

A computational framework for speech emotion recognition in case of multisource data

Alessandra Grossi^{1,2,*}, Giorgio Fratti¹ and Francesca Gasparini^{1,2}

¹*Department of Informatics, Systems and Communication, University of Milano-Bicocca, Building U14, Viale Sarca 336, 20126 Milano, Italy, <https://mmsp.unimib.it/>*

²*NeuroMI, Milan Center for Neuroscience, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy*

Abstract

Although several researches have been carried out in the field of Speech Emotion Recognition (SER), only few of them consider people of different ages or languages. In particular, most of the SER datasets reported in the literature are collected from young adults or take into account a single language, such as English or Chinese. These datasets tend to be poorly heterogeneous and dependent on the context in which they are collected. In general they are composed of acted utterances or they are recorded in situations properly designed to evoke certain emotions. This paper proposes a framework that allows to benefit of complementary information coming from multisource data to train a general SER model. To merge different sources, proper preprocessing steps to normalize the data source, the type of recorded speeches, and the subjects considered are here described. Furthermore we present a domain adaptation strategy that allows to benefit of the general model adapting it to a certain language and/or a certain population age. In particular here we are interested in developing SER models that consider Italian older adults. Preliminary results that consider several sources for training and different language as test set confirm the validity of the proposal.

Keywords

speech emotion recognition, multisource, older adults, domain adaptation, XGboost,

1. Introduction


With the increasing of life expectancy, the promotion of positive psychological well-being of older adults is becoming a primary need. Many older adults live alone in their own homes [1], usually isolated because of health problems or major life events that threaten to limit their social interaction [2]. The negative impact of this isolation on mental and physical health leads to the need to develop systems that can monitor [3] and interact naturally with older adults during their daily lives. In particular, Social Robots, as Companion Type Robots, are being developed specifically to provide companionship and cognitive support to frail people [4] to ensure their health and psychological well-being [5]. Such robots must be able to interact with people in a


AIxAS 2023: Fourth Italian Workshop on Artificial Intelligence for an Ageing Society, 6–9 November 2023, Rome, Italy


*Corresponding author.

†These authors contributed equally.

✉ alessandra.grossi@unimib.it (A. Grossi); giorgio.fratti@unimib.it (G. Fratti); francesca.gasparini@unimib.it (F. Gasparini)

ORCID  0000-0003-1308-8497 (A. Grossi); 0000-0002-6279-6660 (F. Gasparini)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

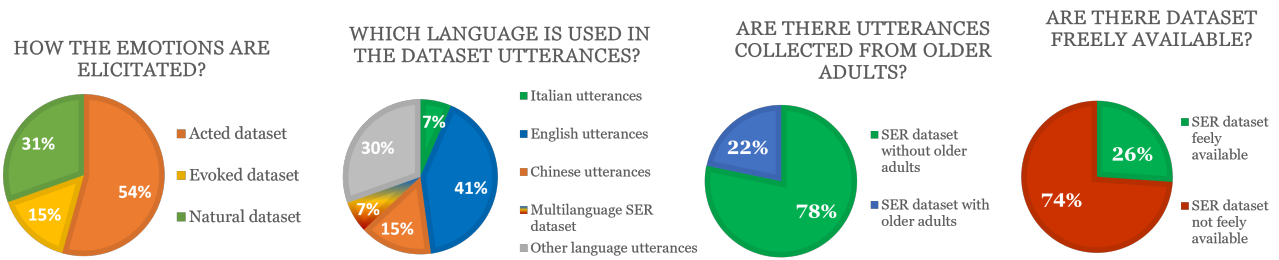


Figure 1: Pie charts depicting different characteristics of the 48 SER datasets analyzed.

natural and realistic way, inferring their emotions and adapting their behaviour accordingly. Similarly, conversational agents were proposed in healthcare domain as mean to help people that live alone or suffer of mental illnesses like depression or anxiety disorder. Examples include voice chatbots like Charlie [6]. These bots aim to provide empathetic support to elderly people and encourage conversations that simulate human interaction. Furthermore, a system that can detect emotions from speech could be incorporated in automated call centres [7] or toll-free helplines for older adults (like the Silver Line) to identify emergency situations, vulnerabilities or social isolation, and take appropriate action. Language and speech are one of the most natural method of communication between humans, and different emotional information can be drawn from the acoustic characteristics of speaker’s voice. Speech Emotion Recognition (SER) is the task of recognizing the speaker’s emotion through the processing and classification of his/her speech signal [8].

