A Hierarchical Approach for Multilingual Speech Emotion Recognition

Marco Nicolini* and Stavros Ntalampiras†[a]

LIM – Music Informatics Laboratory, Department of Computer Science, University of Milan, Italy

Keywords: Audio Pattern Recognition, Machine Learning, Transfer Learning, Convolutional Neural Network, YAMNet, Multilingual Speech Emotion Recognition.

Abstract: This article approaches the Speech Emotion Recognition (SER) problem with the focus placed on multilingual settings. The proposed solution consists in a hierarchical scheme the first level of which identifies the speaker's gender and the second level predicts the speaker's emotional state. We elaborate with three classifiers of increased complexity, i.e. *k*-NN, transfer learning based on YAMNet and Bidirectional Long Short-Term Memory neural networks. Importantly, model learning, validation and testing consider the full range of the big-six emotions, while the dataset has been assembled using well-known SER datasets representing six different languages. The obtained results show differences in classifying all data against only female or male data with respect to all classifiers. Interestingly, a-priori genre recognition can boost the overall classification performance.

1 INTRODUCTION

Speech is fundamental in human-machine communication because, among others, it is one of the primary faucets for expressing emotions. In this context, speech emotion recognition (SER) aims at automatically identifying the emotional state of a speaker using her/his voice and, as such, comprises an important branch of Affective Computing, which studies and develops systems sensing the emotional state of a user (Chen et al., 2023).

Emotion plays a vital role in the way we think, react, and behave: it is a central part of decision making, problem-solving, communicating, or even negotiating. Among various applications, emotion recognition plays a crucial part in human health and the related medical procedure to detect, analyze, and determine the medical conditions of a person. For example, SER can be applied to design a medical robot that provides better health-care services for patients by continuously monitoring the patients' emotional state (Park et al., 2009). Other applications of SER technologies could be deploying emotionally-aware Human Computer Interaction solutions (Pavlovic et al., 1997).

The majority of SER solutions are focused on a single language (Ntalampiras, 2021) and only few language-agnostic methods are present in the literature, for instance the work of Saitta (Saitta and Ntalampiras, 2021) or the work of Sharma (Sharma, 2022). A big part of SER research has focused on finding speech features that are indicative of different emotions (Tahon and Devillers, 2015), and a variety of both short-term and long-term features have been proposed. The emotional space is usually organized in six emotions: angry, disgust, happy, sad, neutral and fear (Miller Jr, 2016).

Speech emotions tend to have overlapping features, making it difficult to find the correct classification boundaries. Given the latter, deep learning methods comprise an interesting solution since they can automatically discover the multiple levels of representations in speech signals (Sang et al., 2018) and as such, there is a constantly growing interest in research in applying deep learning based methods to automatically learn useful features from emotional speech data. For instance, Mirsamadi (Mirsamadi et al., 2017) applies recurrent neural networks to automatically discover emotionally relevant features from speech and to classify emotions; Kun Han (Han et al., 2014) applies Deep Neural Network to SER; the work of Scheidwasser et al. establish a framework for evaluating the performance and generalization capacity of different approaches for SER utterances but their method is dependent on the language and train the different models (mainly deep learning methods) on one dataset of the 6 chosen benchmark datasets per time (Scheidwasser-Clow et al., 2022).

The approach in this work focuses on a language-

[a] https://orcid.org/0000-0003-3482-9215

Nicolini, M. and Ntalampiras, S.
A Hierarchical Approach for Multilingual Speech Emotion Recognition.
DOI: 10.5220/0011714800003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 679-685
ISBN: 978-989-758-626-2; ISSN: 2184-4313

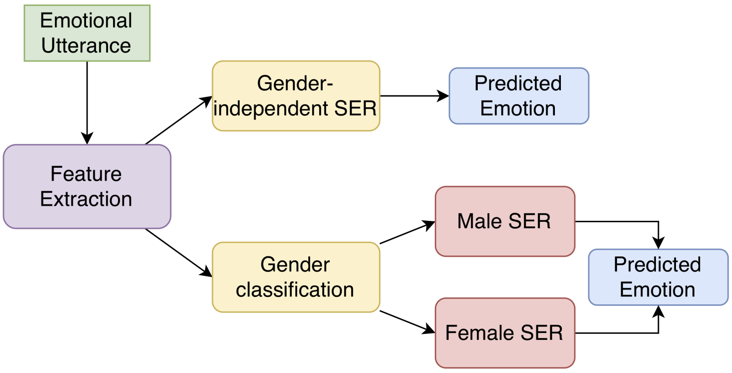

Figure 1: The block diagram of the classification hierarchy adopted in this work.

