Sentiment recognition of Italian elderly through domain adaptation on cross-corpus speech dataset

Francesca Gasparini[1], Alessandra Grossi[1]

[1]*Department of Computer Science, Systems and Communications, University of Milano - Bicocca, Italy*

Abstract

The aim of this work is to define a speech emotion recognition (SER) model able to recognize positive, neutral and negative emotions in natural conversations of Italian elderly people. Several datasets for SER are available in the literature. However most of them are in English or Chinese, have been recorded while actors and actresses pronounce short phrases and thus are not related to natural conversation. Moreover only few speeches among all the databases are related to elderly people. Therefore, in this work, a multi-language and multi-age corpus is considered merging a dataset in English, that includes also elderly people, with a dataset in Italian. A general model, trained on young and adult English actors and actresses is proposed, based on XGBoost. Then two strategies of domain adaptation are proposed to adapt the model either to elderly people and to Italian speakers. The results suggest that this approach increases the classification performance, underlining also that new datasets should be collected.

Keywords

Speech emotion recognition, Sentiment recognition, Domain adaptation, cross-corpus SER, cross-language SER

1. Introduction

Emotions play a relevant role in defining individuals' behaviours and coordination in human-human interactions [1]. In particular, humans find speech conversations more natural and effective than its written form as way to express themselves [2]. During conversations, people try to convey their thought not only by words but also by bodily, vocal or facial expressions [1, 3]. Specifically in vocal expressions the affective state of individuals is expressed both by the linguistic and acoustic information carried by the speech [4]. For instance, the same sentence said with different intonations can express different emotions by the speaker and, thus, can lead to a different response from the listener [5]. Therefore, in order to create a natural interaction between humans and computers, the machine must be able to understand emotions from the speaker's voice and consequently adapt. Speech Emotion Recognition (SER) consists of the task of processing and classifying speech signals in order to recognize the emotional state of

AIXAS2022: Italian Workshop on Artificial Intelligence for an Ageing Society, Udine, Italy, November 29, 2022

✉ francesca.gasparini@unimib.it (F. Gasparini); alessandra.grossi@unimib.it (A. Grossi)
iD 0000-0002-6279-6660 (F. Gasparini); 0000-0003-1308-8497 (A. Grossi)

CEUR Workshop Proceedings (CEUR-WS.org)

the speaker [5, 6]. Systems based on SER have different fields of application, such as health care [7], e-learning tutoring [8], automotive [9] or entertainment [10, 11]. In particular, these kinds of systems can be employed for the definition of diagnostic tools able to help therapists in detecting psychological disorders [12] or for automatically recognising mental state alteration in drivers [13]. Automatic emotion detection systems can also be used in the call center or mobile communications to detect the emotions of callers and to help agents improving the quality of service [14, 15], or in human-robot interactions to support a more natural and social communication between human and machine [16, 17].
Several researches have been carried out in the field of Speech Emotion Recognition during the last three decades [18]. In particular, many of these analysis are performed considering only one between linguistic or acoustic information of speech while in recent analysis a multi-modal approach is examined [19].

In our study, we focus only on acoustic information. In this field, both traditional machine learning and deep learning approaches have been taken into account in previous literature. In general, the traditional pipeline in a SER system consists of three steps: signal preprocessing, features extraction and classification [20]. Concerning features extractions, different set of features have been tested: traditional features extracted by audio signals [2], including prosodic (such as pitch, energy and duration), spectral (such as fundamental frequency, Mel Frequency Cepstral Coefficients or Linear Prediction Cepstral Coefficients) and voice quality features (such as jitter or shimmer), as well as deep features extracted by pre-trained networks. In this latter, the audio signals are usually represented as Spectrogram or Scalogram and used as input to pre-trained network to extract features [21, 22]. With reference to classifiers, in several research such as [23, 24], traditional classifiers have been employed. In particular, according to [25], the classical classification techniques preferred in SER system are Gaussian Mixture Model, Hidden Markov Model, Artificial Neutral Network, Decision Trees and Support Vector Machine. In few analysis [26, 27] also ensemble techniques combining several classifiers have been tested. Deep approaches have been also considered in the last years. In particular, framework using Convolutional Neutral Network (CNN) [28], Recurrent Neural Network (RNN) [29] and Long Short-Term memory network (LSTM) [30] have been evaluated, using both traditional features [31] and raw audio signals. In some cases, also mechanism of attention [32, 33] or auto encoding [34] have been added to classifiers in order to increase performance. The main SER approaches have been summarized in review manuscripts such as [6] or [25].
Despite the huge number of analyzes carried out, there are still numerous issues that make difficult to recognize emotions in speech. In [18] some of these challenges and the approaches tested so far to solve them are summarized. In particular, speech emotion recognition algorithms struggle in recognize emotions when people of different language or age are considered.
In literature there are many datasets collected for SER purpose. These corpora can be classified into three groups with reference to how emotional speech is generated [35]: i) Acted datasets, where the data are collected from actors/actresses that try to simulate emotions; ii) Evoked or Elicited datasets, where the subjects are involved into situations especially created to evoke or induce certain emotions; and iii) Spontaneous or Natural datasets, which contain more authentic emotions as collected from real-world situations like call-centers or public places [18]. Most of the datasets available in the literature are composed of recited speeches [36], while only few