Several datasets exist in the literature that try to face the problem of SER, as previously deeply described in [9], where 48 different datasets have been analyzed and summarized in the pie charts depicted in Figure 1. An excel file that synthesizes the whole analysis of these datasets is reported as supplementary material available at the following link <https://mmsp.unimib.it/download-1/>. These datasets have been acquired in different ways that can be grouped as: i) acted, ii) evoked, and iii) natural conversation [10]. Besides the huge number of datasets considered only few of them are available and moreover their characteristics are significantly different to be merged directly in order to provide a huge dataset to train a SER model. These differences are related to: emotional space, language, age, type of collection, devices adopted among others. We here propose a framework intended to normalize these datasets with respect to their different characteristics, in order to benefit of a consistend amount of speech data able to train a general SER model. From this general model a successive adaptation step allows to apply the domain adapted model to a specific SER application, for instance in case of a particular population in terms of language and age.

From here on, the manuscript is organised as follows. Section 2 provides a brief description of the main stages involved in creating a SER computational system that combines different datasets into a heterogeneous, unified multisource dataset. Section 3 presents our proposed normalisation framework, mainly focused on standardising data from different acted datasets. The effectiveness of the proposed framework is evaluated in the 4 section by comparing different classification models. Finally, next steps and conclusions are outlined in the last section.

2. A multisource data SER computational framework

To benefit of complementary information coming from different datasets acquired under different experimental conditions, and using different devices, it is necessary to define a framework that faces all the related issues.

The multisource framework for Speech Emotion Recognition here proposed is depicted in Figure 2, and is composed of four main modules related to defining the datasets to be used, defining the proper signal processing steps to align and normalize the different data sources, select and train an appropriate machine learning model, and finally define a proper strategy to adapt the general model developed to a specific population with respect to age and language. Each single module is described in what follows.

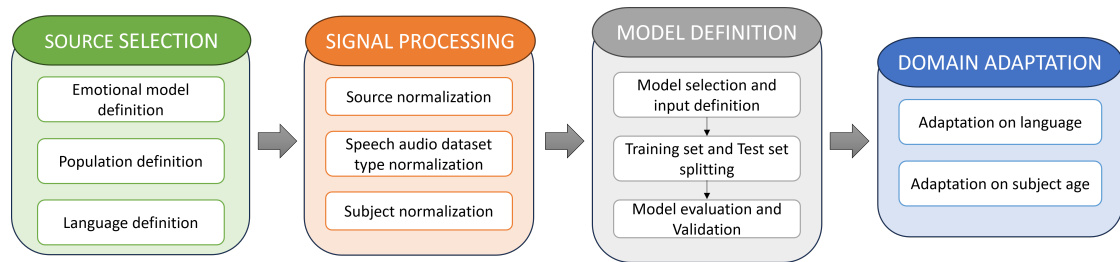


Figure 2: The multisource framework for Speech Emotion Recognition

2.1. Source selection

The first step in defining a unique multisource heterogeneous dataset for training a general SER model is to select the individual sources to be included in it. In this phase the characteristics of each dataset should be taken into account.

As already reported in the introduction, SER datasets vary according to the way in which the emotions are elicited. In particular, they could be divided into acted, evoked, and spontaneous or natural conversation datasets. Furthermore, different data could be labelled with different emotions depending on the emotional model chosen. Two types of model are usually involved in SER analysis: the discrete or categorical emotional models, which include the 6-basic emotions defined by Ekman and Friesen [11] or Plutchik's Wheel of Emotions [12], and the dimensional or continuous emotional models, such as the 3D Valence-Arousal-Dominance space [13]. When considering the union of different datasets, the use of similar types of emotional models makes it possible to simplify the merging between data, thus avoiding the imbalance problems due to the definition of a mapping between categorical and continuous spaces [14].

Also the characteristics of the speakers can influence the choice of the datasets to include in the general model. Most of the available data are collected from young adults and considering a single language, usually English. However, the human voice changes with age and subject's gender [15]. In addition, personality traits or cultural aspects, such as language, may influence the way a person expresses his/her emotions. In particular, the presence of dialects has to be taken into account in the definition of a model that can be applied to different contexts or

populations.

