Are we all good actors? A study on the feasibility of generalizing speech emotion recognition models

Alessandra Grossi[1,2,*], Alberto Milella[1], Riccardo Mattia[1] and Francesca Gasparini[1,2,*]

[1]University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy
[2]NeuroMI, Milan Center for Neuroscience, Piazza dell'Ateneo Nuovo 1, 20126, Milano, Italy

Abstract

Speech is one of the most natural way for human to express emotions. Emotions play also a pivotal role in human-machine interaction allowing a more natural communication between the user and the system. This led in recent years to a growing interest in Speech Emotion Recognition (SER). Most of the SER classification models are based on specific domains and are not easily generalizable to other situations or use cases. For instance, most of the datasets available in the literature are acted utterances collected from English adults and thus not easily generalizable to other languages or ages. In this context, defining a SER system that can be easily generalised to new subjects or languages has become a topic of relevant importance. The main aim of this article is to analyze the challenges and limitations of using acted datasets to define a general model. As preliminary analysis, two different pre-processing and features extraction pipelines have been evaluated for SER models able to recognize emotions from three well-known acted datasets. Then, the model that achieved the best performance was applied to new data collected in more realistic environments. The training dataset is Emozionalmente, a large italian acted dataset collected using a crowdsourced platform from non-professional actors. This model was tested on two subsets of the Italian speech emotion dataset SER_AMPEL, to evaluate the performance in a more realistic context. The first subset comprises audio clips from movies and TV series performed by older adult dubbers, while the second one consists of natural conversations among individuals of different ages. The analysis of performances and results have highlighted the main difficulties and challenges in generalize a model trained on acoustic features to new real data. In particular, this preliminary analysis has shown the limits of using acted dataset to recognize emotion in real environment.

Keywords

Speech emotion recognition, acted dataset, cross-corpus SER, model generalization, cross age SER, acoustic features

1. Introduction

Emotions are a fundamental aspect of social interaction, providing information about individuals feelings, intentions and status [1, 2]. Similarly, systems capable of recognising the user's emotions can modify their behaviour accordingly, defining a more genuine communication between humans and machines [3]. People can express their emotions in different ways including face expression, body gestures, and speech [4]. In particular, the number of technologies interfacing with persons though voice is increased in the last years. Examples include virtual assistants embedded in consumer devices, conversational chat-bots, domotics control systems, and social robots [5, 6]. Furthermore, voice input interfaces have proved to be well accepted by older adults or people with hearing impairments, being easy to use and learn [7]. In this context, the necessity of defining systems capable of interacting naturally with people through speech has led to a growing interest in the field of Speech Emotion Recognition (SER) [8]. In the last decades, extensive research has been conducted on this topic to recognise emotions from speech considering both acoustic and linguistic information [9]. Nevertheless, the majority of SER classification models described in the literature have been developed for specific domains, which limits their applicability to other situations or use cases. These models are typically

Italian Workshop on Artificial Intelligence for Human Machine Interaction (AIxHMI 2024), November 26, 2024, Bolzano, Italy

[*]Corresponding author.

alessandra.grossi@unimib.it (A. Grossi); a.milella3@campus.unimib.it (A. Milella); r.mattia1@campus.unimib.it (R. Mattia); francesca.gasparini@unimib.it (F. Gasparini)
https://mmsp.unimib.it/alessandra-grossi/ (A. Grossi); https://mmsp.unimib.it/francesca-gasparini/ (F. Gasparini)
0000-0003-1308-8497 (A. Grossi); 0000-0002-6279-6660 (F. Gasparini)

