# An Analysis of the Impact of Spectral Contrast Feature in Speech Emotion Recognition

https://doi.org/10.3991/ijes.v9i2.22983

Shreya Kumar, Swarnalaxmi Thiruvenkadam (✉)
Anna University, Chennai, India
thiruvenkadamswarna@gmail.com

**Abstract**—Feature extraction is an integral part of speech emotion recognition. Some emotions become indistinguishable from others due to high resemblance in their features, which results in low prediction accuracy. This paper analyses the impact of spectral contrast feature in increasing the accuracy for such emotions. The RAVDESS dataset has been chosen for this study. The SAVEE dataset, CREMA-D dataset, and JL corpus dataset were also used to test its performance over different English accents. In addition to that, the EmoDB dataset has been used to study its performance in the German language. The use of the Spectral Contrast feature has increased the prediction accuracy in speech emotion recognition systems to a good degree as it performs well in distinguishing emotions with significant differences in arousal levels, and it has been discussed in detail.

**Keywords**—Chroma, Feature extraction, Mel Frequency Cepstral coefficients, Multi-layer perceptron, Spectral contrast

## 1 Introduction

Speech Emotion Recognition (SER) is a progressive area of study and it plays a remarkable role in applications like telecommunications, lie detection, Human-Computer Interface (HCI), etc. The interaction between humans and machines is the most significant application and intensive research on this area has been going on for several years. Displaying emotions through speech is the most distinctive and natural characteristic of humans. Speech Emotion Recognition (SER) can be defined as the extraction of the speaker's emotional state from his or her speech signal. Feature extraction, Feature Selection, and Classification are the three main stages of speech emotion recognition. An extracted feature of a speech signal contains information, which leads to better accuracy and recognition rate. Therefore, the feature extraction algorithms help improve the recognition rate and accuracy.

The speech features extracted for emotion recognition are roughly classified as 1. Acoustic features, 2. Language features like lexical information, 3. Context information like gender, cultural influences, and 4. Hybrid features integrate two or more of the above-mentioned features. Acoustic features which are the most commonly used mainly consist of prosodic features, spectral features, and voice quality features.

Although there is no agreement on the best features for SER it is generally accepted that prosodic features carry most of the emotional information. Prosodic features in combination with spectral and voice quality parameterizations are often used in this field of study. However, prosodic features give a certain reiterative confusion pattern among emotions. They seem to be able to discriminate high arousal emotions (anger, happiness) from low arousal ones (sadness, boredom) easily. But the confusion level for emotions of the same arousal level is very large [7]. Pitch, loudness, and duration are commonly used as prosody features since they express the stress and intonation patterns of spoken language. The relation between prosody and emotion is studied in many works in the literature [2]-[6].

Studies have been performed on harmony features for speech emotion recognition. It has been found that the first and second-order differences of harmony features also play an important role in speech emotion recognition and harmony features have been used in SER in [15]. It has been studied that Fourier parameter (FP) features are effective in identifying various emotional states in speech signals [14]. Spectral features (such as spectrum centroid, spectrum cut-off frequency, correlation density, and Mel-frequency energy, etc.), and cepstral features (such as Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Prediction (PLP), etc.) [13] have been explored and used in SER. The SER was also carried out by combining acoustic features with phonological representations [1].

Although there are various feature extraction algorithms and techniques, there is no common agreement on what group of features should be considered for better classification. In this paper, it is analyzed that the spectral contrast feature combined with MFCC, Mel, and Chroma, increased the accuracy of predicting the emotions using K-Nearest Neighbor, Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting Classifier algorithms compared to utilizing only MFCC, Mel, and Chroma. However, it provided lesser accuracy on Random Forest and Multi-Layer Perceptron classifiers. It is also noted that when the spectral contrast feature is included with MFCC, Mel, and Chroma, it increased the accuracy significantly between pairs of emotions with different arousal levels. This could be further utilized by applications where it is required to detect high levels of anxiety and depression of a person instantly especially in Helpline services.

