Language-agnostic speech anger identification

Alessandra Saitta and Stavros Ntalampiras
Department of Computer Science
University of Milan, Milan, Italy
alessandra.saitta@studenti.unimi.it, stavros.ntalampiras@unimi.it

***Abstract*—Following the constantly increasing adoption of affective computing based solutions, this paper investigates the feasibility of multilingual anger identification. To this end, we formed such a corpus by suitably combining seven different datasets representing five different languages, i.e. English, German, Italian, Urdu, and Persian. After analyzing the diverse characteristics of the datasets, we designed four classification algorithms, namely Support Vector Machine, Decision Tree-based Bagging scheme, Convolutional Neural Network, and Convolutional Recurrent Neural Network. Such classification mechanisms are trained on appropriate features extracted from time and/or frequency domains, while speech data have been balanced considering every diverse characteristic incorporated in the datasets (language, sex, acted, etc.). Our findings render multilingual anger identification feasible since the proposed audio pattern recognition methodology based on Mel-spectrograms and CRNN achieved quite satisfactory identification rates.**

***Index Terms*—speech emotion recognition, multilingual emotion recognition, audio pattern recognition, deep learning**

I. Introduction

With the intensive development and application of artificial intelligence based solutions in human daily life, automatic speech emotion recognition (SER) is gaining ever-increasing attention by the scientific community [1]–[3]. At the same time, there are already commercial solutions utilizing such technology, e.g. a library produced by a Dutch company Vokaturi[1]. As such tools and methodologies constantly expand and become popular across different countries and cultures, the problem of generalization over an increased degree of diversity types arises. In this direction, one of the primary obstacles to be tackled is the existing language differences, which is the focus of this work.

Keeping these difficulties in mind, our first concern towards developing a language-agnostic methodology for SER was identifying available datasets. The existing datasets are characterized by broad diversity expressed in typologies and structures. Such variety involves not only language differences but also different purposes [4], numerous audio recording modes, diverse retrieval methods, acted or non-acted situations, and heterogeneous typologies of labels. More specifically, emotions are represented both in continuous and discrete ways. For example, the *Core-Affect* space described by Russell [5] refers to an emotion mapping on a valence-arousal plane. On the other hand, scientists such as Paul Ekman use an emotional expression labeling method based on the combination of *Basic emotions* [6]. As such, it is of paramount importance to test the robustness of a SER system with a wide combination of emotionally annotated databases [4], [7]. The specific obstacle needs special attention in a cross-language setting.

Unfortunately, there has not been a lot of work on language-agnostic speech emotion recognition. Even though the area of monolingual SER is characterized by a plethora of approaches where multiple features, datasets, and classifiers have been investigated, the same is not true for multilingual SER [8]–[10]. Nonetheless, there are several works exploring SER, where typically the problem is simplified into a binary one where the speech can be associated either with positive or negative emotional states [11], [12], while the only work considering all *big-six* emotions in a bilingual setting is presented in [13]. This work builds upon the gained experience and moves towards the inclusion of more languages while focusing on the emotional state of *anger*.

The specific problem is of particular interest in human computer interaction (HCI) systems, where it is directly associated with user satisfaction and the quality of provided service. At the same time, it could provide an important functionality in conversational agents as well as summarization solutions. As such, multilingual anger identification could provide useful information in assessing HCI systems targeting diverse audiences. Inline the previously mentioned SER systems, the related literature include monolingual solutions, such as Arabic [14], Greek [15], German [16]–[18], and English [19]. This work wishes to fill the existing gap and design an approach for language-agnostic anger identification.

This work contributes to structuring a standardized multilingual corpus of emotional speech, where the main focus is the anger state. We combined seven datasets representing five different languages and formed a binary arrangement *anger vs. rest*. During the audio pattern recognition stage, we investigated a wide range of time and spectral features combined with Support Vector Machines as well as Mel-scaled spectrograms modeled by deep networks including both convolutional and recurrent layers. During train, validation, and test stages, care was taken in order to elaborate on data that are balanced in terms of languages. After extensive experiments, we provide thorough experimental results and identify the best-performing method.

The rest of this work is organized as follows: Section II formulates the problem and section III provides a brief description of the employed datasets as well as the employed feature sets. Section IV outlines the classification algorithms, while section V presents the experimental set-up and results.