trained on acted datasets, collected in laboratory settings from young adult actors, reciting pre-defined utterances simulating different emotions. Furthermore, only a limited number of data corpus include recordings in different languages, while most of them focus on English or Chinese [10]. In recent years, some datasets have been collected considering more real-life conditions or situations. They include data collected from public speeches [11] or call center conversations [12]. Collecting these data is, however, not a straightforward process because of legal issues (copyright and privacy), noisy environmental (background noise, overlapping voices), and difficulties in labeling the data (multiple emotions can occur in a single conversation). In addition, there is no control over the number of instances collected for each emotion, which leads to datasets that are usually unbalanced in terms of class cardinality and utterances length [13]. It is therefore necessary to define a generic SER model that can be easily adapted to different situations and conditions, including multiple languages or ages. To this end, several approaches have been proposed in literature, including traditional machine learning and deep learning strategies [14]. With reference to traditional machine learning, one of the simplest approaches tries to reduce the discrepancy between corpora by training classification models on the combination of different datasets [15]. These classifiers are usually trained using domain-invariant features extracted considering both semi-supervised or unsupervised techniques [13]. Other studies made use of domain adaptation strategies to mitigate the differences in features distribution due to different languages or ages. For this purpose, both unsupervised [16] and supervised [17, 18] approaches have been considered. With regard to deep learning methods, hybrid neural network frameworks have been usually adopted in cross-corpus SER. In particular, the combination of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), autoencoders, and Deep Belief Networks (DBNs) has been employed with the aim of reducing inter-domain variation [19, 14]. Finally, recent SER studies have evaluated the usefulness of feature extraction via self-supervised learning (SSL) models, including Whisper, wav2vec 2.0 or HuBERT [20, 21]. Although the use of these strategies has led to promising results in the development of general models for emotion recognition from speech, this area of research is still an open problem. In fact, many of the classifiers defined in the literature are trained and tested on acted datasets or on datasets collected in controlled environments. Also, the use of deep learning techniques makes it difficult to explain and interpret the results obtained [14]. Finally, analyses often focus on different language datasets and usually do not take into account voice variation due to individuals aging.

The aim of this study is to identify the challenges of defining a general supervised SER classifier trained on an acted data for emotion recognition in natural conversations. In particular, the analysis has been structured in two steps. First, the performance and limits on emotion recognition on acted data were evaluated taking into account three well-known literature datasets. In this phase, two different preprocessing pipelines have been evaluated to remove noise from the signals and to extract features. The results obtained in this preliminary analysis have been then used to define the general model. In particular, the dataset used for this purpose is Emozionalmente, a large Italian emotional speech corpus collected from non-professional actors through a dedicated web-based platform. Compared to the other acted dataset, Emozionalmente contains a large number of data collected from heterogeneous subjects of different ages and genres. In addition, the use of non-professional equipment and non-actors makes the dataset more similar to natural ones, thus including the issues of collecting data in real environments. The classification model defined on Emozionalmente is finally evaluated on a new dataset containing recordings of natural or acted conversations between Italian subjects of different ages. Several considerations will be drawn from the analysis of the results achieved.

2. Datasets

The analysis was carried out considering three well-known acted datasets (RAVDNESS [22], EMOVO [23], and Emozionalmente [24]), as well as a dataset collected under more natural conditions and situations (SER_AMPEL [10]). A detailed description of the four datasets is presented in the following section.

Table 1
Summary of the number of audios for each emotion in the five datasets considered (RAVDESS, EMOVO, Emozionalmente, SER_AMPEL-AOLD, and SER_AMPEL-NYNG-NOLD). The lack of a specific emotional state in a dataset is denoted by the symbol '-'.

	# subjects	anger	disgust	fear	joy	neutral	sadness	surprise	calm
RAVDNESS	24	192	192	192	192	96	192	192	192
EMOVO	6	84	84	84	84	84	84	84	-
Emozionalmente	431	986	986	986	986	986	986	986	-
SER_AMPEL - AOLD	10	34	11	14	20	9	22	10	-
SER_AMPEL - NYNG-NOLD	40	0	0	0	36	11	10	0	-

2.1. RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [22] is a multi-modal acted datasets collected from 24 North America English-speaking actors (12 males, 12 female, mean age: 26 years) in a controlled environment. During the experiment, each participants had to perform 2 spoken and 2 sung utterances simulating eight emotions (happy, sad, angry, fearful, surprised, disgusted, calm, and neutral) in two distinct levels of intensity (normal and strong). The procedure is repeated twice for each subject. The distribution of the 1,440 recordings in the eight classes is reported in the first row of Table 1. All utterances considered in the dataset are in English and semantically neutral.