agnostic methodology for SER: it tries to generalize patterns in data in order to distinguish emotions in every language existing in the employed dataset. A relevant component of the proposed solution is automatic gender differentiation and how it can improve SER performances of the classifiers. Such a direction is analyzed in the work of Vogt et al. (Vogt and André, 2006) where a framework to improve emotion recognition from monolingual speech by making use of automatic gender detection is presented. The work of Dair (Dair et al., 2021) also analyzes differences with and without gender differentiation on three datasets.

Unlike the previous works existing in the related literature, we design a system considering the full range of the big-six emotions as expressed in a multilingual setting with six languages.

The block diagram of the proposed approach is demonstrated in Fig. 1 where the following three classifiers can be observed: a) gender, b) female, and c) male emotional speech. The gender-independent path is shown as well since it has been employed for comparison purposes.

The following classifiers have been employed: a) *k*-Nearest Neighbor Classifier (*k*-NN) which, despite its simplicity, it is a suitable approach for multi-class problems (Hota and Pathak, 2018), b) transfer learning-based classifier based on YAMNet[1], and c) Bidirectional Long Short-Term Memory (BiLSTM) neural network Classifier. The last two classifiers belong to the deep learning domain; the first one relies on a large-scale convolutional network, while the second is able to encode temporal dependencies existing in the available emotional manifestations. Last but not least, every model was trained on appropriate features extracted from time and/or frequency domains.

The rest of this work is organized as follows: section 2 explains the construction of a multilingual dataset facilitating SER purposes. Section 3 briefly describes the employed features and classifiers, while section 4 presents the experimental protocol and obtained results. Finally, in section 5 we draw our conclusions and outline future research directions.

2 CONSTRUCTING THE MULTILINGUAL SER DATASET

SER literature includes various monolingual datasets, thus a corpus combining ten different datasets was formed. More specifically, the following ones have been employed: 1. SAVEE (Vlasenko et al., 2007), 2. CREMA-D (Cao et al., 2014), 3. RAVDESS (Livingstone and Russo, 2018), 4. TESS (Pichora-Fuller and Dupuis, 2020), 5. EMOVO (Costantini et al., 2014), 6. EmoDB (Burkhardt et al., 2005), 7. ShEMO (Nezami et al., 2019), 8. URDU (Latif et al., 2018), 9. JLcorpus (James et al., 2018), and 10. AESDD (Vryzas et al., 2018a; Vryzas et al., 2018b). These datasets include the so-called 'big six' (Miller Jr, 2016) emotions which are typically considered in the related literature. The following six languages are considered in the final dataset: 1. English (New Zealand, British, American, and from different ethnic backgrounds), 2. German, 3. Italian, 4. Urdu, 5. Persian, and 6. Greek.

Table 1 tabulates the duration in seconds for the various classes included in the dataset. It should be mentioned that all datasets include acted speech. We observe only a slight imbalance as regards the genre aspect, where data representing male emotional speech lasts 25460s and female 25081s. Moreover, data associated with the English language comprises the largest part of the dataset followed by Persian, while speech data representing remaining ones (German, Italian, Greek and Urdu) are within the range [998s,2481s].

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

Table 1: Duration (in seconds) of the diverse classes considered in the present work. All values are truncated.

Data part	Angry	Neutral	Sad	Happy	Fear	Disgust
All Data (50541)	10594	11200	9325	7116	5773	6530
Female (25081)	5023	4706	4948	3807	3018	3576
Male (25460)	5570	6493	4377	3309	2754	2953
English (32154)	5475	5464	5772	5174	4765	5502
German (1261)	335	186	251	180	154	154
Italian (1711)	268	261	313	266	283	317
Urdu (998)	248	250	250	250	0	0
Persian (11932)	3832	5037	2175	766	120	0
Greek (2481)	433	0	562	478	449	556

Table 2: Confusion matrix (in %) as regards to gender classification obtained using the YAMNet-based and *k*-NN approaches. The presentation format is the following: YAMNet/*k*-NN. The highest accuracy is emboldened.