The choice of which datasets to use to define the multisource heterogeneous corpus is therefore relevant and must consider all these aspects. In this context, the use of similar datasets may simplify the integration process, but may result in reduced data variability.

2.2. Signal Processing

A signal processing step should then be applied to the selected signals to minimize discrepancies resulting from the diverse nature of the data. The choice of which processing techniques apply to each data depends on the type of dataset considered, as well as the population and device used to record the audio signals.

The characteristics of the recording devices can affect some of the audio features, deteriorating their quality or reliability [16]. Audio signals may be heterogeneous in terms of number of channels (e.g. mono or stereo), volume, and frequency resolution when recorded using different devices, such as high quality microphones or consumer-grade ones. According to the literature [17], the volume adjustment can be performed standardizing the signals by z-score normalization. This allows to normalize the audio in terms of volume, but also with reference of the characteristics of the subject. In addition, a Sample Rate Conversion can be applied to the raw signals of each dataset in order to define a single temporal resolution suitable for all the data collected. This operation could be performed considering the minimum sampling rate in the original datasets. Finally, to standardize the data according to the number of channels, the stereo audio can be converted into a mono signal by selecting only one of the channels or by averaging, for each sample, the data of the two channels.

The recording environment and the method used to elicit the emotions can also create a mismatch between the signals of different source, thus affecting the preprocessing step. For instance, Fahad et al. [18] report some issues due to the use of speech audio collected during natural environmental conversations, such as the presence of background noise, multiple voices, long period of silence or utterances with different length. Various denoising techniques have been proposed in literature to minimize the data mismatch due to these issues. In particular, noise reduction methods based on filters, estimators or spectral subtraction are usually applied in particular to reduce background noise [8]. Moreover, as further speech enhancement techniques, the audio signal can be filtered using band pass filters or first-class FIR high-pass digital filters [19] in order to select the frequency range related to human voice. Finally, three strategies have been proposed to overcome the problem related to utterances with different length [18]: i) computing global features as summary statistics of local features extracted on the frames in which audio signal is splitted by applying a fixed size sliding window; ii) using padding strategy to standardize the signals in length; iii) dividing audio signal into segments of fixed length. Concerning this latter, short utterances of 0.5 - 1.00 second are preferred [20] to longer utterances as they allow to extract significant features while maintaining the quasi-stationary state of the speech signal.

The final factor to consider in pre-processing is the speakers heterogeneity. Speech audio signals are subjective and vary according to personal characteristics, such as age, gender or vocal tract length of the speaker. Such differences make it necessary to apply subject-based normalization to the audio signals and it becomes mandatory in the case of multisource datasets. Two strategies

have been investigated in previous studies, applied to each subject data: i) the standardization or range normalization applied to the whole signal sequence; ii) the neutral-based normalization, where the parameters are obtained on baseline or neutral signal and then applied to the rest of the audios of the same subject [17].

2.3. Model definition

Once pre-processed, the audio signals of the multisource dataset can be used for the definition of the general model. Several algorithms for speech emotion recognition have been proposed in the literature, including traditional machine learning techniques as well as deep learning approaches. Based on the model chosen, the necessary input must be obtained from the audio signals in the form of feature vectors, images or raw signals.

In particular, in the case of traditional machine learning techniques, four types of acoustic features can be extracted from speech audio signals: prosodic features, like rhythm and intonation, spectral features, voice quality features and Teager Energy Operator (TEO) Based Features [21]. Some of these features are subject independent, such as the Ratio of Spectral Flatness to spectral center (RSS) [22] or the features based on weighted bi-spectrum [23], while others have to be normalized to take into account differences between datasets.

Several deep learning algorithms, such as Convolutional Neural Network, need image as input. Time-frequency representation of the audio signals, including spectrograms or scalograms, are usually employed for this purpose. However, several issues have to be taken into account during this conversion. In particular, the length of the signals and the image range have to be homogeneous to make the data comparable.