2.2. EMOVO

The Italian Emotional Speech Database (EMOVO) [23] is a monolingual acted dataset in which 6 young adults actors (3 male, and 3 female) pronounce 14 semantically neutral utterances trying to mimic different emotions. Specifically, each utterance is pronounced considering seven emotional states, including neutral, disgust, fear, anger, joy, surprise, and sadness. The second row of the Table 1 shows the distribution of the audio in the different emotions. All Emovo utterances are in Italian and can be either semantically correct sentences and "nonsense" phrases. Further information on the dataset is available in [23].

2.3. EMOZIONALMENTE

Emozionalmente [24] is amonolingual crowd-sourced emotional speech dataset consisting of 6902 recordings collected by 431 non-professional actors using a web-based platform. Each participant was free to record his/her own voice while performing 18 Italian sentences simulating seven emotional states (neutral, anger, disgust, fear, joy, sadness and surprise). All the data were collected in out-of-laboratory settings using not-professional equipment, thus making them more similar to natural recording conditions. In addition, only Italian mother tongue subjects were included in the dataset while no constraints on the age were considered. In particular, 51 audios were recorded from young people under the age of 18, while 18 instances were recorded from older adults over the age of 60. The sentences selected for the dataset are all semantically neutral and have a maximum length of 10 words. Once the audio clips had been collected, the emotional content of each was validated by five independent evaluators using the platform. The inter-annotator agreement, measured using Krippendorff's alpha, was 0,476, indicating moderate agreement between the annotators. The labelling methodology yielded a balanced distribution of the seven identified emotion classes, with no single class being over-represented. More information about Emozionalmente dataset can be found in [24].

2.4. SER_AMPEL

SER_AMPEL is a multisource dataset focused on data collected from Italian older people and adults that includes both acted conversations, extracted from movies and TV series, and natural conversations, elicited by questions. In particular, two subsets of SER_AMPEL are taken into account in the analysis proposed: the AOLD acted subset and a part of the NYNG-NOLD evoked subset. The AOLD acted subset consists in 120 TV Series and Movies audio clips dubbed by 10 Italian older actors (number of actresses: 4). The recordings, with an average duration of 15 seconds, were labelled by a group of three experimenters considering the six basic emotional states defined by Ekman, as well as the neutral state. The distribution of the audios in the seven classed is reported in the forth row of Table 1.

Differently from AOLD, the NYNG-NOLD evoked subset comprises recordings of natural conversations between older adults or young adults, in which emotional responses are elicited by specific questions or by listening to music. In particular, in this paper we considerthe part of the dataset where memory-related songs chosen by the participants were employed as a hint to evoke emotions during the dialogues. A total of 57 recordings were collected from 40 subjects, including 28 older adults with more that 65 years (16 male, 12 female) and 12 adults with age between 20 and 55 (5 male, 7 female). Similarly to AOLD, each audio was then labeled using the Ekman 7-emotions categorical model. In this case, the most of the instances were labeled as Joy, while only 7, and 12 recordings were labeled as sad and neutral respectively. Table 1 reports in details the cardinality in the seven classes. More information about the SER_AMPEL dataset can be find in the relative article [10].

3. Model definition

The aim of this work is to analyze the challenges of using acted dataset for definition of SER models to be used in different real situations and contexts. The final objective is to define a model that can be integrated into devices with constrained computational resources such as smartphones, smart speakers, or Internet of Things (IoT) devices.

To this end, a first benchmark strategy based on hand crafted features evaluated on the whole utterances (HCW) has been developed, in accordance with the procedure defined in [25]. Then, a second strategy, still based on hand crafted features, but evaluated on segments of each utterance (HCS), has been developed. The adoption of traditional machine learning models that are based on hand crafted features fulfills the requirements of low computational costs and explainability. In this section the two strategies adopted in the analysis are reported in details.