## 2 Related Work

There are various deep learning methods identified previously to classify emotions and are being improved and applied to increase the accuracy of identifying and classifying these emotions. In [19], characteristics of the emotions are learned from the low-level speech signals and the emotional intensity of words is indicated. For instance, words like “awesome” are meant to carry stronger emotions than words like “human”. Features like MFCC, Mel-frequency, and prosodic features are extracted from the data. This multimodal system’s performance could further be improved when more features of the data are taken into account.

In this paper [21], the speaker is recognized in the emotional environment. Spectral features are extracted from the data and are classified. Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Gaussian Naive Bayes, K-Nearest Neighbor, Random Forest, and a simple Neural Network using Keras are all used for classification. Feature combinations are used to improve accuracy in classification. Different spectral features (Mel Frequency Cepstral Coefficients (MFCC), Shifted Delta Cepstral Coefficients (SDCC), etc.) were analyzed and the feature combination that was considered contributed the highest accuracy of 100% in a neutral environment and 87.0967% in an emotional environment].

In this paper [22], the speech emotion is recognized using multi-scale area attention by taking Log Mel Spectrogram into consideration and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset was augmented with vocal tract length perturbation (VTLP). With area attention and VLTP based data augmentation, an accuracy of 77.5% was achieved. This paper [23] considers the semantic information and paralinguistic information in the signals. The unified feature vector is used to make the final prediction using LSTM.

Different features have been used for SER systems, however, there is no such specific set of features for the classification. This is an area of progressive study. In this survey [20], the global and local features of the SER systems that were analyzed are Prosodic features, Voice Quality features, Spectral features, and Teager Energy Operator (TEO) based features. It is also noted that the consonant regions of a speech signal include more emotional information, compared to utterances that involve vowels. After feature extraction, these SER systems can select from a wide range of classification algorithms and methods. Selections and extractions of other features could improve the recognition rates of the SER systems.

# 3 Methodology

## 3.1 Block diagram

The detailed project flow and architecture diagram has been shown in Fig 1, Fig 2. The acoustic features extracted from the sound files of the datasets are used to train the multi-layer perceptron to predict emotions. The dataset is split into 2 parts: 80% as the training data and 20% as the test data. In real-time, the features selected from the received speech input are used to recognize the emotion. The simulation is done on Jupyter Notebook and majorly uses Keras library and Librosa package from python. The dataset, feature extraction, and classification methods are discussed below.

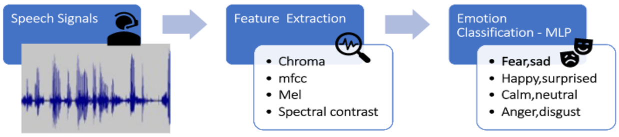


**Fig. 1.** Overview of project flow

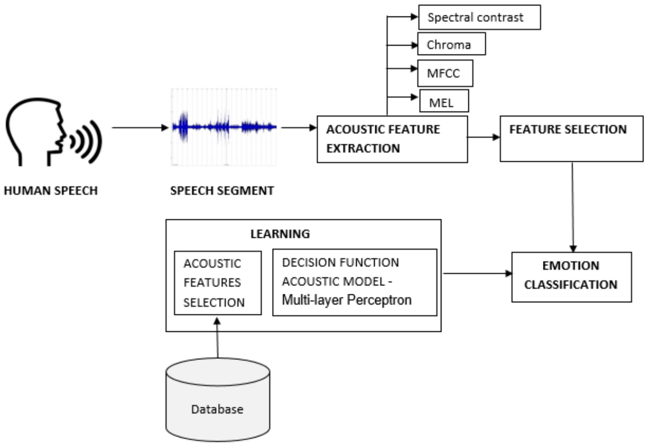

Spectral contrast
Chroma
MFCC
MEL
HUMAN SPEECH
SPEECH SEGMENT
ACOUSTIC FEATURE EXTRACTION
FEATURE SELECTION
LEARNING
ACOUSTIC FEATURES SELECTION
DECISION FUNCTION ACOUSTIC MODEL - Multi-layer Perceptron
EMOTION CLASSIFICATION
Database

**Fig. 2.** Block Diagram

### 3.2 Dataset

The Ryerson Audio-Visual Database of Emotional Speech and song dataset [12] used in the speech emotion recognition has 1440 sound files which contain 24 professional actors (12 female, 12 male) vocalizing two lexically matched statements (1. “Kids are talking by the door”, and 2. “Dogs are sitting by the door”) in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal and strong) with an additional neutral expression.