of them consider natural conversations [37, 38, 39]. Moreover, the considered languages are mainly English and Chinese. It has been demonstrated that language has a strong influence in how emotions are expressed [24], and thus multi-language datasets have been proposed [40, 41]. Age is another factor that influences the acoustic characteristics of the voice, especially in the case of elderly [42, 43]. However, this is still an open field of research and few works face the problem of SER in case of elderly, or varying the age [22, 33, 44, 45], and old subjects are rarely present in available datasets [46, 47, 48, 38].

In this work we consider the problem of SER, considering elderly Italian people. Moreover we focus on positive, neutral and negative emotions. We propose to consider a multi-language, multi-aged approach, considering a cross-corpus dataset, described in Section 2. We start from a general model trained on an English dataset of young and adult subjects, and we refine this model to adapt either to elderly and Italian language, as described in Section 3, adopting two different domain adaptation techniques. In Section 4 preprocessing of raw data, feature extraction and data augmentation, needed to apply the proposed solutions are presented. The results, discussed in Section 5, underline the potentialities and the limits of the proposed approaches, while future perspective are drawn in the Conclusions.

2. Cross-corpus dataset

In this work, we consider two datasets available in the literature, labeled with emotions, and characterized by the presence of elderly subjects or by the presence of Italian sentences: the CRowd-sourced Emotional Multimodal Actors dataset CREMA-D [47] and EMOVO [49].

CREMA-D [47] is a free audio-visual dataset collected to investigate facial and vocal expressions and perception of acted emotions. It consists of 7442 audio and video recordings of professional actors playing 12 utterances each one expressed in six emotional states (happy, sad, anger, fear, disgust and neutral) at different intensity levels. In the first utterance, the actors were directed to simulate each emotion in three levels of intensity (low, medium and high) while, for the other eleven sentences, they were free to express the emotion at their preferred intensity. The sentences selected for the experiment are in English and have a neutral semantic content. In total, 48 actors and 43 actresses of different ages and ethnicity were involved in the experiments, including 6 elderly with more than 60 years and 85 adults aged between 20 and 59 years. For the purpose of our analysis, the two groups of subjects are considered separately with a total of 492 signals for elderly, named hereinafter *CREMA-D-ELD*, and 6950 signals for adults (*CREMA-D-ADULT*). For further details of CREMA-D dataset, please refer to the reference manuscript [47].

EMOVO [49] is an acted free audio speech emotional dataset based on the Italian language. The corpus was collected from six young Italian actors (3 male and 3 female) with a mean age of 27.1 (no elderly actors were involved). Similarly to CREMA-D, in the experimental protocol, 14 utterances had to be performed by the actors simulating different emotional states. In particular, for each utterance, 7 affective states were considered: neutral, disgust, fear, anger, joy, surprise and sadness. The total number of utterances collected in the dataset is 588, with a mean of 98

Table 1
Summary of main CREMA-D and EMOVO characteristics

Dataset Name	Type	Emotions Considered	Language	No. of Utterances	Tot. No. of Subjects	No. of Males	No. of Elderly	Mode
CREMA-D	Acted	happy, sad, anger, fear, disgust and neutral	English	12	91	48	6	Audio/ visual
EMOVO	Acted	joy, surprise, sad, anger, fear, disgust and neutral	Italian	14	6	3	0	Audio