Presented \ **Predicted**	*Female*	*Male*
Female	**94.3**/91.1	5.7/8.9
Male	4.8/3.6	95.2/**96.4**

Aiming at a uniform representation, all data has been resampled to 16 kHz and monophonic wave format. When building a SER system, the specific dataset present various challenges, i.e.

- different languages presenting important cultural gaps,
- imbalances at the genre, language, and emotional state levels,
- diverse recording conditions, and
- different recording equipment.

The first two obstacles were addressed by appropriately dividing the data during train, validation, and test phases so that the obtained models are not biased to one or more subpopulations existing within the entire corpus. As regards to the last two, the proposed approach aims at creating a standardized representation of the audio signal so that the effect of recording conditions and equipment is minimized.

To the best of our knowledge, this is the first time in the SER literature that the full range of the big-six emotional states expressed in six different languages is considered.

3 THE CONSIDERED CLASSIFICATION MODELS

This section describes briefly the considered classification models along with their suitably-chosen feature sets characterizing the available audio data. Importantly, each classification model has been separately trained and tested on the following settings: a) the entire dataset, b) female data, and c) male data. Data division in train, validation, and test sets have been kept constant among every classifiers so as to obtain a reliable comparison among the considered models. At the same time, *k*-NN and YAMNet-based models have been trained to distinguish between man and female utterances.

3.1 *k*-NN

The standard version of the *k*-NN classifier has been used with the Euclidean distance as similarity metric. Despite its simplicity, *k*-NN has been able to offer satisfactory performance in SER (Venkata Subbarao et al., 2022), thus we assessed its performance on the present challenging multilingual setting.

Feature Extraction. The short-term features feeding the *k*-NN model are the following: a) zero crossing rate, b) energy, c) energy's entropy, d) spectral centroid and spread, e) spectral entropy, f) spectral flux, g) spectral rolloff, h) MFCCs, i) harmonic ratio, j) fundamental frequency, and k) chroma vectors.

Table 3: Average classification accuracy and balanced accuracy results (in %) using 10 fold evaluation. The presentation format is the following: accuracy/balanced accuracy. The highest accuracy and balanced accuracy are emboldened.

Data subpopulation	***k*-NN**	**YAMNet**	**BiLSTM**
all data	59/59	**74.9**/46.9	62.0/**59.1**
female data	65.2/65.1	**79.9**/52	68.6/**67**
male data	51.6/**51.6**	**71.7**/47.1	56.7/51.3

Table 4: Confusion matrix (in %) as regards to SER classification obtained using the BiLSTM, *k*-NN and YAMNet models approaches with all data. The presentation format is the following: BiLSTM/*k*-NN/YAMNet. The highest accuracy is emboldened.

Pred. **Pres.**	*angry*	*disgust*	*fear*	*happy*	*neutral*	*sad*
angry	**83.1**/68.7/60.9	3.8/8.9/12.1	0.9/3.2/23.8	5.8/13.4/3.2	5.5/4.8/-	0.9/0.9/-
disgust	11.6/5.0/1.4	45.0/64.3/**76.6**	4.3/7.4/19.3	4.6/8.0/2.8	16.8/7.3/-	17.7/8.0/-
fear	10.5/6.0/0.6	6.8/17.6/5.1	41.2/47.4/**94.3**	8.6/10.7/-	7.6/5.2/-	25.3/13.1/-
happy	22.0/13.3/14.8	4.9/14.9/34.4	8.8/8.8/26.2	43.2/**50.2**/24.6	16.1/9.2/-	5.0/3.5/-
neutral	2.8/2.7/-	5.3/12.3/40	1.0/4.6/31.4	3.2/4.6/17.1	**75.6**/67.3/8.6	12.1/8.6/2.9
sad	1.5/3.5/-	3.0/11.5/72	4.0/11.5/12	1.8/3.7/16	24.1/13.6/-	**65.7**/56.2/-

We opted for the mid-term feature extraction process, based short-term features, meaning that mean and standard deviation statistics on these short term features are calculated over mid-term segments. More information on the adopted feature extraction method can be found in (Giannakopoulos and Pikrakis, 2014).