3.1. HCW strategy

In the initial analysis, the audio signals of the three acted datasets Emovo, Ravdess, and Emozionalmente have been preprocessed considering a procedure similar to [25]. As first step, a 5th-order Chebyshev Type I bandpass filter with a lower passband frequency of 100 Hz and a higher passband frequency of 8000 Hz has been applied to the audios to select the frequency range related to human speech. A total of 70 Low Level Descriptors (LLD) have been then extracted from each of the preprocessed utterances, including both prosodic (Pitch, Energy, Zero-Crossing Rate, Spectral Skewness, and Spectral Kurtosis) and spectral features (Mel-Frequency Cepstral Coefficients (MFCCs)). Therefore, for each coefficient, four statistical features have been extracted as the maximum, minimum, mean, and standard deviation of the coefficient values. A summary of the strategy is depicted in Figure 1.

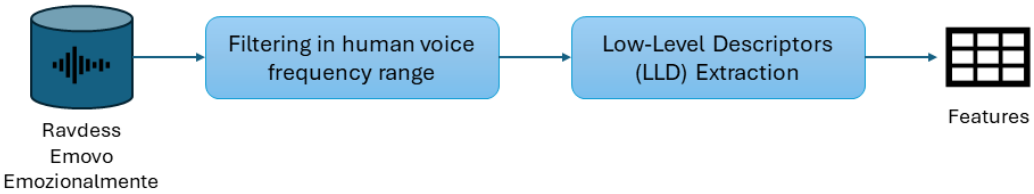
Ravdess
Emovo
Emozionalmente
Filtering in human voice frequency range
Low-Level Descriptors (LLD) Extraction
Features

Figure 1: Pipeline applied to the three acted datasets (Emovo, Ravdess, and Emozionalmente) to extract the features to train the classification models in case of HCW strategy.

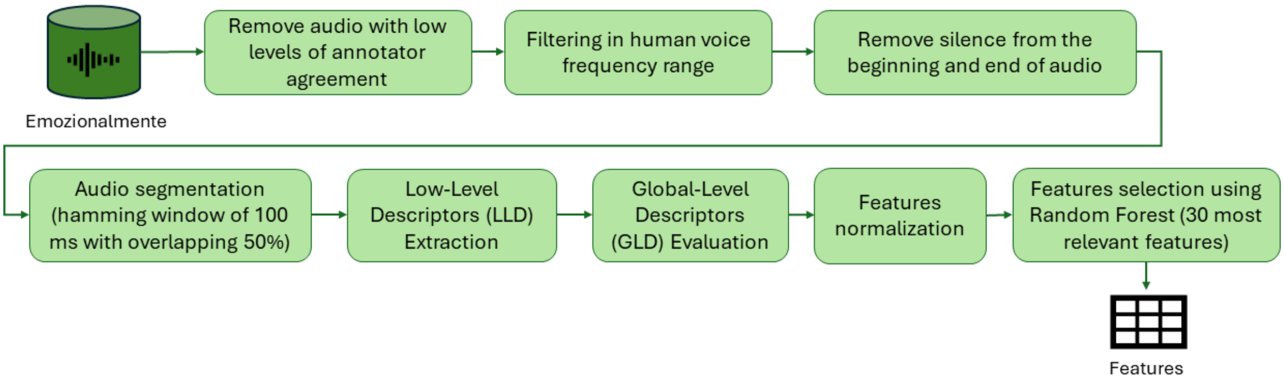
Emozionalmente
Remove audio with low levels of annotator agreement
Filtering in human voice frequency range
Remove silence from the beginning and end of audio
Audio segmentation (hamming window of 100 ms with overlapping 50%)
Low-Level Descriptors (LLD) Extraction
Global-Level Descriptors (GLD) Evaluation
Features normalization
Features selection using Random Forest (30 most relevant features)
Features

Figure 2: Pipeline applied to the acted dataset Emozionalmente to extract the features in the HCS strategy.