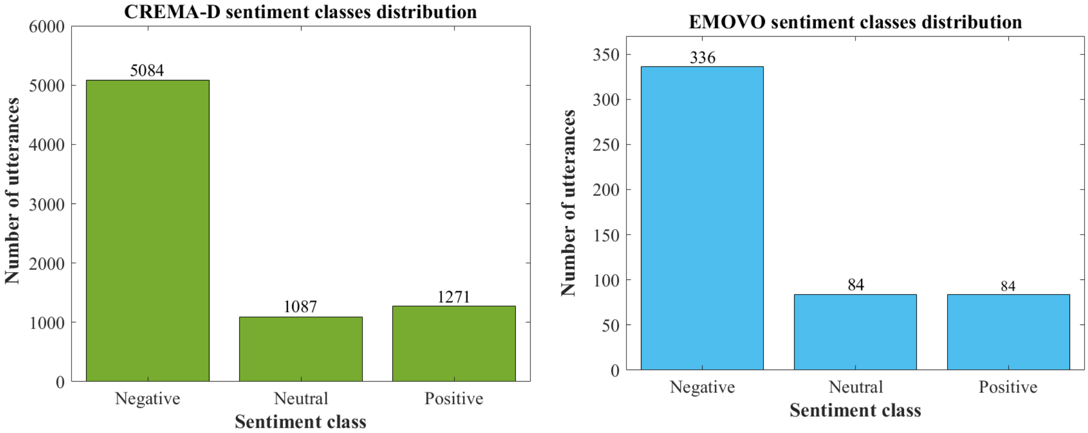

Figure 1: Number of utterances mapped as negative, positive or neutral in the two datasets CREMA-D (left) and EMOVO (right)

signals per actor. More details about EMOVO can be found in [49]

In Table 1 the main information about these two datasets are summarized.

In both the selected datasets, the signals are labeled using the six basic emotions defined by Ekman. In order to use these datasets in our analysis, each emotion has been converted into its respective sentiment according with the mapping defined in [50]. In particular, we have considered anger, fear, disgust and sadness as negative sentiments, happy (or joy) as positive sentiment and neutral as neutral sentiment. All the EMOVO signals labeled as “surprise” has been instead excluded from the analysis as difficult to be mapped into a single sentiment class [50]. The distribution of the utterances in the three sentiment classes is shown in Figure 1 for the two datasets considered.

Concerning the sentiment analysis, two other datasets are usually adopted in Speech Sentiment Recognition researches: Multimodal EmotionLines Dataset (MELD) [50] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [51]. The first [50] is a data corpus composed by more then 13000 utterances from 1433 dialogues from the TV-series Friends and labeled with three sentiment class: negative, positive and neutral. CMU-MOSEI [51], instead, contains 23453 annotated video-clips from 250 different topics, gathered from online video sharing websites and labeled with sentiment in Likert scale. Despite both the datasets are directly labeled with sentiment, they were excluded from our analysis. In particular, concerning MELD, the dataset has been discarded due to the presence, in several audios, of laugh tracks or multiple voices overlapping the main actor's speech. This makes the audio signal very noisy and makes it difficult to identify which part of the audio is related to the labelled sentiment. With reference to CMU-MOSEI, instead, the dataset has been excluded from the study because of the lack of the subject's age that makes impossible to separate signals collected from elderly from the one's collected from young or adults.

3. Data Adaptation strategies

The proposed analysis considers two research hypotheses:

- *Domain adaptation based on age*, training a general Speech Sentiment Recognition model using speech data collected from English young and adults subjects and adapting this model on new data collected from English elderly subjects.
- *Domain adaptation based on language*, trying to refine a pre-trained Speech Sentiment Recognition model on English young and adults subjects to recognize new data collected from Italian young and adults people.

In all the experiments performed, the gradient boosted decision trees algorithm implemented as XGBoost [52] has been selected as classification model while two different instance weighting domain adaptation strategies have been tested:

- the *Kullback-Leiber Important Estimation Procedure (KLIEP) strategy* [53] that assigns a weight to the training instances during the classifier learning task in order to minimize the Kullback-Leibler divergence between train and target distributions. In our analysis we have considered the supervised implementation of this algorithm using "rbf" as Kernel with two different gamma: 0.1 and 1.
- *the Transfer AdaBoost for Classification (TrAdaBoost)* [54] is a supervised domain adaptation strategy that extends boosting-based learning algorithms to the field of transfer learning. In particular, at each iteration, the algorithm trains a new weak classifier giving less importance to the training instances poorly predicted in previous iterations while emphasising the target samples correctly recognized. The final model is the combination of the last half computed estimators weighted according to their relevance. The number of iterations selected in our experiments is 10.