### 3.3 Feature extraction

The feature extraction part involves extracting the following features from the sound file.

**MFCC**: MFCC was first introduced and applied to speech emotion recognition in [8]. Mel frequency Cepstral coefficients are coefficients that collectively make up MFCC, which is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. MFCC is derived by taking the Fourier transform of a windowed signal, mapping the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows. Followed by taking logs of the powers at each of the Mel frequencies then taking discrete cosine transform of the list of Mel log powers, as if it were a signal. The MFCCs are amplitudes of the resulting spectrum.

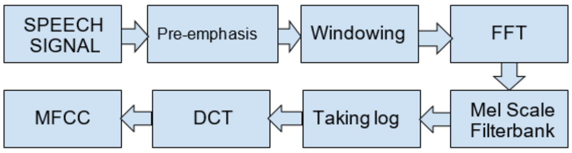


**Fig. 3.** MFCC Derivation

**Mel**: Mel spectrogram is a spectrogram with Mel scale as its y-axis. It is derived by sampling the input with windows, making hops each time to sample the next window then computing fast Fourier transform and generating a Mel scale by taking the entire frequency spectrum and separating it into evenly spaced frequencies.

**Chroma**: The term Chroma closely relates to the 12 different pitch classes. One main property of Chroma features is that they capture the harmonic and melodic characteristics of music while being robust to changes in timbre and instrumentation.

**Spectral Contrast**: Spectral contrast considers the spectral peak, the spectral valley, and their difference in each frequency sub-band. This feature includes more spectral information than MFCC. It represents the relative spectral distribution instead of the average spectral envelope.

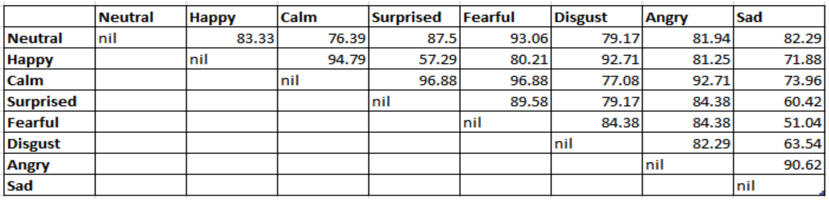
### 3.4 Classification method

The classifier used for the study is Multilayer perceptron which is a class of feed-forward artificial neural networks that refers to networks composed of multiple layers of the perceptron. The MLP model used consists of an input layer, three hidden layers with hidden layer sizes (300, 100, 50), and an output layer. Except for the input node, each node is a neuron that uses the Rectifier Linear unit (ReLu) as an activation function and Adam optimizer for weight optimization. It uses a supervised learning technique called backpropagation for training.

## 4 Experimental Results

The most commonly used features for Speech Emotion Recognition are MFCC, MEL, and chroma. The results of using spectral contrast as a feature along with MFCC, MEL, and chroma are compared with the accuracies obtained from excluding the spectral contrast on the RAVDESS dataset on different machine learning algorithms. The usage of spectral contrast feature for speech emotion recognition for the RAVDESS dataset has increased the accuracy for 8 pairs of emotions; neutral and happy, happy and surprised, sad and fearful, sad and surprised, angry and fear, disgust and fear, fear and surprise, and sad and calm which have considerable differences in their arousal levels.