To train and validate the general model, several evaluation strategies have been proposed in literature. In case of speech, the use of traditional techniques as hold-out cross validation or k-fold cross validation can lead to biased results. In fact, audio signals from the same subjects or related to similar utterances may appear in both training and test set, thus making the model biased on this type of data. Evaluation strategies such as Leave One Subject out (LOSO), Leave One Utterance Out (LOUO) or Subject independent k-fold cross validation are mandatory in case of multisource dataset analysis. Finally, the emotional model selected can affect also the evaluation metrics used to assess the classifier. Metrics such as the per-class f1-score or macro f1-score have to be preferred in case of multi-class classification as they take into account the issues due to unbalance among the different classes.

Concerning this latter, the use of different datasets can lead to the definition of classes not balanced in number of instances. Data augmentation strategies or undersampling methods have been proposed in the state of art as solution for this problem. However, the creation of synthetic data, as well as the random selection of subset, could affect the performances of the classifier reducing the generalizability of the model or adding bias due to the similarity between the original and the new data.

2.4. Domain Adaptation

The use of multisource dataset in the definition of the SER classifier allows to reduce the generalization error, creating models able to capture meaningful patterns of the speech data. However,

these models are not always proper for describing scenarios characterized by few, unlabeled data, as occurs with certain languages or age groups. Recent studies [24, 14] have presented multiple transfer-learning methodologies that reuse knowledge acquired from differing but correlated tasks (source domain) to enhance recognition accuracy for a novel task (target domain). Several approaches presented in literature have aimed to enhance deep learning SER models performance by fine-tuning pre-trained networks, primarily based on images, using speech data gathered from specific domains[25]. Furthermore, pre-trained networks can also be employed as features extraction methods as outlined in [26, 27]. Finally, feature-based domain adaptation strategies have also been tested by researchers to adapt pre-trained machine learning classifiers to new labelled data, as reported in [14, 28].

3. Experiments on the proposed framework

In the following analysis, two aspects of the proposed framework are considered: i) the benefits obtained from a multisource approach to build a general model for SER, addressing all the normalization aspects required, and ii) the adoption of a domain adaptation strategy to fine tune a general model for a different specific scenario.

Due to the lack of available datasets, especially containing a significant number of Italian elderly, for this preliminary study some assumptions have been made:

- Only acted datasets have been taken into account for the experiments.
- Ekman's universal emotions, including angry, fear, disgust, surprise, happy, sad, and neutral, have been selected as emotional model.
- Only domain adaptation with respect to language is here considered.

Based on these assumptions, four acted datasets have been involved to test the performance of the general model: analysis: CREMA_D, RAVDESS, SAVEE, and EMOVO. The Crowd-sourced Emotional Multimodal Actors Dataset (*CREMA_D*) [29] and the Ryerson Audio-Visual Database of Emotional Speech and Song (*RAVDESS*) [30] are two multi-modal acted datasets collected in a controlled environment. In the *CREMA_D* dataset, 96 professional actors (48 male and 43 female) performed 12 semantically neutral phrases while simulating six distinct emotions (Anger, Disgust, Fear, Happiness, Neutral, and Sadness). Participants of different ethnic and ages were involved in the dataset, including 6 older adults. Similarly, the *RAVDNESS* dataset consists in 7,356 recordings obtained from 24 young adult actors (12 males, 12 females) with a mean age of 26 years. Each participant performed 60 spoken utterances and 44 sung utterances, expressing eight emotions (happy, sad, angry, fearful, surprised, disgusted, calm, and neutral) under two distinct levels of intensity (normal and strong). In both the datasets, the utterances are in English and the audio signals were recorded using a sampling frequency of 48 kHz.

The Surrey Audio-Visual Expressed Emotion (*SAVEE*) database [31] is an acted dataset where 4 male actors, aged from 27 to 31 years, pronounce 15 English phonetically-balanced TIMIT sentences in seven different emotions, including the six Ekman universal emotions as well as the neutral state. Unlike *CREMA_D* and *RAVDNESS*, in *SAVEE* only a subset of the utterances is common to all the emotions, while most of them change according to the emotional state expressed. The audios are recorded using a sampling rate of 44.1 kHz.

Finally, a SER italian dataset have been involved into the analysis to evaluate the performance of a general model when adapted to recognize emotions from a specific population or scenario. The Italian Emotional Speech Database (EMOVO) is an acted dataset that includes 588 speech audios collected from 6 professional actors (3 male and 3 female) while they were playing 14 sentences mimicking 6 different emotions (disgust, anger, fear, surprise, joy, and sadness) plus neutral state. All the utterances are in Italian and they include both semantically correct and "nonsense" phrases. The audios are recorded considering a sampling frequency of 48 kHz as well as a bit depth of 16 bit.