The application of these two strategies requires to split the data into three distinct sets: i) Training (or Source) set made up of a large amount of labeled data used to train the general model; ii) Target set consisting of few samples belonging to a new but related domain that are used to adapt the general model to this new data distribution and iii) Test set composed by

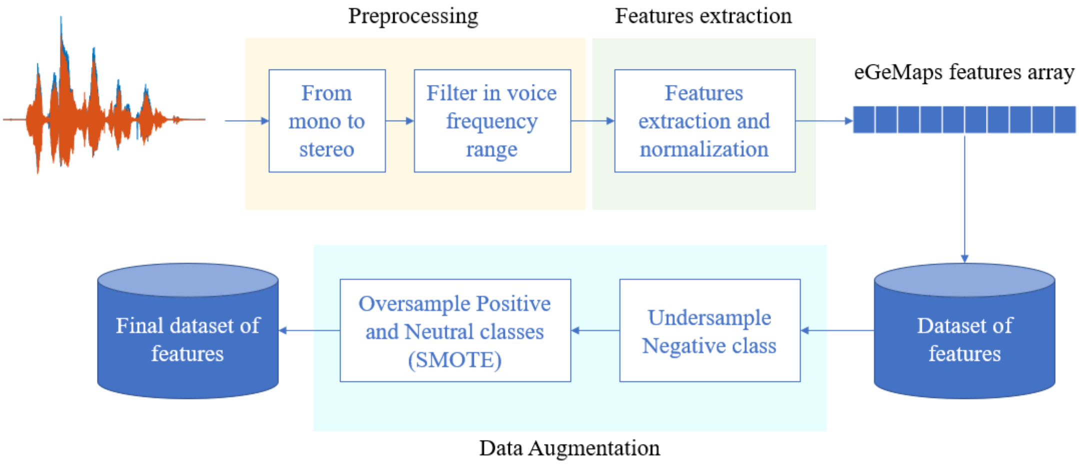
Preprocessing
Features extraction
From mono to stereo
Filter in voice frequency range
Features extraction and normalization
eGeMaps features array
Dataset of features
Undersample Negative class
Oversample Positive and Neutral classes (SMOTE)
Final dataset of features
Data Augmentation

Figure 2: Pipeline used, in the analysis, for extracting features from the signals of Training and Target datasets. Concerning the Test set, the Data Augmentation step is not applied.

data similar to Target set and used to evaluate the model performances. In our experiments, the definition of these three sets changes according to the research hypothesis considered. In particular, in multi-age analysis, the data of CREMA-D-ADULT have been used as Training set while Target and Test sets have been defined as subsets of CREMA-D-ELD. Instead, in multi-language analysis, the training of the general model is performed using CREMA-D-ADULT data while Target and Test sets are both defined as partitions of EMOVO data.

Different validation strategies have been tested to partition the data of CREMA-D-ELD and EMOVO into Target and Test set:

- Leave One Subject Out (LOSO) Cross Validation strategy, where the folds are partitioned according to subject and thus, at each iteration, all the data of a single subject are used as Test set while the data of the remaining subjects are used as Target set.
- Leave One Utterance Out (LOUO) Cross Validation strategy, where the folds are defined according to the pronounced utterances, thus at each iteration, all the data related to a single utterance are used to test the model while the data of the remaining utterances are used as Target set.

To test the performances of our classification models, several well-known evaluation metrics are computed [55] including the accuracy, single class F1-score, evaluated as the harmonic mean of single class precision and recall, and macro F1-score [56] computed as the unweighted mean of the single class F1-score.

4. Model input data

To apply the strategies of domain adaptation described in the previous section, preprocessing, feature extraction, and data augmentation to balance the classes have been performed on raw data. The whole process is depicted in Figure 2.

4.1. Preprocessing

The audio signals of each dataset are preprocessed to extract only the information concerning the target speaker's voice. In particular, the audio clips were first converted from stereo to mono by averaging samples across the two channels. Then, each signal was filtered using a pass-band Butterworth filter with lower cutoff frequency at 300 Hz and upper cutoff frequency at 3000 Hz to removes the spectral components out of the voice frequency range [57].

4.2. Feature extraction

From the pre-processed signals, the eGeMAPS acoustic feature set was extracted using the python library implementation of openSMILE toolkit [58]. The eGeMAPS feature set (extended Geneva Minimalistic Acoustic Parameter Set) [59] is a set of audio features proposed for affective analysis in voice signals. It consists of 25 Low Level Descriptor (LLD) features including energy, frequency, cepstral, spectral and dynamic parameters. In order to summarize the variation of these parameters over the time windows, some high level functional features are extracted using statistical functions as arithmetic mean, standard deviation or percentile. Applying these statistics, a total of 88 features have been extracted for each considered signal. The extracted features have been normalized by z-scoring in order to reduce inter signals differences.

