Biologically-Inspired Speech Emotion Recognition Using Rate Map Representations: An Application to the ShEMO Persian Speech Database

İlyas Özer[1,2]
[1]*Bandırma Onyedi Eylül University, Dept. of Computer Engineering, Bandırma, Balıkesir, Turkey, iozer@bandirma.edu.tr, ORCID: 0000-0003-2112-5497*
[2] *AINTELIA Artificial Intelligence Technologies Company, Bursa, Turkey, iozer@aintelia.com*

This paper presents an innovative Speech Emotion Recognition (SER) model, inspired by the human auditory system, for analyzing and interpreting emotions in speech. Our proposed model utilizes a rate map representation to encode the spectro-temporal characteristics of auditory nerve activity, closely mimicking the intricate processes of human auditory perception. This model comprises several stages: pre-emphasis of the audio signal, cochlear filtering using a Gammatone Filter bank (GTF), neuromechanical transduction modeled by the Dau inner hair cell model, and the assembly of a rate map representation through integration of responses from each frequency channel. We apply this model to the ShEMO database, an extensive collection of Persian emotional speech, to detect and classify a spectrum of emotions. Our experimental results, obtained using deep learning architectures, demonstrate the effectiveness of the proposed model. We report the highest classification metrics with the Mobilenet architecture, achieving a performance of 71.57% and an F1 score of 51.52%. Overall, this study contributes to the field of speech emotion detection by offering a biologically-inspired model, validated with a substantial dataset, and yielding promising results in emotion classification using advanced machine learning techniques.

Keywords: *Speech Emotion Recognition, Rate Map, Machine Learning.*

1. Introduction

In the realm of human communication, speech serves as a multifaceted channel conveying not only linguistic information but also rich emotional cues that play a pivotal role in interpersonal interactions [1]. Recognizing and understanding emotions in speech is a fundamental aspect of human communication and has garnered substantial attention in both psychological and technological domains. The burgeoning field of SER has emerged as a critical area of study, with far-reaching implications for applications ranging from human-computer interaction to mental health assessment [2, 3].

Emotions, integral to the human experience, manifest in speech through an intricate interplay of prosody, tone, pitch, and other acoustic features. As technology continues to advance, the integration of SER into various domains holds significant promise, offering opportunities to enhance human-machine interactions, augment mental health diagnostics, and revolutionize communication interfaces [4-6].

The recognition of emotions in speech is a complex task, influenced by cultural nuances, individual differences, and contextual factors. Researchers and practitioners in the field grapple with the challenges of developing robust and culturally sensitive models that can accurately decode the emotional content embedded in spoken language. Moreover, the dynamic and context-dependent nature of emotions poses a substantial hurdle, requiring sophisticated methodologies to capture the subtleties and nuances inherent in human emotional expression.

In recent years, significant strides have been made in the development of machine learning and deep learning models for SER. These advancements have paved the way for more accurate and efficient emotion recognition systems,

capable of handling diverse datasets and adapting to the nuances of individual speakers. However, the challenges persist, necessitating ongoing research to address issues such as cross-cultural variability, data scarcity for certain emotions, and the interpretability of complex model architectures [7-11].

Speech emotion detection is a crucial aspect of SER systems, where the primary goal is to identify the emotional state of the speaker from their speech. This process is complex and multidimensional, largely due to the inherently variable and continuous nature of speech, which carries both emotional and informational content. The efficiency and accuracy of SER systems hinge on how well they can interpret these subtleties in speech, which often vary based on physical and environmental factors.

The process of speech emotion identification in SER systems is fundamentally a pattern recognition task, typically involving three core steps: signal preprocessing, feature extraction, and classification. Each step is integral to the system's overall performance and effectiveness.

Signal preprocessing involves preparing the raw speech signal for further analysis. It often includes noise reduction, normalization, and segmentation. The aim is to enhance the quality of the speech signal and to isolate relevant portions for feature extraction, ensuring that subsequent stages are working with clear, focused data.

Perhaps the most crucial step, feature extraction involves identifying and isolating specific characteristics from the speech signal that are relevant to emotional content. Features may include pitch, tone, tempo, and various other speech properties. The challenge here lies in selecting the right features that effectively capture the emotional nuances within the speech. The choice of features significantly impacts the system’s ability to accurately recognize emotions.