In this preliminary experiments, CREMA_D and RAVDNESS have been selected as training and/or test set, while SAVEE has been used only for testing. EMOVO instead has been employed to test the performance of the model when a domain adaptation strategy is applied.

Using these datasets, two different pipelines are here compared: a basic not normalized framework (hereinafter referred to as *Basic_framework*), and a framework with normalization and pre-processing (*Norm_framework* from hereon), are depicted in Figure 3

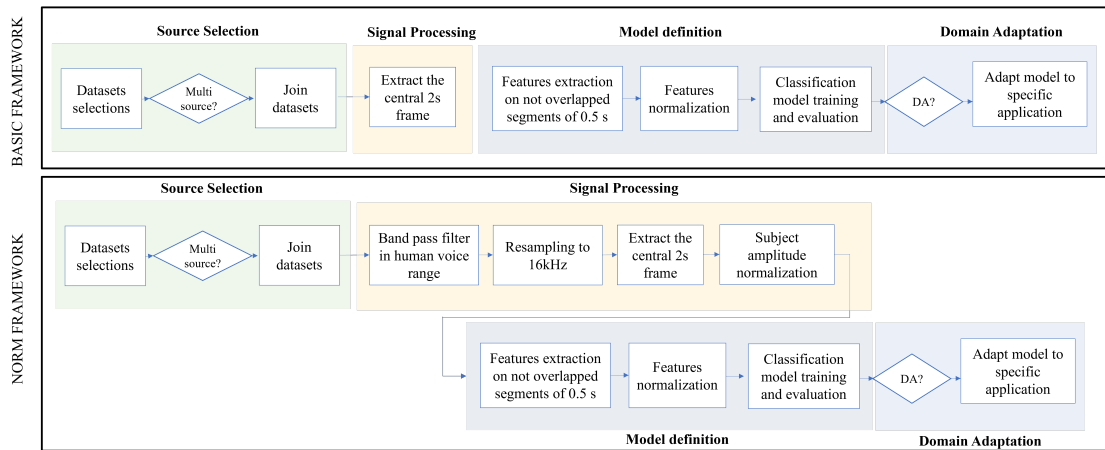


Figure 3: The two computational frameworks compared in the analysis

Basic_framework. Since the audios of the considered dataset are already mono signals, the first step in the *Basic_framework* pipeline is the segmentation in frames. To limit the variance in the length of the utterances, only the central 2-seconds frame of each audio is considered. A zero padding strategy has been applied to the signals shorter than 2-seconds to standardize them in number of samples. Furthermore, to take into account temporal variation of audio signal, in all the analysis performed, the 2-second audio signals have been segmented in four segments, using a not-overlapping window of 0.5 seconds. A total of 140 acoustic features were used in the definition of the model: 35 acoustic features for each segment in which audio signal have been divided.

In particular, both temporal and spectral features have been considered for this analysis:

- *MFCC*: the first 20 Mel-Frequency Cepstral Coefficients, representing the envelope of

the short-time power spectrum, were evaluated on each 0.5 second frame as descriptors of the shape of the vocal tract. A window size of 2048 samples and a hop length of 512 samples were selected for the computation of the Fast Fourier Transform (FFT).

- *Chroma features*: 12 chroma features (or pitch class profile) were evaluated for each frame as indicators of the harmonic and melodic characteristics of voice. Similarly to MFCC, the window size and the hop length of the FFT used to compute the chromagram were setted respectively to 2048 samples and 512 samples. In our analysis, the chroma spectrogram is wrapped averaging, for each frame and along the time-axis, the values of all the pitches belonging to same pitch classes.
- *RMS*: The global energy of each 0.5 second frame of the signal has been calculated by taking the Root Mean Square Energy of their amplitude. This prosodic features is often evaluated in SER and it allows to describe the audio loudness.
- *Spectral centroid*: The speech brightness is evaluated using the Spectral Centroid feature, which is the average of the signal frequency values weighted by the magnitude of each frequency. For the evaluation of the frequency content, a Fast Fourier Transform using a moving window of size 2048 samples and an overlap of 25% has been employed.
- *ZCR*: the Zero-Crossing Rate (ZCR) feature represent the number of time the signal cross from positive to negative and viceversa, resunting thus a good measure of the frequency content of the signal. A current analysis, a single ZCR is evaluated for each frames.