4.3. Data Augmentation

Only for Training (or Source) and Target dataset, the feature extraction step has been followed by data augmentation. In both the datasets, the cardinality of the negative sentiment class is four times greater then positive or neutral ones. This is due to an imbalance among the number of emotions mapped as negative (angry, fear, sadness, disgust) and the number of emotions mapped as positive (happy) and neutral (neutral) in the selected emotion-sentiment transformation. In order to create more balanced classes, a two steps procedure have been applied to training and target data according to the experiment considered . First the majority class have been under-sampled, discarding randomly half of the negative instances. In this process, the discarded elements have been selected trying to keep balanced the number of elements for each negative emotions. Then an oversampling strategy based on SMOTE algorithm [60] has been applied to increase the number of samples in the two minority classes (positive and neutral). SMOTE (Synthetic Minority Oversampling TEchnique) [60] is an oversampling method that random generates new synthetic data for the minority class starting from the original data points. In

Table 2
Experiments carried out in cross-age analysis varying the Domain Adaptation Strategy considered (second column) and the Validation strategy adopted (last column). In all the experiments, the general model has been trained using the data of CREMA-D-ADULT while different subsets of CREMA-D-ELD dataset have been selected as Target and Test set.

	DA Strategy	**Training Set**	**Target set**	**Test set**	**Validation Strategy**
XGBoost classifier		*CREMA-D-ADULT*	*CREMA-D-ELD*		
	No Domain Adaptation	all 85 young and adults subjects	no target dataset	6 elderly x 12 utterances	Training / Test independent
	KLIEP	all 85 young and adults subjects	5 elderly x 12 utterances	1 elderly x 12 utterances	LOSO
			11 utterances x 6 elderly	1 utterance x 6 elderly	LOUO
	TrAdaBoost	all 85 young and adults subjects	5 elderly x 12 utterances	1 elderly x 12 utterances	LOSO
			11 utterances x 6 elderly	1 utterance x 6 elderly	LOUO

particular, at each iteration, the algorithm selects one of the k-nearest neighbors of a random minority class element and create new artificial elements linear interpolating the two instances using a random number between zero and one. The procedure is repeated until the cardinality of the classes is balanced.

5. Results and discussion

The aim of this work is define a classification model able to automatically recognize three sentiment states (positive, neutral and negative) using acoustic features extracted from speech when different age and language are considered. In particular, two different experiments have been carried out to evaluate the research hypothesis described in Section 3: domain adaptation on elderly and domain adaptation on language.

5.1. Domain adaptation on elderly

In the first analysis, a multi-age corpus sentiment classification is considered. As described in Section 3, the two parts of CREMA-D dataset have been used respectively for Training set (CREMA-D-ADULT) and Target and Test set (CREMA-D-ELD). For each domain adaptation strategy, two different evaluation methods are tested: LOSO and LOUO. The results achieved in these experiments are compared with the performances reached by the XGBoost model when no domain adaptation strategy is applied. In this case, thus, the classifier is trained on CREMA-D-ADULT data and tested on the independent dataset CREMA-D-ELD. The classification settings considered in the analysis are summarized in Table 2. For each of these analyses, Table 3 reported the classification performance achieved by the XGBoost classifiers in terms of accuracy, macro

Table 3
Cross-age performance comparison using CREMA-D-ADULT as training set and CREMA-D-ELD as target and test set. The analysis are performed varying the Domain Adaptation strategy (second column) and Performance Evaluation method (third column). Three evaluation metrics are considered: macro F1-score, accuracy and single class F1-score.

Classifier	DA Strategy	Validation Strategy	Macro F1-score	Negative F1-score	Neutral F1-score	Positive F1-score	Accuracy
XGBoost classifier	No Domain Adaptation	Training / Test independent	62%	0,79	0,55	0,51	70%
	KLIEP	LOSO	60%	0,76	0,57	0,49	67%
		LOUO	60%	0,76	0,56	0,47	67%
	TrAdaBoost	LOSO	62%	0,79	0,56	0,52	70%
		LOUO	62%	0,78	0,55	0,52	69%

F1-score and single class F1-score. The results show how, in case of elderly, the use of domain adaptation techniques does not significantly increase the performances of the classification model with reference to the benchmark case without adaptation. A macro F1-score value of 62%, in fact, is achieved both when TrAdaBoost or no domain adaptation is applied. Lower performances are instead obtained using the KLIEP domain adaptation algorithm with F1-score value near to 60%. Similar results are reached using both LOSO and LOUO evaluation strategies. Considering the values of per-class F1-scores reached emerges how, in all the experiments performed, the Negative class appears easier to be recognized than Neural and Positive ones. This difference can be due to the presence of a higher number of different instances in the negative class than in the other two classes where several instances were artificially created using SMOTE data augmentation strategy.
From these preliminary results, it seems that data adaptation does not increase the performance of the proposed SER model. This is probably related to several aspects. The elderly here considered are actors or actresses, and thus they are not so significantly different from a population of young and adult persons. Moreover the elderly are only 6, of which only one is a female. A more realistic dataset should be consider to proper verify this research question.