[1] https://vokaturi.com/

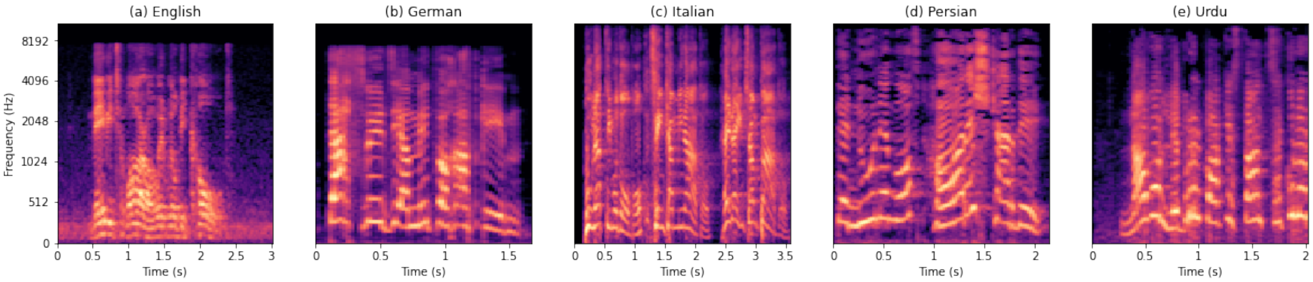

Fig. 1: Mel-Spectrograms representing the emotional state of anger across different languages.

Finally, in section VI, we draw our conclusions along with several promising research directions.

II. Problem Formulation

This work assumes the availability of corpus TS encompassing single-channel recordings of emotional speech representing various states in dictionary $\mathcal{D}$ including *anger*. Importantly, we relax the monolingual assumption since the language could be any of the following: a) English, b) German, c) Italian, d) Urdu, and e) Persian. Moreover, we assume that such emotionally annotated speech recordings follow consistent, yet unknown probability density functions denoted as $P_i, i \in \mathcal{D}$ [20], [21]. The final goal is to identify anger in novel speech recordings.

III. The dataset construction

For this study, we constructed a multi-language dataset combining 7 different speech corpora. Interestingly, 5 of them contain acted utterances (SAVEE, RAVDESS, CREMA-D, EmoDB, EMOVO) while the rest are databases of semi-natural spoken sentences extracted from different media such as talk shows and online radio broadcasts. It should be noted that each utterance of these datasets is labeled with a categorical emotion. Following the problem formulation outlined in section II, we have considered 2 labels summarized as ”Anger” and ”Other Emotion”.

A. Speech databases

This subsection briefly describes the datasets employed towards forming a multilingual corpus of emotional speech. Mel-Spectrograms representing the emotional state of anger across different languages are demonstrated in Fig. 1.

1) SAVEE: The Surrey Audio-Visual Expressed Emotion (SAVEE) Database [22] contains videos and audio tracks of 480 utterances in British English spoken by 7 male actors and evaluated by a group of postgraduates and young researchers at the University of Surrey. The corresponding labels are the following: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, and *neutral*.

2) RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song [23] is a corpus that includes audio and visual tracks of both speech and songs in American English performed by 24 actors of both sexes. Following the problem formulation outlined in section II, we considered audio files containing spoken utterances labeled as *calm, happy, sad, angry, fearful, surprised, and disgusted*.

3) CREMA-D: The *Crowd-Sourced Emotional Multimodal Actors Dataset* [24] is a dataset of 7,442 audio-visual recordings of sentences spoken in English by 91 actors of different ethnic backgrounds. The utterances are evaluated thanks to a crowdsourcing campaign taking into consideration the discrete emotion *happiness, sadness, anger, fear, disgust, and neutral state*.

4) EmoDB: This is a German emotional speech database [25] and one of the most famous datasets used for emotion classification. It contains almost 500 utterances spoken by 5 actors and 5 actresses and labeled as *neutral, angry, bored, disgusted, anxious/fearful, happy, sad*.

5) EMOVO: This dataset refers to an Italian emotional speech corpus [26]. It is composed of almost 600 utterances pronounced by both male and female actors. The audio recordings are labeled taking into account 7 emotions:*disgust, joy, fear, anger, surprise, sadness, neutral*.

6) ShEMO: Sharif Emotional Speech Database [27] is a Persian emotional speech dataset that contains 3000 semi-natural utterances extracted from online radio plays and labeled considering the emotions *anger, fear, happiness, sadness, surprise, and neutral state*, by a group of 12 annotators of both sexes.