Norm_Framework. In addition to the standard operations applied to the audio signals by the Basic_Framework, the Norm_Framework incorporates further pre-processing steps for standardizing the data concerning variations in data sources or subjects. In particular, before applying the segmentation, a Butterworth band pass filter of 6-order is applied to the signals to select the frequency band related to human voice (300 Hz - 4.5 kHz).

In the Norm_Framework, the segmentation task is followed by a subject normalization step performed using the neutral-based normalization. The z-score parameters were evaluated for each subject based on his/her signals labeled as neutral emotion, and then used to standardise his/her remaining audios. Finally, a down-sampling step is applied to all the signals of both training and test set in order to standardize the data to a single Sampling Frequency. A fixed 16 KHz sampling frequency has been chosen in accordance with [32] and [33].

Notice that only a subset of the processing tasks outlined in Chapter 2 have been taken into account in the two proposed pipelines. In particular, we focused on the steps necessary to standardise data in acted datasets, as the only considered in this analysis.

For each experiment performed, the features extracted from the audio signals have been then used to train a multi-class gradient boosted decision trees algorithm implemented as XGBoost [34]. All the six universal emotions (angry, fear, disgust, surprise, happy, sad), plus the neutral state have been considered in the classification task. To assess and compare the performance of the different classifiers, a 5-folds subject independent cross validation strategy has been applied. In this method, the dataset is randomly partitioned in five subsets so that the data of the same subject never occurs into two different folds. At each iteration all observations from one of these groups of subjects are used to test the model, while the remaining observations are used as training set. Several well-known evaluation metrics have been computed from

the resulting confusion matrix. In particular, the overall performance of the classifier were measured using the Accuracy and Macro-F1 scores. The last experiment performed, concerns the use of a domain adaptation module to analyze the ability of the general model in adapting to a specific scenario represented by a small dataset. In this analysis, the domain adaptation is used to adapt the general model, trained on English utterances, to recognize data collected in Italian language from young adults. The Transfer AdaBoost for Classification (TrAdaBoost) supervised domain adaptation strategy has been tested for this purpose. In according to [14], the split of the data into Target and Test sets has been performed considering a Leave One Subject Out Cross validation strategy.

4. Discussion

The initial experiments aim to verify the benefits obtained from a multisource approach to build a general model for SER.

To this end we here consider as test sets: CREMA_D, RAVDESS, and a multisource dataset defined as union of CREMA_D and RAVDESS. The same CREMA_D and RAVDESS, together with the SAVEE dataset are then considered as test sets.

In Table 2 the classification results achieved in the different trials are summarized. The application of the proposed normalized framework allows to increase significantly the performance of the classification model in almost all the analysis carried out. In particular, Macro F1-Score values between 55% and 58% are obtained when the same dataset is employed as training and test set using the subject independent 5-fold cross validation. These values outperform the results achieved by the same model in the Basic_framework benchmark case (about 44%). When different datasets are used as training and test set, as expected, the general performance of the classification models decreases. However, also in these cases, the use of the Norm_framework allows to improve the performance of the models, especially when the SAVEE dataset is considered as test set. Herein, the use of multisource dataset as training set allows to achieve the best result when the normalization framework is applied with a Macro F1-Score value of 30%. It is worth noting how the use of multisource training datasets enhances performance compared to using a single dataset for training. This emphasises the significance of adopting a more diversified and heterogeneous set of data when training SER models.

The second set of experiments focuses on the domain adaptation (DA) step. The TrAdaBoost DA module here considered has been applied to fine tune the general model to Italian language. The performance of this module, compared with the performance obtained without domain adaptation, are reported in Table 2. Although the performance is not high, the results obtained show an increase both by including the normalisation procedure and applying DA, suggesting that this approach is noteworthy and should be further investigated. It is not easy to compare the results of our framework with others in the state of the art, mainly because even if several datasets are considered, they are different from the ones here adopted. Moreover, it is also difficult to find models that are validated using subject independent cross validation. Finally, up to our knowledge there are no other works that deal with the domain adaptation approach with respect to language or age.