3.2. HCS strategy

A second pipeline has been developed to consider the variation of the speech signal over time. This model has been trained only on Emozionalemente, as this dataset could be considered more similar to natural conversation, than the other two datasets, and thus more appropriate to classify the SER_AMPEL speeches. The initial stage of the process involved the selection of a subset of the audio data from the dataset, in order to keep only the recordings with higher values of agreement between the annotators. In particular, the audios selected for the analysis are the ones where at least the 60% of the annotators where able to correctly recognize the emotion performed. This led to a reduction in the overall number of recordings from 6902 to 2253. Then to isolate only the frequencies associated with the human voice, the audio clips were filtered using a Butterworth band pass filter of 4th order with cut-off frequencies of 300 Hz and 8000 Hz. The silence at the beginning and end of the signals were trimmed using a 5% energy threshold. Finally, to take into account differences in utterance length, each audio was segmented into 100 ms frames with a 50% overlap, using a Hamming window. From each frame, 43 low-level descriptors (LLD) has been extracted, including Root Mean Square (RMS), Zero Crossing Rate (ZCR), Energy, Pitch, and 13-Mel-Frequency Cepstral Coefficients (MFCCs). These features have been aggregated by computing the mean, standard deviation, minimum, and maximum, resulting a total of 171 global level descriptors for each recording. The features were finally normalized using min-max scaling to the range $[0, 1]$. Finally, to reduce features dimensionality, a Random Forest with 10-fold cross-validation was employed to select the 30 most important features based on their contribution to classification accuracy. The whole procedure is shown in the framework of Figure 2.

4. Results and Discussion

The acoustic features extracted by the two pipelines (HCW and HCS) were used as input to train and test different classification models, including support vector machine (SVM), gradient-boosted decision tree (XGBoost), decision trees and linear discriminant analysis (LDA). For each model, the hyperparameters were optimized using a cross-validation strategy. According to the dataset considered, seven different emotional states were taken into account in the model definition.

In particular, for Emovo and Emozionalmente, the classifiers were trained to recognize the six basic Ekman emotions and the neutral state. Note that the analysis on Emozionalmente was carried out considering as labels the emotions that the actors wanted to convey (intended emotions), instead of the emotional states identified by the annotators (perceived emotions)[26], to be coherent with the labels of the other datasets. For Ravdess, the emotion “calm” was included instead of neutrality in order to keep balanced the classes as number of instances.

Two evaluation strategies were investigated for the analysis: a 5-fold cross-validation strategy, where data from the same subject can be included in both the training and test sets, and a subject-independent

Table 2
Best performances achieved on the three acted datasets (Emovo, Ravdness, and Emozionalmente), varying the classification strategy applied (HCW or HCS) and the evaluation methods adopted (Cross Validation and Subject Independent). Details about the dataset considered are reported in the first part of the table. The last rows show the results yielded by the two evaluation methods in terms of Macro F1-scores, accompanied by information on the classification model, and number of instances considered to achieve them.

Dataset	EMOVO	RAVDNESS	Emozionalmente	
No. of subjects	6	24	431	
Type of dataset	acted	acted	acted	
Language	italian	english	italian	
Subject ages	young adults	young adults	from very young to older adults	
Type of actors	professional actors	professional actors	not-professional actors	
Emotional states	anger, disgust, fear, neutral, joy, sadness, surprise	anger, disgust, fear, calm, joy, sadness, surprise	anger, disgust, fear, neutral, joy, sadness, surprise	
Classification strategy	HCW	HCW	HCW	HCS
No. of instances	588	1344	6902	2253
Classifier	SVM	SVM	SVM	XGBoost
Macro F1-score cross validation	69%	68%	38%	46%
Macro F1-score subj. independent	21%	47%	25%	39%

cross-validation strategy, where training and testing of the models are performed on different subjects. The performances of the classifiers were compared using the macro F1-Score evaluation metric, which is computed from the total confusion matrix. For the sake of analysis, only the best results obtained are reported in Table 2.