Classification is the final step involves classifying the extracted features into different emotional states. This is typically achieved through machine learning algorithms. The choice of algorithm and its tuning are crucial, as the classifier must be able to handle the complexity and variability of speech effectively. It should accurately associate specific patterns in the speech features with corresponding emotional states.

The complexity of speech emotion detection in SER systems is further compounded by the fact that speech signals can vary based on one’s physical state (like health or stress levels) and environmental conditions (like background noise or room acoustics). These factors introduce additional variability and potential for misinterpretation, making the task of feature extraction and classification even more challenging.

In summary, the success of SER systems in speech emotion detection relies heavily on the effective handling of each of these steps. With advancements in signal processing and machine learning algorithms, SER systems continue to evolve, striving for more accurate and reliable emotion detection from speech, which has a wide array of applications in areas like customer service, healthcare, and human-computer interaction.

SER systems, integral in understanding the emotional states from vocal expressions, have evolved significantly with advancements in signal processing and machine learning. This study emphasizes the importance of feature selection and classification algorithms in enhancing the accuracy of SER systems. Speech, a variable-length continuous transmission, encapsulates both emotion and information, with its characteristics often fluctuating due to physical and environmental conditions. These variabilities necessitate the choice of robust features and effective classification methods.

Various studies have investigated different features like prosodic, voice quality, Teager energy operator (TEO), and spectral features for SER tasks. For instance, a study proposed a parallel architecture utilizing prosodic, TEO, and spectral features, leading to significant classification success on multiple datasets using neural networks. Another research integrated paralinguistic features with prosodic ones, employing gender-dependent tests and support vector machines (SVMs) for emotion detection [12-16].

The exploration of time-frequency representations has also yielded promising results in SER. Studies employing 3D log Mel-spectrogram features with convolutional neural networks (CNNs), SVMs, and long short-term memory (LSTM) networks have achieved high classification performance. However, the challenge remains in representing audio signals effectively as images, a critical aspect in using CNNs, widely applied in image classification.

This study focuses on auditory firing rate map features, inspired by the human auditory system. These features encapsulate the spectro-temporal representation of the auditory nerve firing rate, processed through several stages including pre-emphasis, cochlear filtering via the gammatone filter (GTF), and integration using the Hamming window. The GTF, in particular, provides a more precise frequency resolution in the lower-frequency range, advantageous for SER tasks.

Rate maps, typically used in computational auditory scene analysis and speaker recognition, reflect the average energy of each frequency channel. They also offer smoothing on features, thereby reducing variance and enhancing class separability. This aspect is crucial as it allows focusing on regions with concentrated energy, pertinent for emotion detection in the human voice. The emphasis on spectral peaks over valleys can improve SER performance, as peaks are less affected by noise and less correlated, making modeling more realistic.

In conclusion, the auditory firing rate map features, by closely modeling the human auditory system, present a novel and effective approach for emotion detection in speech. These features, focusing on energy concentrations in various frequency channels, provide a robust framework for SER, potentially improving the accuracy and reliability of these systems.

2. Proposed Model

The rate map representation, drawing inspiration from the human auditory system, encodes the spectro-temporal characteristics of auditory nerve activity. This complex encoding process involves several distinct stages, each integral to capturing the intricate dynamics of auditory perception.

Initially, the audio signal undergoes a pre-emphasis stage, where it's filtered to simulate the acoustic modulation performed by the human outer and middle ears. This step enhances certain frequency components of the signal, tailored to mimic the natural resonance characteristics of the ear.

Following this, the signal is subjected to cochlear filtering. This stage is crucial for replicating the cochlea's innate ability to select frequencies. Utilizing a GTF bank, this process emulates the cochlea's frequency-specific responses. Phase compensation within this filter bank is crucial for accurately representing the timing of frequency-specific auditory responses.

The third stage involves neuromechanical transduction, modeled using the Dau inner hair cell model. This stage is pivotal in translating the mechanical vibrations, triggered by sound waves, into neural signals. The cochlea, conceptualized as a frequency analyzer, is instrumental in this process. The model delineates the conversion of basilar membrane movements into nerve impulses, with a focus on the inner hair cell's role in this mechanotransduction.

Finally, the rate map representation is assembled. This involves integrating the responses from each frequency channel. The impressions from inner hair cells are first smoothed, then segmented into overlapping periods, over which a Hamming window is applied. The final step is the averaging of these samples within each segmented time frame, culminating in the creation of the rate map.