Table 1

Comparison of the multi-class XgBoost performance varying the computational pipeline applied to normalize the audio signals (Basic and Norm). For each analysis, the dataset used to train the model is depicted on the rows, while on the column is reported the dataset used as test set. The results are reported considering two evaluation metrics: accuracy, and Macro F1-Score.

		Test	CREMA_D		RAVDESS		SAVEE	
			Accuracy	Macro F1-Score	Accuracy	Macro F1-Score	Accuracy	Macro F1-Score
Training								
CREMA_D	Basic		44%	43%	32%	25%	14%	4%
	Norm		56%	55%	41%	38%	29%	27%
RAVDESS	Basic		20%	10%	45%	45%	14%	5%
	Norm		30%	25%	57%	58%	21%	16%
Multisource (RAVDESS+CREMA_D)	Basic		-	-	-	-	14%	4%
	Norm		-	-	-	-	32%	30%

Table 2

Comparison of the multi-class XgBoost performance using the Multisource dataset to train the general model, and EMOVO as target and test set. The no Domain Adaptation (first rows) strategy is compared with the Domain Adaptation using TrAdaBoost (last rows). Each analysis is performed considering the two computational frameworks: without applying the pre-processing pipeline (Basic) and with pre-processing pipeline (Norm). The results are reported in term of Accuracy and Macro F1-score.

DA strategy		Target/Test	EMOVO	
			Accuracy	Macro F1-Score
No Domain Adaptation	Basic		24%	21%
	Norm		31%	32%
Domain Adaptation using TrAdaBoost	Basic		29%	34%
	Norm		34%	35%

5. Conclusion

The lack of huge emotionally labelled speech dataset makes it necessary to define strategies to merge individual and heterogeneous data into a single multisource dataset, to benefit from complementary information. The positive increase in performance obtained in this work highlighted the potential of defining a general computational framework capable of identifying emotions from unobserved data acquired in different experimental conditions. An important future development in this direction is to include in the multisource training set, natural conversations, which better reflect the real scenarios of applicability. A second interesting outcome of this work is the increase of performance obtained applying a domain adaptation module that fine tuned the general model to a specific scenario. In this work we have considered only DA on a different language, but we plan in the future to test our proposal also on a dataset of audios recorded from Italian older adults, composed of acted utterances and natural

conversations, that we have already collected but not completely labelled, and that is described in [9].

Acknowledgments

This research is partially supported by the FONDAZIONE CARIPLO “AMPEL: Artificial intelligence facing Multidimensional Poverty in ELderly” (CUP H45F20000840007 Ref. 2020-0232) and by the co-funding European Union – Next Generation EU, 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) PE00000015 – CUP: H43C22000840006.

References

- [1] A. Ahmad, P. Mozelius, Human-computer interaction for older adults: a literature review on technology acceptance of ehealth systems, *Journal of Engineering Research and Sciences (JENRS)* 1 (2022) 119–126.
- [2] S. Hutson, S. L. Lim, P. J. Bentley, N. Bianchi-Berthouze, A. Bowling, Investigating the suitability of social robots for the wellbeing of the elderly, in: *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*, Springer, 2011, pp. 578–587.
- [3] D. H. García, P. G. Esteban, H. R. Lee, M. Romeo, E. Senft, E. Billing, Social robots in therapy and care, in: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2019, pp. 669–670.
- [4] J. Broekens, M. Heerink, H. Rosendal, et al., Assistive social robots in elderly care: a review, *Gerontechnology* 8 (2009) 94–103.
- [5] L. Ragno, A. Borboni, F. Vannetti, C. Amici, N. Cusano, Application of social robots in healthcare: Review on characteristics, requirements, technical solutions, *Sensors* 23 (2023) 6820.
- [6] S. Valtolina, L. Hu, Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness, in: *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter, 2021*, pp. 1–5.
- [7] M. Bojanić, V. Delić, A. Karpov, Call redistribution for a call center based on speech emotion recognition, *Applied Sciences* 10 (2020) 4653.
- [8] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, *Speech Communication* 116 (2020) 56–76.
- [9] F. Gasparini, A. Grossi, Ser_ ampel: a multi-source dataset for speech emotion recognition of italian older adults, in: *Proceedings of the 12th Italian Forum Ambient Assisted Living, 2023*.
- [10] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, E. Ambikairajah, A comprehensive review of speech emotion recognition systems, *IEEE access* 9 (2021) 47795–47814.
- [11] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion., *Journal of personality and social psychology* 17 (1971) 124.
- [12] 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* 89 (2001) 344–350.