7) Urdu: This is the first collection of spontaneous emotional speech in the Urdu language [7]. The corpus is formed by 400 audio traces extracted from Urdu TV talk shows and labeled by a group of students of NUST and CIIT universities. They have considered four discrete emotions: *anger, happiness, sadness, and neutrality.*

B. Features Extraction

In this study, we have extracted two sets of features from the signal of the audio tracks following the specifications of each classification mechanism (see section IV). To this end, every signal has been sampled at $22050 Hz$ using the Librosa library [28] to homogenize the corpus.

The first feature set is composed of numeric properties of the audio signal and is used by a Bagging algorithm and a Support Vector Machine (SVM). The second feature set contains Mel-Spectrograms characterizing the structure of every signal. These features are used to train and test a Convolutional

TABLE I: The datasets considered in the present work.

Name	Reference	Language	Acted/Real	Speakers	Emotional states	No. utterances
SAVEE	[22]	British English	Acted	7 male	Anger, disgust, fear, happiness, neutral, sadness, and surprise	480
RAVDESS	[23]	American English	Acted	12 female, 12 male	Anger, calm, disgust, fear, happiness, sadness, and surprise	1440
CREMA-D	[24]	English	Acted	48 male, 43 female	Anger, disgust, fear, happiness, neutral, and sadness	7442
EmoDB	[25]	German	Acted	5 male, 5 female	Anger, boredom, disgust, anxiety/fear, happiness, neutral, and sadness	535
EMOVO	[26]	Italian	Acted	3 male, 3 female	Anger, disgust, fear, joy, neutral, sadness, and surprise	588
ShEMO	[27]	Persian	Real	56 male, 31 female	Anger, fear, happiness, neutral, sadness, and surprise	3000
Urdu	[7]	Urdu	Real	27 male, 11 female	Anger, happiness, sadness, and neutral state	400

Neural Network (CNN) and a Convolutional Recurrent Neural Network (CRNN).

1) The first feature set: The first set is composed of the following features [28]:

1) **Mean Mel Frequency Cepstral Coefficients**: dimensionality equal to 20;
2) **Mean Zero Crossing Rate**: dimensionality equal to 1;
3) **Mean Spectral Centroid**: dimensionality equal to 1;
4) **Mean Spectral Roll-off**: dimensionality equal to 1;
5) **Difference in energy between the frequency band between 50 and 1000 Hz and the frequency band between 1000 and 4000 Hz**: let us denote with HI = mean of the absolute FFT amplitudes for frequencies in $[1000Hz, 4000Hz]$ and LO = mean of the absolute FFT amplitudes for frequencies in $[50Hz, 1000Hz]$, thus the feature will be computed as $|HI - LO|$ (dimensionality equal to 1);
6) **Mean quadratic power of the signal**: for an audio signal **s** with length N where s_i is an element of **s**, the quadratic power is computed as follows: $\frac{\sum_{i=1}^{N} s_i^2}{N}$ (dimensionality equal to 1);
7) **Mean fundamental frequency of the signal**: the fundamental frequencies are estimated with the YIN algorithm. The minimum frequency taken in consideration by the function corresponds to a C2 note while the maximum to a C7 one (dimensionality equal to 1);
8) **Minimum and maximum pitch**: Minimum and maximum fundamental frequencies, dimensionality equal to 2;
9) **Mean Linear Prediction Coefficients of grade 2**: here, we applied a second-grade filter since we intended to cause a sharpening effect on the differences in the signals (dimensionality equal to 3).

This dataset is formed by 34 features and 13885 samples.

2) The second feature set: This representation matches the properties of deep networks able to capture and reveal localized patterns which are of significant importance in audio pattern recognition applications [29], [30]. We converted the audio signals to suitable Mel-Spectrograms, the values of which are then converted in dB. It should be noted that the extraction algorithm generates 128 Mel-bands. The standard extraction method is followed including the computation of short time Fourier transform. The images are extracted using Matplotlib [31] and has $0.72X0.72$ dimensions. Representative Mel-spectrograms demonstrating the emotional state of anger across the considered languages are shown in Fig. 1.

IV. The classification models

This section outlines briefly the four different classification algorithms that were explored. The interested reader can find additional details in the provided references.

A. Support Vector Machine

The specific classifier aims at discovering the optimal hyperplane separating the two classes so that the margin between the support vectors is maximized. Such support vectors include class-specific representatives coming from the training data; these are relevant cornerstones during the determination of the optimal hyperplanes discriminating the two classes. In terms of parameterization, the most important design choice is the kernel function which maps the data in a higher dimensional space easing class separation. Here we experimented with linear, polynomial, radial basis function, and sigmoid. Finally, it should be mentioned that the final SVM was determined after an optimization scheme based on 5-fold cross-validation [32], which resulted in the use of a polynomial kernel of fifth grade.