4.1. Analysis of the HCW strategy

The SVM classifier achieves the highest performance in almost all experiments which are based on the HCW strategy. In particular, Macro F1-score values near to 69% have been obtained for the recognition of emotions in Emovo and Ravdess when the cross-validation strategy is applied, while a recognition of only 38% is achieved when the same procedure is used to analyze the data of Emozionalmente. This difference in performance may be due to the distinctive characteristics of Emozionalmente. In fact, the dataset was collected in an uncontrolled environment by amateur actors using a non-professional acquisition device, thus sensitive to noise and artefacts. Furthermore, not always the participants were good in expressing emotions. In [24] is reported an overall accuracy of 66% reached by human evaluators in recognizing the emotions expressed in the audios of Emozionalmente. This highlights the challenges in recognizing emotions from data collected by not-professional actors, putting a limit to the performance that can be achieved by the classifiers when recordings with low level of annotator agreement are included in the dataset analyzed.

In addition, the results reported in Table 2 show that the application of 5-folds Cross Validation method is not appropriate for evaluating the capability of the classifiers to generalise to novel data. In fact, the use of the subject independent evaluation strategy decreases significantly the general performances of all the analysis carried out. The drop in performances is related to the number of subjects. In the dataset Emovo (with the lowest number of actors, i.e. 6) the model decreases its performance from 69% to only 21%. On the other hand, the heterogeneity and the high number of participants in Emozionalmente made it possible to reduce the relative drop in the overall performance.

4.2. Analysis of the HCS strategy

The last column of Table 2 reports the values of Macro F1-score obtained on Emozionalmente when the HCS strategy is applied. The application of global features extracted from the audio segments, the reduction of the dataset by considering only recordings with high agreement values and the use of the XGBoost classifier led to an increase in the Macro F1-score value from 38% to 46% for 5-fold cross-validation and from 25% to 39% for subject-independent evaluation strategy. However, it is worth noting that the subset of selected audios includes all recordings with a higher agreement between the intended and perceived emotion.

4.3. Testing on realistic conversations: the SER_AMPEL dataset

As further experiment, the XGBoost classifier trained on Emozionalmente using the HCS strategy has been also applied to data collected in the SER_AMPEL dataset. The idea was to assess whether the generalising ability of the model in emotion recognition remains even when applied to new data collected in more realistic situations. In Tables 3 and 4 the confusion matrices resulting from the application of the classifier to the data of SER_AMPEL-AOLD and SER_AMPEL-NYNG-NOLD are reported. An overall Macro F1-score value of 18% have been achieved by the classifier when applied to the data collected from senior dubbers, while a recognition rate of 26% have been reached by the model when applied to natural conversations among young and older adults.

In particular, in the AOLD dataset, most of the audios are misclassified as "surprise" or "fear", probably due to the emphasis given by the dubbers on performing the emotions. To this end, it is important to take into account that the almost all the audios are taken from sitcoms or films in which the acting is strongly emphasised in the emotional states of the characters. Similarly, the classifier completely fails in recognizing the "anger" emotion, usually predicting it as "sadness", "surprise" or "fear". In the dataset, this emotion is acted out in different ways and with different levels of intensity, thus making it difficult for the algorithm to correctly identify it. Finally, also the "neural" state is usually misclassified. This can be due to the tone of voice and the cadence used by the dubbers during the performing, which has also led the experimenters to identify varying emotional states in the audio recordings.

The "neutral" state is instead better recognized in the SER_AMPEL-NYNG-NOLD dataset with a recall of 55%. However, it should be noted that the precision value for this class is not high (18%). A possible explanation concern the fact that most of the subjects in the dataset are older adults. With aging the larynx and vocal cords change, thus modifying the characteristic of the voice (variation in pitch, reduced volume, voice tremors, and increased number of silence). This characteristic may have negatively influenced the features extracted, led the classifier to wrongly classify specific emotions, like "joy" or "sadness", as "neutral" state. Similarly, some audios were also erroneously classified as "sad" instead

Table 3

Confusion Matrix resulting from the application of the classifier trained on Emozionalmente to the data of SER_AMPEL-AOLD. Tha matrix is normalized over the true (rows) conditions. The values corresponding to the correct predictions are highlighted in bold. The overall Macro F1-score value achieve is 18%.