Throughout these stages, the Dau model plays a critical role in estimating the likelihood of a spike in auditory nerve firing rates, based on the response of the GTF bank. This model intricately details the neurotransmitter dynamics at the hair cell synapse, offering a nuanced view of auditory nerve activation probability.
Overall, this multi-step process effectively captures the essence of auditory nerve firing patterns, providing a sophisticated and biologically-inspired representation of auditory processing.

3. Persian Speech Emotion Detection with ShEMO Database

The ShEMO database represents a significant stride in research on Persian emotional speech, offering a comprehensive and freely available resource for academic purposes. This database comprises 3000 semi-natural utterances, which

equates to about 3 hours and 25 minutes of speech data, all extracted from online radio plays. It's an extensive collection that captures the nuances of Persian speech across a spectrum of emotions, including anger, fear, happiness, sadness, surprise, and a neutral state [17].

One of the key features of ShEMO is its inclusivity of samples from 87 native-Persian speakers, providing a rich and varied dataset that is essential for robust emotion detection research. Each of these speech samples is meticulously annotated with both orthographic and phonetic transcriptions, following the standards of the International Phonetic Alphabet. This level of detail not only aids in the understanding of emotional content but also facilitates phonetic analysis, which is crucial in speech emotion research.

The process of labeling and categorizing the emotional content in ShEMO involves the efforts of twelve annotators. These individuals are tasked with identifying the underlying emotional state of each utterance. The use of majority voting to finalize the labels ensures that the dataset is marked with a high level of accuracy and consensus. The inter-annotator agreement, measured by the kappa statistic, stands at 64%, indicating a substantial level of agreement and further cementing the reliability of the data.

In addition to providing this rich dataset, the study associated with ShEMO delves into benchmark results using common classification methods in the field of speech emotion detection. These benchmarks are crucial in evaluating the effectiveness of various approaches in accurately categorizing emotions in speech. Remarkably, the Support Vector Machine (SVM) algorithm emerges as the most effective model in this study. It demonstrates superior performance in both gender-independent and gender-dependent models, with an overall gender-independent accuracy of 58.2%. Notably, the accuracy for female voices reaches 59.4%, while for male voices, it is slightly lower at 57.6%.

The ShEMO database is not just a collection of Persian speech samples; it is a pivotal tool in advancing the understanding and technology behind speech emotion recognition. The thoroughness in its creation and annotation, combined with insightful findings on classification methods, particularly the efficacy of SVM, positions it as a valuable asset for researchers and technologists in the field of speech emotion detection.

4. Performance Evaluation Metrics

In SER studies, evaluating the performance of the employed model is crucial for determining its effectiveness and reliability. This study adopts various performance evaluation criteria, including classification accuracy (ACC), sensitivity (SENS), specificity (SPEC), precision (PREC),and F1-Score. Each of these metrics offers insights into different aspects of the model's performance.
Classification accuracy is a straightforward metric that calculates the proportion of correctly classified instances out of all instances. It is given by the formula:

$$Accuracy(N) = \frac{\sum_{i=1}^{|N|} estimate(n_i)}{|N|}, \quad n_i \in N \tag{1}$$

$$Estimate(n) = \begin{cases} 1, & if\ estimate(n) = cn \\ 0, & otherwise \end{cases} \tag{2}$$

$$Classification\ Accuracy(ML) = \frac{\sum_{i=1}^{|k|} Accuracy(N_i)}{|k|} \tag{3}$$

Here, n denotes the number of test data sets, "cn" represents the n value's class, Estimate (n) is the result of the n classification process, and k represents the total number of groups in the data set.

However, in scenarios where the dataset is imbalanced, ACC might not be the most reliable metric as it could be skewed by the majority class. Sensitivity (also known as recall) measures the proportion of actual positive cases that are correctly identified by the model:

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

Where, TP is symbolize true positive predictions and FN means false negative predictions.

Specificity, on the other hand, measures the proportion of actual negative cases that are correctly identified. To represent samples where FP is falsely judged positive, SPEC can be expressed as follows:

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

Precision assesses the proportion of positive identifications that were actually correct:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

The F-Score, or F1 Score, is the harmonic mean of precision and sensitivity. It is particularly useful when seeking a balance between precision and sensitivity:

$$F - Score = \frac{2TP}{2TP + FP + FN} \quad (7)$$

By employing these various performance metrics, the study aims to provide a thorough and nuanced understanding of the model's effectiveness in classifying emotions from speech data. This comprehensive approach ensures that the strengths and limitations of the model are appropriately identified and addressed.