B. Bagging classification scheme

Here, we employed collaborative schemes, which are typically used in most of real-world applications where discovering a single best classifier can be a task of increased difficulty. The present scheme uses 90 decision trees as base estimators, trained on maximum the 80% of the training set. Similarly to the SVM, Bagging parameters were optimized based on 5-fold cross-validation [32].

C. Deep Neural Networks

Following the current trend in audio pattern recognition, the next models comprise two neural networks constructed with

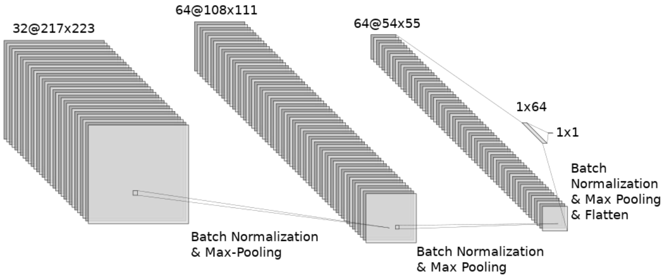
32@217x223
64@108x111
64@54x55
1x64
1x1
Batch Normalization & Max-Pooling
Batch Normalization & Max Pooling
Batch Normalization & Max Pooling & Flatten

Fig. 2: The structure of the considered CNN.

Keras[2] and Tensorflow[3] libraries. The first one is a convolutional neural network while the second includes a recurrent structure aiming at capturing existing temporal dependencies.

1) CNN: The CNN used in this work is composed of 3 convolutional layers. The first convolutional layer has a 32-dimensional output space while the following are 64-dimensional. Fig. 2 illustrates the detailed structure of the considered CNN. The convolutional network is followed by a dropout layer (35% of the inputs are dropped), a flattening layer, a 64-neurons dense layer, another dropout layer (35%), and a final dense layer composed of a single neuron with the sigmoid activation function. In addition, between the second and the third convolutional layer we inserted a 25% dropout layer. The activation function used for the other neurons of the network is a ReLu [33].

2) CRNN: The convolutional, as well as the completing components of the convolutional recurrent neural network, share the same structure of the respective two parts of the previously explained CNN. Interestingly, the middle layer of this network is a recurrent neural network formed by 3 Long Short Term Memory (LSTM) layers and 2 dense ones. Fig. 3 demonstrates the structure of the considered CRNN.

V. Experimental Set-Up & results

To evaluate our models we used 80% of the composed dataset as a training set and the rest of it as a test set. During the subdivision of the corpus, we have taken into consideration the fact that the number of utterances has to be balanced across the previously cited languages.

The statistical measures obtained by submitting the test set to the supervised learning algorithms (see Table II) show that the SVM has been outperformed by the other models. This fact is particularly noticeable from Cohen's kappa coefficients and the recall measures, and it could be related to an overfitting phenomenon that occurred with the SVM approach. The bagging collaborative scheme is the third approach in terms of achieved rates. On the other hand, the implemented deep neural networks achieved the best results. In particular, the incorporation of the recurrent layer to the convolutional network has improved the performance of our deep models. This could indicate that there are not only localized patterns but also temporal structures characterizing the emotional state of anger in a language-agnostic manner. Interestingly the CRNN was able to capture such structures and subsequently identify them in novel speech data (see Table III).

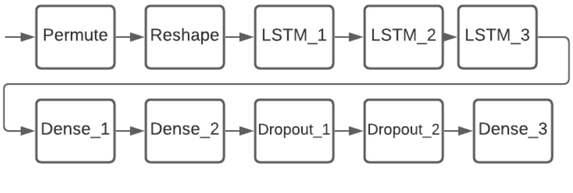

Fig. 3: Recurrent component of the CRNN.

TABLE II: Statistical performance indicators of the considered models when applied on the test-set. The highest rates are emboldened.

Classifiers / ***Accuracy***	**SVM**	**Bagging**	**CNN**	**CRNN**
	80.74%	88.73%	89.41%	**89.77%**
Balanced Accuracy	55.19%	77.55%	83.06	**85.28%**
Precision	**85.33%**	83.61%	75.63%	74.25%
Recall	10.88%	58.16%	72.22%	**77.6%**
F-measure	19.31%	68.60%	73.89%	**75.89%**
Cohen_kappa	0.1525	0.6201	0.6725	**0.6941**

Overall, we argue that the achieved performance is more than satisfactory given the increased task difficulty, and more importantly, it demonstrates the feasibility of cross-language SER frameworks.