Parameterization. Short- and mid-term window and hop sizes, have been discovered after a series of early experimentations on the various datasets. The configuration offering the highest recognition accuracy is the following: 0.2, 0.1 seconds for short-term window and hop size; and 3.0, 1.5 seconds for mid-term window and hop size respectively. Overall, the both feature extraction levels include a 50% overlap between subsequent windows.

Moreover, parameter k has been chosen using test results based on the ten-fold cross validation scheme; depending on the considered data population, the obtained optimal values range in [5, 21] (see section 4 for more information).

3.2 Transfer Learning Based on YAMNet

YAMNet is a deep neural network model developed by Google and trained on 512 classes of generalized audio events belonging to the AudioSet ontology[2]. As such, the learnt representation may be useful in diverse audio classification tasks including SER. To this end, we elaborated on the Embeddings layer of model and employed it as a feature set which comprises the input to a dense layer, with as many neurons as the classes that have to be classified (2 genres or 6 emotions). The final prediction is based on a softmax layer, while a dictionary of weights can also be employed to compensate cases of imbalanced data classes.

3.3 Bidirectional LSTM

Long short-term memory network is a specific type of recurrent neural network, which is particularly effective in capturing long-term temporal dependencies. Given the fact that audio signal are characterized by their evolution in time, such a property may be significant, thus such a classifier was included in the experimental set-up. More specifically, we considered a bidirectional LSTM (BiLSTM) layer learning bidirectional long-term dependencies in sequential data. BiLSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems (Sajjad et al., 2020). BiLSTMs train two instead of one LSTMs on the input sequence: the first on the input sequence as-is and the second on a reversed copy of the input sequence.

Feature Extraction. In this case, the considered feature sets, able to preserve the temporal evolution of the available emotional manifestations, were the following: a) Gammatone cepstral coefficients (GTCC), b) Delta Gammatone cepstral coefficients (delta GTCC), c) delta-delta MFCC, d) Mel spectrum, and e) spectral crest. The window length has been chosen after extensive experimentations performed on the different data subpopulations, e.g. genre, emotions, etc., while there is no overlapping between subsequent windows.

[2] https://research.google.com/audioset/index.html

Table 5: Confusion matrix (in %) as regards to SER classification obtained using the BiLSTM, *k*-NN and YAMNet models approaches with female data. The presentation format is the following: BiLSTM/*k*-NN/YAMNet. The highest accuracy is emboldened.

Pred. / **Pres.**	*angry*	*disgust*	*fear*	*happy*	*neutral*	*sad*
angry	**85.4**/75.5/63.2	2.4/6.6/16.1	0.9/2.7/19.5	6.6/9.9/1.1	3.6/4/-	1.1/1.3/-
disgust	9.1/7.8/0.8	56.7/67.4/**88.4**	2.0/4.1/10.1	4.7/5.3/0.8	13.5/7/-	13.9/7/-
fear	7.8/6.7/0.4	5.3/9.4/3.1	51.6/58.7/**96.5**	8.2/7.6/-	5.3/11.3/-	21.9/11.3/-
happy	16.2/15.5/14.8	4.5/9.7/59.3	6.5/8.7/25.9	**57.8**/54.8/-	10.1/8/-	4.8/3.3/-
neutral	2.7/3.4/-	8.0/12.1/40	0.5/2.4/50	2.9/2.5/-	**74.8**/71.7/10	11.1/7.8/-
sad	1.8/4.5/-	3.1/7.4/71.4	3.6/8.6/28.6	1.3/2.7/-	14.6/13.9/-	**75.5**/63/-

Table 6: Confusion matrix (in %) as regards to SER classification obtained using the BiLSTM, *k*-NN and YAMNet models approaches with male data. The presentation format is the following: BiLSTM/*k*-NN/YAMNet. The highest accuracy is emboldened.