5.2. Domain adaptation on language

The second part of our study focused on speech sentiment recognition when multi-language-corpus datasets are taken into account. The trials tested for this analysis are summarized in Table 4. Two different datasets were used: the English dataset CREMA-D-ADULT, used to train the model, and the Italian dataset EMOVO, as Target and Test set. Furthermore, similarly to elderly, the results obtained varying the domain adaptation technique (KLIEP and TrAdaBoost) and evaluation strategy (LOSO, LOUO) were compared with the performance reached by the classification model trained without domain adaptation. The values of accuracy, macro-F1 score and per-class F1-scores achieved in the different experiments are reported in Table 5.

Table 4
Experiments carried out in cross-language analysis varying the Domain Adaptation Strategy considered (second column) and the Validation strategy adopted (last column). In all the experiments the general model has been trained using the data of CREMA-D-ADULT while different subsets of EMOVO dataset have been selected as Target and Test set.

	DA Strategy	Training Set	Target set	Test set	Validation Strategy
XGBoost classifier		*CREMA-D-ADULT*	*EMOVO*		
	No Domain Adaptation	all 85 young and adults subjects	no target dataset	6 subjects x 14 utterances	Training / Test independent
	KLIEP	all 85 young and adults subjects	6 subjects x 14 utterances	1 subject x 14 utterances	LOSO
			13 utterances x 6 subjects	1 utterance x 6 subjects	LOUO
	TrAdaBoost	all 85 young and adults subjects	5 subjects x 14 utterances	1 subject x 14 utterances	LOSO
			13 utterances x 6 subjects	1 utterance x 6 subjects	LOUO

Table 5
Cross-language performance comparison using CREMA-D-ADULT as training set and EMOVO as target and test set. The analysis are performed varying the Domain Adaptation strategy (second column) and Performance Evaluation method (third column). Three evaluation metrics are considered: Macro F1-score, Accuracy and single class F1-score.

Classifier	DA Strategy	Validation Strategy	Macro F1-score	Negative F1-score	Neutral F1-score	Positive F1-score	Accuracy
XGBoost classifier	No Domain Adaptation	Training / Test independent	35%	0,71	0,28	0,07	57%
	KLIEP	LOSO	33%	0,64	0,19	0,16	48%
		LOUO	32%	0,68	0,22	0,06	51%
	TrAdaBoost	LOSO	44%	0,68	0,25	0,39	56%
		LOUO	85%	0,91	0,90	0,74	88%

From the analysis of the results, it emerges how the best performances in both the validation strategies were obtained applying the TrAdaBoost domain adaptation method. In particular, the two macro F1-score values of 44% and 85% generated using respectively LOSO and LOUO validation strategies outperform the value of 35% reached when no domain adaptation is considered. Similarly to elderly, the lowest general performances were instead reached applying the KLIEP domain adaptation strategy with macro F1-score values near to 32% in both the analysis performed. Another general consideration regards the single classes recognition. In almost all the trials, the use of domain adaptation techniques allowed to better recognize the instances of Positive class, reaching often more balanced classification performances in identify the three sentiments. Nevertheless, the Negative sentiment is still the class better recognized from all the classification models examined, thus confirming what has already been observed on the elderly analysis.

Finally, the last remark concerns the performance differences between the two validation strategies applied. In particular, the partition of Target and Test set using utterances allows to achieve better results than the one based on subjects. This can be explained by the fact that, in addition to language, the division by utterance also takes into account the difference between people with regard to personal vocal characteristics or how they express their emotions. Using this method, data from each of the analyzed subjects appear in each of the folds generated, allowing the classification model to better learn about vocal timbre differences or differences in the individuals' personalities. However, it is worth to underline that both the datasets analyzed are acted, making perhaps more similar how the same subject expresses the same emotion, also in different sentences. For this reason, in future analyzes, it may be necessary to validate the hypotheses here proposed on new natural datasets collected in real situations.

All the analysis were run on a computer with Intel Core™ i7-7700HQ Processor using 16 GB of RAM and 2.80 GHz CPU. In both the experiments, the proposed techniques take approximately 110 ms to extract features and classify a new instance lasting about 2 seconds. Not relevant variations in computational time have been detected when different domain adaptation strategies are applied. The high execution speed of the algorithms could allow their integration into near real-time systems. In this case, streams of audio directly collected from the speaker might be divided into segments of about two seconds and sent to the algorithm for processing and classifying, generating thus a response to the user with a delay of two seconds. The implementation of such systems implies, however, that the classifier used in the process was already trained and adapted to the new data. These operations are highly time-consuming and often require an execution time of minutes to be performed. For this reason, the proposed domain adaptation techniques seem not suitable in the definition of classifiers continuously adapting to newly acquired data, appearing instead useful in the development of near real-time systems based on cyclic updated pre-trained classifiers or batch processing system. Future analysis will be carried out in this regard.