4. Experimental results and discussion

In this study, an SER approach inspired by the human auditory system is proposed. In the proposed model, audio recordings were first denormalized. Preemphasis was applied to denormalized audio recordings. A gammatone filter was then applied, modeling the frequency selection property of the cochlea. The lower frequency value of the filter was set to 50 Hz and the upper frequency was 10000 Hz. The number of channels was used as 128. Afterwards, neuromechanical transduction was performed using the Dau model. The signal at the output of the specific basilar membrane segment was half-wave rectified and low-pass filtered at 1 kHz in the Dau model. In IHC, this stage approximates the translation of mechanical oscillations of the basilar membrane into receptor potentials. For high carrier frequencies, low-pass filtering practically preserves the signal envelope. Feedback loops were used to model the effects of adaptation. This level compresses steady signals virtually logarithmically while converting quick fluctuations in the input more linearly. Following the feedback loops, the signal was low-pass filtered at 8 Hz with a time constant of 20 ms. The results of simulations presented in the companion study Dau et al., 1996 [11] suggested this number. Finally, each frequency channel's rate map is calculated separately. IHC measurements are first adjusted using a leaky integrator for this operation. In this case, the time constant is commonly 8 ms. A Hamming window is then applied to each channel, which is separated into overlapping periods. Finally, each time frame's samples are averaged.

After the extraction of rate map features as mentioned above, the log scale was applied. Following these steps, a threshold value function and image conversion were performed as stated in [11]. 5 different deep learning architectures were used for the classification process. Table 1 shows the results obtained with these architectures. The highest classification metrics were obtained with Mobilenet. In this model, the performance was 71.57% and the F1 score was

51.52%. On the other hand, the lowest results were achieved in the VGG16 model. It is considered that the reason for the better performance of Mobilenet is that the number of parameters is less than other models.

Table 1. *Classification Results.*

	ACC	SENS	SPEC	PREC	F1
Mobilenet	**71.57**	**50.18**	**93.51**	**54.90**	**51.52**
Vgg16	64.97	36.30	91.58	40.59	35.75
Resnet50	69.37	47.65	92.92	54.34	49.49
Xception	69.88	48.72	93.38	50.78	49.28
InceptionV3	67.17	38.74	92.16	43.24	37.26

Table 2 shows the class-based results obtained with Mobilenet. The highest F1 score was obtained for the feeling of anger. The highest F1 score was obtained for the feeling of nervous. On the other hand, the F1 score for fear was 0%. It is considered that the main reason for this is due to the high amount of imbalance in the number of samples in each class in the data set. Machine learning models generally tend to lean toward the majority class. The confusion matrix in Figure 1 shows how many samples of each class are in the test set.

Table 2. *Class Based Classification Results.*

	SENS	SPEC	PREC	F1
Anger	83.74	88.40	79.07	81.34
Fear	0.00	99.31	0.00	0.00
Happiness	33.33	97.64	50.00	40.00
Neutral	77.34	86.08	74.41	75.85
Sadness	70.10	91.09	60.71	65.07
Surprise	36.59	98.55	65.22	46.88