Pred. / **Pres.**	*angry*	*disgust*	*fear*	*happy*	*neutral*	*sad*
angry	**80.6**/64/58.4	3.8/10.5/11.2	0.7/4.6/25.5	6.9/13.6/5	7.2/6.2/-	0.9/1.1/-
disgust	13.8/4.9/1.9	33.0/55.7/**66.5**	6.5/10.5/27.3	7.2/10.9/4.3	17.4/6.9/-	22.0/11.1/-
fear	14.7/5.7/0.3	6.9/23.8/4.4	30.4/34.6/**95.3**	11.1/11.5/-	9.1/6.5/-	27.7/18/-
happy	26.1/13/11.8	6.1/18.5/23.5	11.5/10.8/23.5	31.4/**41.2**/38.2	19.9/12.3/-	5.0/4.3/2.9
neutral	2.7/2.4/-	5.1/11.6/56	1.9/5.2/28	3.5/5.5/16	**75.8**/65.5/-	11.0/9.6/-
sad	2.2/1.8/-	3.6/14.6/61.1	5.3/13/-	2.5/4.6/33.3	29.6/17.1/-	**56.7**/48.8/5.6

4 EXPERIMENTAL PROTOCOL AND RESULTS

We followed the 10-fold cross evaluation experimental protocol while care was taken so that every classifier operated in identical training, validation and testing folds. The achieved average accuracy with respect to every classifier are summarized in Table 3. Furthermore, Tables 4-6 are the confusion matrices for gender-independent and depended SER. Overall, the following observations can be made.

First, regarding gender discrimination both YAMNet and *k*-NN perform well, while YAMNet offered the highest rate as we see in the respective confusion matrix (Table 2). Interestingly, such results are comparable to the state of art on gender discrimination (Chachadi and Nirmala, 2021).

Second, regarding SER the highest unbalanced accuracy is obtained using the YAMNet model; however, in confusion matrices Tables 4-6, we observe that YAMNet-based classification performs well for specific emotional states, e.g. *fear* and poorly for others, e.g. *happy*.

Third, we focus on the BiLSTM models: they manage to overperform the rest of the considered classifiers, i.e. BiLSTM reaches 62% average accuracy on all the data with per classes accuracy measures that do not fall below 56.1% (Table 4). This may be due to their ability to capture temporal dependencies existing in emotional speech, which are important for speech processing in general (Latif et al., 2022). At the same time, BiLSTM models trained on male or female data provide satisfactory performance (Tables 5, 6) with a balanced accuracy of 62.7%.

Fourth, *k*-NN models results are not very far from the BiLSTM ones, while the associated confusion matrices confirm the interesting capacity of the classifier to distinguish between the various classes. The best k parameter obtained for the all data, female data, male data, gender data models are: *k*=11, *k*=21, *k*=13, *k*=5. These results are in line and confirm the finding that distributed modeling types may in effective in multilingual settings (Ntalampiras, 2020).

Finally, regarding gender-dependent classification: results show a common pattern that can be found in literature (for example in the work of Vogt (Vogt and André, 2006)), i.e. performances improve when female emotional speech is considered (in BiLSTM more than 6%) but not in male (e.g. in BiLSTM 6% worse). Since gender discrimination achieved almost perfect accuracy (more than 94%) the hierarchical classifier combining gender and emotion recognition can improve the overall recognition rate of a gender-independent SER.

In order to enable reliable comparison with other solutions and full reproducibility, the implementation of the experiments presented in this paper is publicly available at https://github.com/NicoRota-0/SER.

5 CONCLUSION AND FUTURE DEVELOPMENTS

In this work, multilingual audio gender-based emotion classification has been analyzed. Importantly, we proposed a SER algorithm offering state-of-art results while considering the full range of the big six emotional states as expressed in six languages. Interestingly, it has been demonstrated that a gender-based emotion classifier can outperform a general emotion classifier.

Future work could assess the performance reached by such modeling architectures on each language separately. Moreover, these models could be part of a more complex system to recognise human emotions that use biosensors measuring physiological parameters, e.g. heart rate, given the accelerated spread of IoT devices as stated in the work of Pal (Pal et al., 2021). Other additional work could investigate the one-vs-all emotion classification scheme using the present models; an example is the work of Saitta et al. (Saitta and Ntalampiras, 2021). An alternative approach would be adding a language classifier before emotion detection (with or without gender detection) to access if it can achieve better results.