6. Conclusion

The sentiment emotion recognition task is still an open field of research, especially when considering different languages and ages. In particular in the case of our interest, Italian elderly, no datasets are available in the literature. Domain adaptation techniques could partially solve this lack of data. However our preliminary results indicate that there is the urgency of a more realistic collection of data, that also faces the need of considering different ages. Domain adaptation techniques seem to better perform in case of cross-language datasets, paving the way for further researches in this direction. In particular, after a proper data collection, future experiments could be conducted considering both language and local dialects, particularly widespread among the elderly population. For what concerns the lack of performance increase applying domain adaptation models in the case of multi-age corpus, conclusions can not be drawn, due to the peculiarity of the datasets available (where the collected speeches were recorded by professional actors) and given the low presence of elderly people. Finally, in the presented study only audio signals have been taken into account. In the last years, the use of acoustic or textual features extracted from speech has been often paired with the use of other data collected from the speakers. In particular, in several literature datasets, visual signals such as face expressions or body movements, physiological signals or behavioural biometric data have been collected together with audio to consider a multimodal approach of Speech Emotion Recognition [6]. In future works, similar strategies could be applied also in case of elderly Italian people in order to create more robust and accurate cross-corpus emotional classifiers.

Acknowledgments

This research is supported by the FONDAZIONE CARIPLO "AMPEL: Artificial intelligence facing Multidimensional Poverty in ELderly" (Ref. 2020-0232).

References

[1] J. Lange, M. W. Heerdink, G. A. Van Kleef, Reading emotions, reading people: Emotion perception and inferences drawn from perceived emotions, Current Opinion in Psychology 43 (2022) 85–90.

[2] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern recognition 44 (2011) 572–587.

[3] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, International Journal of Speech Technology 21 (2018) 93–120.

[4] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artificial Intelligence Review 43 (2015) 155–177.

[5] X. Wu, Q. Zhang, Design of aging smart home products based on radial basis function speech emotion recognition., Frontiers in Psychology 13 (2022) 882709–882709.

[6] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76.

[7] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE transactions on Biomedical Engineering 47 (2000) 829–837.

[8] A. Hua, D. J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, A. Purandare, Using system and user performance features to improve emotion detection in spoken tutoring dialogs, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 2, 2006, pp. 797–800.

[9] D. Cevher, S. Zepf, R. Klinger, Towards multimodal emotion recognition in german speech events in cars using transfer learning, arXiv preprint arXiv:1909.02764 (2019).

[10] R. Nakatsu, J. Nicholson, N. Tosa, Emotion recognition and its application to computer

agents with spontaneous interactive capabilities, in: Proceedings of the seventh ACM international conference on Multimedia (Part 1), 1999, pp. 343–351.

[11] A. Alhargan, N. Cooke, T. Binjammaz, Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals, in: Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 479–486.

[12] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, N. B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Transactions on Biomedical Engineering 58 (2010) 574–586.

[13] F. Al Machot, A. H. Mosa, K. Dabbour, A. Fasih, C. Schwarzlmüller, M. Ali, K. Kyamakya, A novel real-time emotion detection system from audio streams based on bayesian quadratic discriminate classifier for adas, in: Proceedings of the Joint INDS'11 & ISTET'11, IEEE, 2011, pp. 1–5.

[14] P. Gupta, N. Rajput, Two-stream emotion recognition for call center monitoring, in: Eighth Annual Conference of the International Speech Communication Association, Citeseer, 2007.

[15] C. Vaudable, L. Devillers, Negative emotions detection as an indicator of dialogs quality in call centers, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 5109–5112.

[16] F. Hegel, T. Spexard, B. Wrede, G. Horstmann, T. Vogt, Playing a different imitation game: Interaction with an empathic android robot, in: 2006 6th IEEE-RAS International Conference on Humanoid Robots, IEEE, 2006, pp. 56–61.

[17] C. Jones, A. Deeming, Affective human-robotic interaction, in: Affect and emotion in human-computer interaction, Springer, 2008, pp. 175–185.

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

[20] A. Thakur, S. Dhull, Speech emotion recognition: A review, Advances in Communication and Computational Technology (2021) 815–827.

[21] 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.

[22] 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.

[23] M. Swain, S. Sahoo, A. Routray, P. Kabisatpathy, J. N. Kundu, Study of feature combination using hmm and svm for multilingual odiya speech emotion recognition, International Journal of Speech Technology 18 (2015) 387–393.