		Predicted							
		anger	**disgust**	**fear**	**joy**	**neutrality**	**sadness**	**surprise**	**Recall**
Actual	**anger**	**0,00**	0,00	0,47	0,09	0,00	0,15	0,29	0,00
	disgust	0,00	**0,27**	0,64	0,00	0,00	0,09	0,00	0,27
	fear	0,00	0,07	**0,64**	0,00	0,00	0,07	0,21	0,64
	joy	0,00	0,10	0,40	**0,10**	0,05	0,15	0,20	0,10
	neutrality	0,00	0,00	0,33	0,00	**0,11**	0,22	0,33	0,11
	sadness	0,00	0,05	0,36	0,00	0,09	**0,27**	0,23	0,27
	surprise	0,00	0,00	0,40	0,00	0,00	0,40	**0,20**	0,20
	Precision	Nan	0,43	0,16	0,40	0,25	0,27	0,07	

Table 4

Confusion Matrix resulting from the application of the classifier trained on Emozionalmente to the data of SER_AMPEL-NYNG-NOLD. The overall Macro F1-score value achieve is 26%. The matrix is normalized over the true (rows) conditions. The values corresponding to the correct predictions are highlighted in bold. Given the restricted number of emotional categories included in the dataset, only the rows corresponding to the available classes are displayed in the table.

		Predicted							
		anger	**disgust**	**fear**	**joy**	**neutrality**	**sadness**	**surprise**	**Recall**
Actual	**joy**	0,00	0,06	0,00	**0,06**	0,64	0,25	0,00	0,06
	neutrality	0,00	0,00	0,09	0,00	**0,55**	0,36	0,00	0,55
	sadness	0,00	0,00	0,00	0,00	0,40	**0,60**	0,00	0,60
	Precision	Nan	0,00	0,00	1,00	0,18	0,32	Nan	

of "neutral" or "joy". Furthermore, the audios were collected during conversations with an average duration of 15 seconds. Listening to these audios revealed that the emotions are not static but vary during the dialogue, thus making difficult to identify a single emotional state distinctive of the whole recording.

5. Conclusions

There are several challenges still open in the task of emotion recognition from speech. In particular, the lack of properly labeled data available to train classification models is a primary issue that has to be taken into account. The current work has proved that the use of acted dataset can only partially solve this problem. Although they allow to have data balanced in the number of elements for each class, not always the emotions intended by the speaker are correctly recognized by the listeners. In particular, when considering only audio speeches where there is a high agreement among the perceived and intended emotions the performance of the classifiers are higher.

Furthermore, the number of speakers included in the datasets influences the generalization capability of the trained model. This can be quantitatively evaluated only when a subject independent validation strategy is applied, suggesting to avoid the use of traditional k-fold strategies to provide a fair estimation of the model performance.

These classifiers seems to struggle when applied to real audios collected in non controlled environments characterised by the presence of noise and heterogeneous subjects. In particular, the acted datasets available in literature do not seem to be suitable for recognizing the voice characteristics of older adults, as well as their way of expressing emotions. Finally, audios collected in real situations have different time duration, often including varying emotional states in a single conversation. This makes it difficult to identify a unique emotion representative of the whole recordings using features extracted from the entire audio. The employ of large acted datasets collected in conditions more similar to real life, as in the case of Emozionalmente, seems to reduce some of these issues. However, it does not resolve them completely.

In future analysis, different set of features and models will be evaluated to try to reduce the dependence on the subject, considering for instance subject-independent features or domain adaptation models. The emotional variation over time during conversations necessitates a finer analysis of the recordings, e.g. by studying the evolution of emotions in small audio segments or by including models that take into account temporal information. Also a finer analysis in the frequency domain, such as considering wavelet transform or other temporal-frequency representation can underline which frequency bands are more significant with respect to each emotion. Approaches based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), could provide features that are more representative of emotions across different environments and speakers, potentially improving generalization.

Furthermore, in the current analysis only the acoustic information of the audio is taken into account.