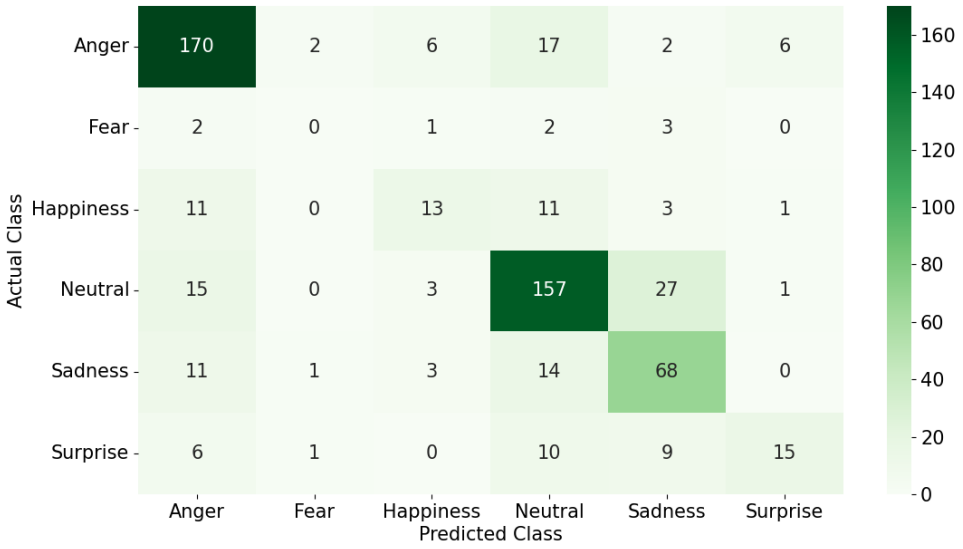

Figure 1*. Confusion Matrix of Mobilenet Model.*

Finally, Figure 2 shows the radar chart of the results obtained with each model. As can be seen from the graph, the best results for all metrics belong to the Mobilenet model.

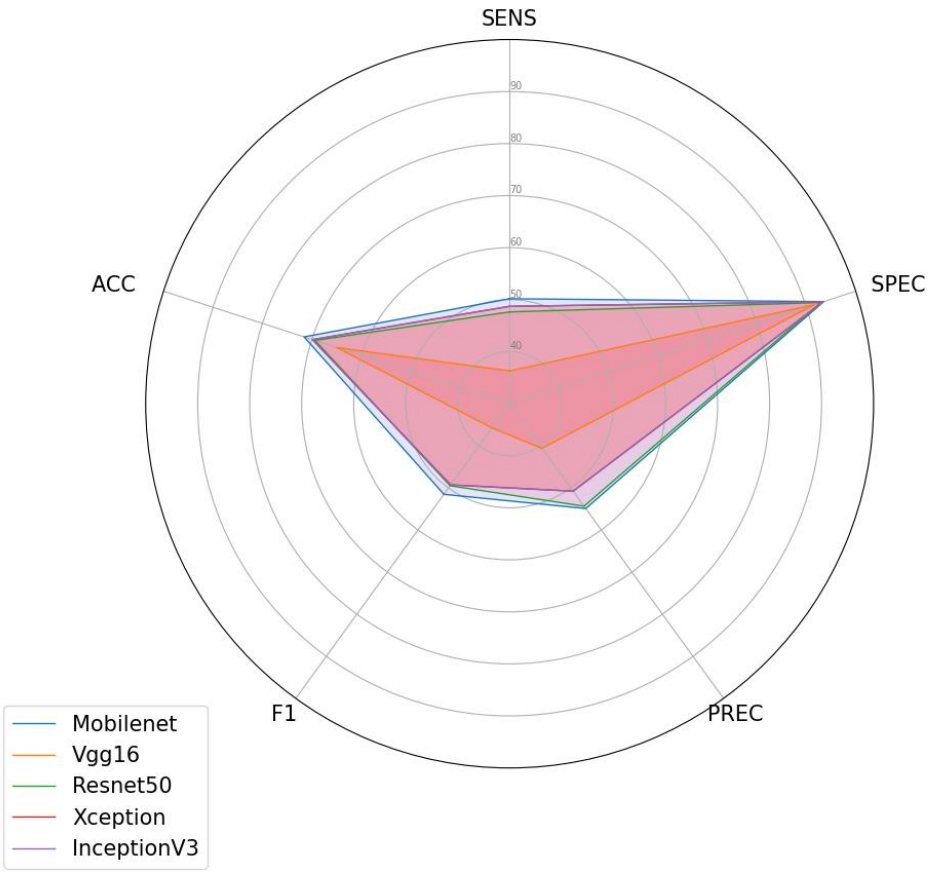

Figure 2. *Radar chart of all models.*

5. Conclusion

This study presents a novel SER model inspired by the human auditory system, demonstrating significant advancements in the field of emotion detection from speech. Our proposed model, through its intricate rate map representation, effectively encodes the spectro-temporal characteristics of auditory nerve activity, closely mirroring the complexity of human auditory perception. The model's multi-stage process, including pre-emphasis, cochlear filtering, neuromechanical transduction, and rate map assembly, ensures a sophisticated and biologically-inspired representation of auditory processing.