[24] S. Latif, A. Qayyum, M. Usman, J. Qadir, Cross lingual speech emotion recognition: Urdu vs. western languages, in: 2018 International Conference on Frontiers of Information Technology (FIT), IEEE, 2018, pp. 88–93.

[25] 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.

[26] M. Lugger, M.-E. Janoir, B. Yang, Combining classifiers with diverse feature sets for robust

speaker independent emotion recognition, in: 2009 17th European Signal Processing Conference, IEEE, 2009, pp. 1225–1229.

[27] B. Schuller, M. Lang, G. Rigoll, Robust acoustic speech emotion recognition by ensembles of classifiers, in: Tagungsband Fortschritte der Akustik-DAGA# 05, München, 2005.

[28] A. M. Badshah, J. Ahmad, N. Rahim, S. W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: 2017 international conference on platform technology and service (PlatCon), IEEE, 2017, pp. 1–5.

[29] K. Aghajani, I. Esmaili Paeen Afrakoti, Speech emotion recognition using scalogram based deep structure, International Journal of Engineering 33 (2020) 285–292.

[30] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, N. Dehak, Deep neural networks for emotion recognition combining audio and transcripts, arXiv preprint arXiv:1911.00432 (2019).

[31] H. S. Kumbhar, S. U. Bhandari, Speech emotion recognition using mfcc features and lstm network, in: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2019, pp. 1–3.

[32] B. T. Atmaja, M. Akagi, Speech emotion recognition based on speech segment using lstm with attention model, in: 2019 IEEE International Conference on Signals and Systems (ICSigSys), IEEE, 2019, pp. 40–44.

[33] Q. Jian, M. Xiang, W. Huang, A speech emotion recognition method for the elderly based on feature fusion and attention mechanism, in: Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), volume 12167, SPIE, 2022, pp. 398–403.

[34] M. Neumann, N. T. Vu, Improving speech emotion recognition with unsupervised representation learning on unlabeled speech, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7390–7394.

[35] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2012) 99–117.

[36] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the recola multimodal corpus of remote collaborative and affective interactions, in: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 2013, pp. 1–8.

[37] S. Steidl, Automatic classification of emotion related user states in spontaneous children's speech, Logos-Verlag Berlin, Germany, 2009.

[38] W. Fan, X. Xu, X. Xing, W. Chen, D. Huang, Lssed: a large-scale dataset and benchmark for speech emotion recognition, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 641–645.

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

[40] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, B. Schuller, Categorical vs dimensional perception of italian emotional speech (2018).

[41] V. Hozjan, Z. Kacic, A. Moreno, A. Bonafonte, A. Nogueiras, Interface databases: Design and collection of a multilingual emotional speech database., in: LREC, 2002.

[42] D. Deliyski, Steve An Xue, Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications, Educational gerontology 27 (2001)

159–168.

[43] J. Sundberg, M. N. Thörnvik, A. M. Söderström, Age and voice quality in professional singers, Logopedics Phoniatrics Vocology 23 (1998) 169–176.

[44] D. Verma, D. Mukhopadhyay, Age driven automatic speech emotion recognition system, in: 2016 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2016, pp. 1005–1010.

[45] G. Soğancıoğlu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadée, A. A. Salah, A. Karpov, Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition, arXiv preprint arXiv:2009.03432 (2020).

[46] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, et al., The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks (2020).

[47] 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.

[48] M. K. Pichora-Fuller, K. Dupuis, Toronto emotional speech set (TESS), 2020. URL: https://doi.org/10.5683/SP2/E8H2MF. doi:10.5683/SP2/E8H2MF.

[49] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, Emovo corpus: an italian emotional speech database, in: International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[50] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint arXiv:1810.02508 (2018).

[51] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[52] 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.

[53] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, M. Kawanabe, Direct importance estimation with model selection and its application to covariate shift adaptation, Advances in neural information processing systems 20 (2007).

[54] W. Dai, Q. Yang, G.-R. Xue, Y. Yu, Boosting for transfer learning, volume 227, 2007, pp. 193–200. doi:10.1145/1273496.1273521.

[55] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, arXiv preprint arXiv:2008.05756 (2020).

[56] Z. C. Lipton, C. Elkan, B. Naryanaswamy, Optimal thresholding of classifiers to maximize f1 measure, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2014, pp. 225–239.

[57] B. Birch, C. Griffiths, A. Morgan, Environmental effects on reliability and accuracy of mfcc based voice recognition for industrial human-robot-interaction, Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 235 (2021) 1939–1948.

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

[59] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al., The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, IEEE transactions on affective computing 7 (2015) 190–202.

[60] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.