However, different literature works have highlighted the significance of the words used during a speech to express emotions. The use of acted datasets, based on fixed and pre-selected utterances, does not allow to perform a proper analysis of the textual content of a speech. To this end, future studies will be conducted to assess the potential and challenges of defining multi-modal SER models based on both acoustic and linguistic components.

Acknowledgment

This study work was funded by the European Union – Next Generation EU programme, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-It (Ageing Well in an Ageing Society). This work was also partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] T. Johnstone, K. R. Scherer, Vocal communication of emotion, Handbook of emotions 2 (2000) 220–235.

[2] H. Nordström, Emotional communication in the human voice, Ph.D. thesis, Department of Psychology, Stockholm University, 2019.

[3] Y. Chavhan, M. Dhore, P. Yesaware, Speech emotion recognition using support vector machine, International Journal of Computer Applications 1 (2010) 6–9.

[4] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th international conference on Multimodal interfaces, 2004, pp. 205–211.

[5] W. Alsabhan, Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention, Sensors 23 (2023) 1386.

[6] M. Spezialetti, G. Placidi, S. Rossi, Emotion recognition for human-robot interaction: Recent advances and future perspectives, Frontiers in Robotics and AI 7 (2020) 532279.

[7] F. Portet, M. Vacher, C. Golanski, C. Roux, B. Meillon, Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects, Personal and Ubiquitous Computing 17 (2013) 127–144.

[8] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, C. Cleder, Automatic speech emotion recognition using machine learning, Social Media and Machine Learning [Working Title] (2019).

[9] B. T. Atmaja, A. Sasou, M. Akagi, Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion, Speech Communication 140 (2022) 11–28.

[10] A. Grossi, F. Gasparini, Ser_ampel: A multi-source dataset for speech emotion recognition of italian older adults, in: Italian Forum of Ambient Assisted Living, Springer, 2023, pp. 70–79.

[11] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, U. Lee, K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, Scientific Data 7 (2020) 293.

[12] D. Morrison, R. Wang, L. C. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech communication 49 (2007) 98–112.

[13] M. S. Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digital signal processing 110 (2021) 102951.

[14] S. Zhang, R. Liu, X. Tao, X. Zhao, Deep cross-corpus speech emotion recognition: Recent advances and perspectives, Frontiers in neurorobotics 15 (2021) 784514.

[15] F. Eyben, A. Batliner, B. Schuller, D. Seppi, S. Steidl, Cross-corpus classification of realistic emotions–some pilot experiments (2010).

[16] N. Liu, Y. Zong, B. Zhang, L. Liu, J. Chen, G. Zhao, J. Zhu, Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5144–5148.

[17] P. Song, S. Ou, W. Zheng, Y. Jin, L. Zhao, Speech emotion recognition using transfer non-negative matrix factorization, in: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 5180–5184.

[18] F. Gasparini, A. Grossi, Sentiment recognition of italian elderly through domain adaptation on cross-corpus speech dataset, arXiv preprint arXiv:2211.07307 (2022).

[19] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Transfer learning for improving speech emotion classification accuracy, arXiv preprint arXiv:1801.06353 (2018).

[20] L. Pepino, P. Riera, L. Ferrer, Emotion recognition from speech using wav2vec 2.0 embeddings, arXiv preprint arXiv:2104.03502 (2021).

[21] M. Osman, D. Z. Kaplan, T. Nadeem, Ser evals: In-domain and out-of-domain benchmarking for speech emotion recognition, arXiv preprint arXiv:2408.07851 (2024).

[22] S. R. Livingstone, 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 13 (2018) e0196391.

[23] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, et al., Emovo corpus: an italian emotional speech database, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[24] F. Catania, Speech emotion recognition in italian using wav2vec 2, Authorea Preprints (2023).

[25] A. Koduru, H. B. Valiveti, A. K. Budati, Feature extraction algorithms to improve the speech emotion recognition rate, International Journal of Speech Technology 23 (2020) 45–55.

[26] L. Turchet, J. Pauwels, Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021) 305–316.