Applied to the extensive ShEMO database, our model addresses the nuances of Persian emotional speech, utilizing a dataset that encompasses a wide range of emotions and voices. The thorough annotation process and the substantial inter-annotator agreement highlight the reliability and richness of the ShEMO database, making it an invaluable tool for speech emotion research.

Our experimental results reveal the effectiveness of various deep learning architectures in emotion classification, with the Mobilenet architecture showing the most promise. The superior performance of Mobilenet, in comparison to other models like VGG16, underscores the importance of selecting appropriate model parameters in SER tasks. However, the challenges posed by class imbalances, as evidenced by the disparate F1 scores for different emotions, highlight an area for further research and optimization.

In conclusion, this study contributes to the growing body of knowledge in SER by introducing a biologically-inspired model that showcases the potential of machine learning in understanding and interpreting human emotions through speech. The success of the model on the ShEMO database not only underscores its effectiveness in dealing with diverse

emotional states and speech patterns but also paves the way for further advancements in the field. Future research should aim to address the limitations noted, particularly in dealing with class imbalances, to enhance the accuracy and applicability of SER systems in real-world scenarios.

REFERENCES

[1] Ilyas, Ozer. "Pseudo-colored rate map representation for speech emotion recognition." *Biomedical Signal Processing and Control* 66 (2021): 102502.

[2] Bhavan, Anjali, Pankaj Chauhan, and Rajiv Ratn Shah. "Bagged support vector machines for emotion recognition from speech." *Knowledge-Based Systems* 184 (2019): 104886.

[3] Özseven, Turgut. "A novel feature selection method for speech emotion recognition." *Applied Acoustics* 146 (2019): 320-326..

[4] Sun, Linhui, et al. "Speech emotion recognition based on DNN-decision tree SVM model." *Speech Communication* 115 (2019): 29-37..

[5] Mustafa, Mumtaz Begum, et al. "Speech emotion recognition research: an analysis of research focus." *International Journal of Speech Technology* 21 (2018): 137-156.

[6] Rázuri, Javier G., et al. "Speech emotion recognition in emotional feedbackfor human-robot interaction." *International Journal of Advanced Research in Artificial Intelligence* (IJARAI) 4.2 (2015): 20-27.

[7] Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." *IEEE access* 8 (2020): 79861-79875..

[8] Lu, Guanming, et al. "Speech emotion recognition based on long short-term memory and convolutional neural networks." *Journal of Nanjing University of Posts and Telecommunications* 38.5 (2018): 63-69.

[9] Ozer, Ilyas, Zeynep Ozer, and Oguz Findik. "Noise robust sound event classification with convolutional neural network." *Neurocomputing* 272 (2018): 505-512.

[10] FADEL, Mariem Mine CHEİKH MOHAMED, and Ö. Z. E. R. Zeynep. "Trafikle İlgili Seslerin İşitsel Modeller ve Konvolüsyonel Sinir Ağları Kullanılarak Sınıflandırılması." *Mühendislik Bilimleri ve Araştırmaları Dergisi* 5.2 (2023): 233-242..

[11] Ozer, Ilyas, Zeynep Ozer, and Oguz Findik. "Lanczos kernel based spectrogram image features for sound classification." *Procedia computer science* 111 (2017): 137-144..

[12] Rao, K. Sreenivasa, Shashidhar G. Koolagudi, and Ramu Reddy Vempada. "Emotion recognition from speech using global and local prosodic features." *International journal of speech technology* 16 (2013): 143-160..

[13] Valstar, Michel, et al. "Avec 2016: Depression, mood, and emotion recognition workshop and challenge." *Proceedings of the 6th international workshop on audio/visual emotion challenge.* 2016.

[14] Jiang, Pengxu, et al. "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition." *IEEE Access* 7 (2019): 90368-90377.

[15] Zhao, Jianfeng, Xia Mao, and Lijiang Chen. "Speech emotion recognition using deep 1D & 2D CNN LSTM networks." *Biomedical signal processing and control* 47 (2019): 312-323.

[16] Martin, Olivier, et al. "The eNTERFACE'05 audio-visual emotion database." *22nd international conference on data engineering workshops (ICDEW'06). IEEE*, 2006.

[17] Mohamad Nezami, Omid, Paria Jamshid Lou, and Mansoureh Karami. "ShEMO: a large-scale validated database for Persian speech emotion detection." *Language Resources and Evaluation* 53 (2019): 1-16.