REFERENCES

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Chachadi, K. and Nirmala, S. R. (2021). Voice-based gender recognition using neural network. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, pages 741–749. Springer Singapore.

Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W., and Hirota, K. (2023). K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, 70(1):1016–1024.

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

Dair, Z., Donovan, R., and O'Reilly, R. (2021). Linguistic and gender variation in speech emotion recognition using spectral features. *arXiv preprint arXiv:2112.09596*.

Giannakopoulos, T. and Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, Inc., USA, 1st edition.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

Hota, S. and Pathak, S. (2018). KNN classifier based approach for multi-class sentiment analysis of twitter data. *International Journal of Engineering and Technology*, 7(3):1372.

James, J., Tian, L., and Watson, C. I. (2018). An open source emotional speech corpus for human robot interaction applications. In *INTERSPEECH*, pages 2768–2772.

Latif, S., Qayyum, A., Usman, M., and Qadir, J. (2018). Cross lingual speech emotion recognition: Urdu vs. western languages. In *2018 International Conference on Frontiers of Information Technology (FIT)*, pages 88–93. IEEE.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., and Schuller, B. W. (2022). Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition. *IEEE Transactions on Affective Computing*, pages 1–1.

Livingstone, S. R. and Russo, F. A. (2018). 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(5):e0196391.

Miller Jr, H. L. (2016). *The Sage encyclopedia of theory in psychology*. SAGE Publications.

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE.

Nezami, O. M., Lou, P. J., and Karami, M. (2019). Shemo: a large-scale validated database for persian speech emotion detection. *Language Resources and Evaluation*, 53(1):1–16.

Ntalampiras, S. (2020). Toward language-agnostic speech emotion recognition. *Journal of the Audio Engineering Society*, 68(1/2):7–13.

Ntalampiras, S. (2021). Speech emotion recognition via learning analogies. *Pattern Recognition Letters*, 144:21–26.

Pal, S., Mukhopadhyay, S., and Suryadevara, N. (2021). Development and progress in sensors and technologies for human emotion recognition. *Sensors*, 21(16):5554.

Park, J.-S., Kim, J.-H., and Oh, Y.-H. (2009). Feature vector classification based speech emotion recognition for service robots. *IEEE Transactions on Consumer Electronics*, 55(3):1590–1596.

Pavlovic, V., Sharma, R., and Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):677–695.

Pichora-Fuller, M. K. and Dupuis, K. (2020). Toronto emotional speech set (TESS). Scholars Portal Dataverse.

Saitta, A. and Ntalampiras, S. (2021). Language-agnostic speech anger identification. In *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pages 249–253. IEEE.

Sajjad, M., Kwon, S., et al. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875.

Sang, D. V., Cuong, L. T. B., and Ha, P. T. (2018). Discriminative deep feature learning for facial emotion recognition. In *2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Scheidwasser-Clow, N., Kegler, M., Beckmann, P., and Cernak, M. (2022). Serab: A multi-lingual benchmark for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7697–7701. IEEE.

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6907–6911. IEEE.

Tahon, M. and Devillers, L. (2015). Towards a small set of robust acoustic features for emotion recognition: challenges. *IEEE/ACM transactions on audio, speech, and language processing*, 24(1):16–28.

Venkata Subbarao, M., Terlapu, S. K., Geethika, N., and Harika, K. D. (2022). Speech emotion recognition using k-nearest neighbor classifiers. In Shetty D., P. and Shetty, S., editors, *Recent Advances in Artificial Intelligence and Data Engineering*, pages 123–131, Singapore. Springer Singapore.

Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007). Combining frame and turn-level information for robust recognition of emotions within speech. pages 2249–2252.

Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiaion. In *LREC*, pages 1123–1126.

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C. A., and Kalliris, G. (2018a). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, 66(6):457–467.

Vryzas, N., Matsiola, M., Kotsakis, R., Dimoulas, C., and Kalliris, G. (2018b). Subjective evaluation of a speech emotion recognition interaction framework. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, pages 1–7.