VI. Conclusions

This work systematically investigates the feasibility of multilingual anger identification based on speech recordings.

TABLE III: Confusion matrices demonstrating the recognition rates of the considered solutions.

Presented / ***Predicted***	**Not Angry**	**Angry**
Not Angry	2179	11
Angry	524	64

(a) SVM

Presented / ***Predicted***	**Not Angry**	**Angry**
Not Angry	2123	67
Angry	246	342

(b) Bagging

Presented / ***Predicted***	**Not Angry**	**Angry**
Not Angry	2068	134
Angry	160	416

(c) CNN

Presented / ***Predicted***	**Not Angry**	**Angry**
Not Angry	2047	155
Angry	129	447

(d) CRNN

[2] https://github.com/fchollet/keras

[3] www.tensorflow.org

After putting together a suitable corpus encompassing five languages, we experimented with a gamut of features and classification algorithms. It was shown that the solution based on Mel-spectrograms modeled via a CRNN outperformed the rest offering a high recognition rate (89.77%) which potentially imply the existence of temporal structures characterizing the emotional state of anger in a language-agnostic manner.

After such a successful step in multilingual SER, our future work includes a) incorporation of datasets representing additional languages, b) focus on temporal modeling of speech, e.g. by considering Temporal Convolutional Neural Networks [34], and c) addressing the full problem of multilingual SER considering all *big-six* emotional states.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research. This work was carried out within the project entitled Advanced methods for sound and music computing funded by the Piano Sostegno alla Ricerca of University of Milan.

References

[1] B. W. Schuller, "Speech emotion recognition," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018. [Online]. Available: https://doi.org/10.1145/3129340

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020. [Online]. Available: https://doi.org/10.1016/j.specom.2019.12.001

[3] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers in Computer Science*, vol. 2, May 2020.

[4] S. M. Feraru, D. Schuller *et al.*, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[5] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

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

[7] S. Latif, A. Qayyum, M. Usman, and 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.

[8] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[9] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5135–5139.

[10] N. B. Wunarso and Y. E. Soelistio, "Towards indonesian speech-emotion automatic recognition (i-spear)," in *2017 4th International Conference on New Media Studies (CONMEDIA)*, 2017, pp. 98–101.

[11] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5800–5804.

[12] C. Chang and C. Lee, "Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5820–5824.

[13] S. Ntalampiras, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 7–13, Feb. 2020. [Online]. Available: https://doi.org/10.17743/jaes.2019.0045

[14] A. Khalil, W. Al-Khatib, E. El-Alfy, and L. Cheded, "Anger detection in arabic speech dialogs," in *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, 2018, pp. 1–6.

[15] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Anger detection in call center dialogues," in *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2015, pp. 139–144.

[16] J. Pohjalainen and P. Alku, "Automatic detection of anger in telephone speech with robust autoregressive modulation filtering," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7537–7541.

[17] P. Chhabra, G. Vyas, J. Chatterjee, and S. H. Voß, "An automatic system for recognition and assessment of anger using adaptive boost," in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, 2016, pp. 151–154.

[18] J. Deng, F. Eyben, B. Schuller, and F. Burkhardt, "Deep neural networks for anger detection from real life speech data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 1–6.

[19] S. Mohamed, P. Beckett, and M. Lech, "Effect of fixed point computations on anger classification in speech signals," in *2015 IEEE Symposium on Computer Applications Industrial Electronics (ISCAIE)*, 2015, pp. 59–64.

[20] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," in *ICASSP*, 2020, pp. 621–625.

[21] S. Ntalampiras, "Generalized sound recognition in reverberant environments," *JAES*, vol. 67, no. 10, pp. 772–781, Oct. 2019.

[22] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

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

[24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[25] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[26] G. Costantini, I. Iaderola, A. Paoloni, and 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.

[27] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "Shemo: a large-scale validated database for persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, Mar 2019. [Online]. Available: https://doi.org/10.1007/s10579-018-9427-x

[28] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[29] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[30] S. Ntalampiras, "Deep learning of attitude in children's emotional speech," in *2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2020, pp. 1–5.

[31] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

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

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[34] J. Yan, L. Mu, L. Wang, R. Ranjan, and A. Y. Zomaya, "Temporal convolutional networks for the advance prediction of ENSO," *Scientific Reports*, vol. 10, no. 1, May 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-65070-5