, Lu L, Zhang H, Tao J and Cai L. (2002). *Music type classification by spectral contrast feature*. In Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE Int Conf. vol. 1, pp. 113–116. IEEE, 2002
17. Koolagudi SG, Rao KS (2012) Emotion recognition from speech: a review. Int J Speech Technol 15:99–117
18. Kossaifi J et al (2021) SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. In IEEE Trans Patt Anal Mach Intell. 43(3):1022–1040. https://doi.org/10.1109/TPAMI.2019.2944808
19. Lausen A, Hammerschmidt K (2020) Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. Humanit Soc Sci Commun 7:2. https://doi.org/10.1057/s41599-020-0499-z
20. Madhu N (2009) Note on measures for spectral flatness. Electron Lett 45(23)Confusion matrix for all samples and best model:1195
21. Marchi E, Ringeval F, and Schuller B. (2014) Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody," Speech and Automata in Health Care (Speech Technology and Text Mining in Medicine and Healthcare). De Gruyter, Boston/Berlin/Munich. pp. 207-236
22. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, EB, and Nieto O. (2015) librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, pp. 18–25
23. Mehrabian A (1971) (1971) *Silent Messages*. Wadsworth Publishing Co., Belmont, CA
24. Montoro JM, Gutierrez-Arriola J, Colas J, Enriquez E, and Pardo JM. (1999). *Analysis and modeling of emotional speech in Spanish*. In Proc. int. conf. on phonetic sciences (pp. 957-960)
25. Muhammadi J, Rabiee HR, and Hosseini A. (2013). *Crowd Labeling: a survey*. arXiv: Artificial Intelligence.
26. Parada-Cabaleiro E, Costantini G, Batliner A et al (2020) DEMoS: an Italian emotional speech corpus. Lang Res Eval 54(2):341–383. https://doi.org/10.1007/s10579-019-09450-y
27. Picard R (1997) (1997) *Affective Computing*. The MIT Press, Cambridge
28. Poblet M, Garcia-Cuesta E, Casanovas P (2018) Crowdsourcing roles, methods and tools for data-intensive disaster management. Inf Syst Front 20(6):1363–1379. https://doi.org/10.1007/s10796-017-9734-6
29. Polzehl T, Schmitt A, and Metze Florian. (2010). Approaching multi-lingual emotion recognition from speech - on language dependency of acoustic/prosodic features for anger recognition. In Speech Prosody'2010 Conference, paper 442, Chicago, IL, USA May 10–14
30. Poorna SS, Nair GJ (2019) Multistage classification scheme to enhance speech emotion recognition. Int J Speech Technol. 22(2):327–340

31. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci Special Issue: Cognition and Consiousness in Nonhuman Species. 1(4):515–526
32. Quadflieg S, Wend B, Mohr A, Miltner WH, Straube T (2007) Recognition and evaluation of emotional prosody in individuals with generalized social phobia: A pilot study. Behav Res Ther. 45(12):3096–3103
33. Real Academia Española y Asociación de Academias de la Lengua Española. (2005). *Diccionario panhispánico de dudas*. Madrid: Santillana
34. Rodriguez IA (2016) Cálculo de frecuencias de aparición de fonemas y alófonos en espanol actual utilizando un transcriptor automatico. Loquens 3(1):e029
35. Rozgic V, Ananthakrishnan S, Saleem S, Kumar R, Vembu AN, Prasad R. (2012) Emotion recognition using acoustic and lexical features. In: INTERSPEECH. Portland, USA
36. Sailunaz K, Dhaliwal M, Rokne J, and Alhajj R. (2018) *Emotion detection from text and speech: a survey*. SocNetw Anal Min. 8(1)
37. Scherer KR, Banziger T and Roesch E. (2010). *A Blueprint for Affective Computing: A source book and manual*. Oxford University press.
38. Schuller B, Steidl S, Batliner A, Hirschberg J, Burgoon JK, Baird A, and Evanini K. (2016). *The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language*. In 17TH Ann Conf Int Speech Comm Assoc (Interspeech 2016),. Vols 1–5 (Vol. 8, pp. 2001–2005). ISCA.
39. Schuller B, Wöllmer M, Eyben F, and Rigoll G. (2009) *Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs*. The Role of Prosody in Affective Speech (S. Hancil, ed.), vol. 97 of Linguistic Insights, Studies in Language and Communication, pp. 285-307, Peter Lang Publishing Group
40. Shen P, Changjun Z. and Chen X. (2011) *Automatic Speech Emotion Recognition Using Support Vector Machine*. Int Conf Electr Mech Eng Inf Technol
41. Snow R, O' Connor, Jurafsky D. and Ng A. (2008). *evaluating Non-Expert annotations for natural language tasks*, Proceedings of EMNLP-08.
42. Staugaard SR (2010) Threatening faces and social anxiety: A literature review. Clin Psychol Rev 30(6):669–690
43. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 21(1):93–120
44. Tacconi D, Mayora O, Lukowicz P, Arnrich B, Setz C, Troster G, and Haring C (2008) Activity and emotion recognition to support early diagnosis of psychiatric diseases. In 2008 Second Int Conf Perv Comput Technol Healthcare, pp. 100-102
45. Trémeau F (2006) A review of emotion deficits in schizophrenia. Dialogues Clin Neurosci 8(1):59–70
46. Tseng HH, Huang YL, Chen JT, Liang KY, Lin CC, Chen SH (2017) Facial and prosodic emotion recognition in social anxiety disorder. Cogn Neuropsychiatry. 22(4):331–345
47. Vaidyanathan PP (2008) *The Theory of Linear Prediction*. Chapter 8. California Institute of Technology. Morgan and Claypool Publishers Series
48. Xu Z, Meyer P, Fingscheidt T (2018) "On the Effects of Speaker Gender in Emotion Recognition Training Data," Speech Communication; 13th ITG-Symposium. Oldenburg, Germany, pp 1–5