- [13] A. Mehrabian, J. A. Russell, *An approach to environmental psychology.*, the MIT Press, 1974.
- [14] F. Gasparini, A. Grossi, Sentiment recognition of italian elderly through domain adaptation on cross-corpus speech dataset, in: *Proceedings of the Italian Workshop on Artificial Intelligence for an Ageing Society 2022*, volume 3367 of *AI*IA*, CEUR, 2022, pp. 12–28.
- [15] A. Dehqan, R. C. Scherer, G. Dashti, A. Ansari-Moghaddam, S. Fanaie, The effects of aging on acoustic parameters of voice, *Folia Phoniatrica et Logopaedica* 64 (2013) 265–270.
- [16] F. Busquet, F. Efthymiou, C. Hildebrand, Voice analytics in the wild: Validity and predictive accuracy of common audio-recording devices, *Behavior Research Methods* (2023) 1–21.
- [17] R. Böck, O. Egorow, I. Siegert, A. Wendemuth, Comparative study on normalisation in emotion recognition from speech, in: *Intelligent Human Computer Interaction: 9th International Conference, IHCI 2017, Evry, France, December 11-13, 2017, Proceedings 9*, Springer, 2017, pp. 189–201.
- [18] M. S. Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, *Digital signal processing* 110 (2021) 102951.
- [19] X. Wu, Q. Zhang, Design of aging smart home products based on radial basis function speech emotion recognition, *Frontiers in Psychology* 13 (2022) 882709.
- [20] J. Chang, X. Zhang, Q. Zhang, Y. Sun, Investigating duration effects of emotional speech stimuli in a tonal language by using event-related potentials, *IEEE Access* 6 (2018) 13541–13554.
- [21] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern recognition* 44 (2011) 572–587.
- [22] E. H. Kim, K. H. Hyun, S. H. Kim, Y. K. Kwak, Improved emotion recognition with a novel speaker-independent feature, *IEEE/ASME transactions on mechatronics* 14 (2009) 317–325.
- [23] C. Yogesh, M. Hariharan, R. Yuvaraj, R. Ngadiran, S. Yaacob, K. Polat, et al., Bispectral features and mean shift clustering for stress and emotion recognition from natural speech, *Computers & Electrical Engineering* 62 (2017) 676–691.
- [24] S. Sahoo, P. Kumar, B. Raman, P. P. Roy, A segment level approach to speech emotion recognition using transfer learning, in: *Asian Conference on Pattern Recognition*, Springer, 2019, pp. 435–448.
- [25] M. N. Stolar, M. Lech, R. S. Bolia, M. Skinner, Real time speech emotion recognition using rgb image classification and transfer learning, in: *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, IEEE, 2017, pp. 1–8.
- [26] G. Boateng, T. Kowatsch, Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning, in: *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 12–16.
- [27] S. Akinpelu, S. Viriri, Robust feature selection-based speech emotion classification using deep transfer learning, *Applied Sciences* 12 (2022) 8265.
- [28] A. Hassan, R. Damper, M. Niranjana, On acoustic emotion recognition: compensating for covariate shift, *IEEE Transactions on Audio, Speech, and Language Processing* 21 (2013) 1458–1468.
- [29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, *Crema-d: Crowd-*

- sourced emotional multimodal actors dataset, *IEEE transactions on affective computing* 5 (2014) 377–390.
- [30] 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.
 - [31] S. Haq, P. J. Jackson, J. Edge, Audio-visual feature selection and reduction for emotion classification, in: *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08)*, Tangalooma, Australia, 2008.
 - [32] S. Sarker, K. Akter, N. Mamun, A text independent speech emotion recognition based on convolutional neural network, in: *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, 2023, pp. 1–4.
 - [33] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al., A database of german emotional speech., in: *Interspeech*, volume 5, 2005, pp. 1517–1520.
 - [34] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.