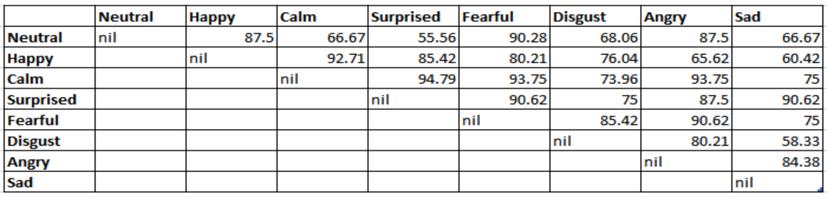
**Table 1.** Accuracies between pairs of emotions before using the spectral contrast feature



| | Neutral | Happy | Calm | Surprised | Fearful | Disgust | Angry | Sad |
|---|---|---|---|---|---|---|---|---|
| Neutral | nil | 83.33 | 76.39 | 87.5 | 93.06 | 79.17 | 81.94 | 82.29 |
| Happy | | nil | 94.79 | 57.29 | 80.21 | 92.71 | 81.25 | 71.88 |
| Calm | | | nil | 96.88 | 96.88 | 77.08 | 92.71 | 73.96 |
| Surprised | | | | nil | 89.58 | 79.17 | 84.38 | 60.42 |
| Fearful | | | | | nil | 84.38 | 84.38 | 51.04 |
| Disgust | | | | | | nil | 82.29 | 63.54 |
| Angry | | | | | | | nil | 90.62 |
| Sad | | | | | | | | nil |

The comparisons of accuracies before and after using the spectral contrast feature for each pair of emotions in the RAVDESS dataset can be seen in Table 1 and Table 2 respectively.

**Table 2.** Accuracies between pairs of emotions after using the spectral contrast feature



| | Neutral | Happy | Calm | Surprised | Fearful | Disgust | Angry | Sad |
|---|---|---|---|---|---|---|---|---|
| Neutral | nil | 87.5 | 66.67 | 55.56 | 90.28 | 68.06 | 87.5 | 66.67 |
| Happy | | nil | 92.71 | 85.42 | 80.21 | 76.04 | 65.62 | 60.42 |
| Calm | | | nil | 94.79 | 93.75 | 73.96 | 93.75 | 75 |
| Surprised | | | | nil | 90.62 | 75 | 87.5 | 90.62 |
| Fearful | | | | | nil | 85.42 | 90.62 | 75 |
| Disgust | | | | | | nil | 80.21 | 58.33 |
| Angry | | | | | | | nil | 84.38 |
| Sad | | | | | | | | nil |

Using Multi-layer perceptron, the overall accuracy for all the eight emotions with spectral contrast feature was found to be 45%, and without spectral contrast, it was 56.39%. With logistic regression, the overall accuracy has increased to 51.67% with spectral contrast feature, from 46.11% without the spectral contrast feature. Similarly, the gradient boosting classifier gives an accuracy of 52.78%, while without spectral contrast feature it gives 51.11%, thus increasing the overall accuracy using the spectral contrast feature. Table 4 and Table 5 show the accuracy comparison between pairs of emotions from different datasets that have been used in this study.

## 5 Discussion

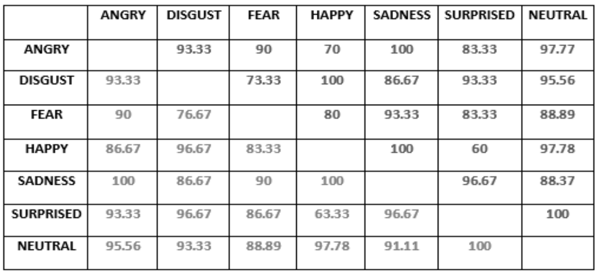
To evaluate the performance of the system, we also used datasets with different accents of the English language. English is spoken in many accents worldwide. The analysis was carried out on the datasets SAVEE [10] for British accent, JL-corpus for New Zealand accent, and Crema-d which has a collection of accents like Caucasian and African-American. It was found that the overall accuracy for the SAVEE dataset increased from 65% to 70%, and overall accuracy decreased for Crema-d [11] from 46.64% to 44.8% while using spectral contrast. In addition to English language datasets, we used the German Language dataset EmoDB [9] to analyze the performance of this system for different languages.

**Table 3.** Accuracy for other classification algorithms with and without spectral contrast feature

| | **Without Spectral contrast feature** | **With Spectral contrast feature** |
|---|---|---|
| KNN | 47.50% | 47.78% |
| Logistic Regression | 46.11% | 51.67% |
| Random forest | 44.44% | 43.33% |
| SVM - Linear | 29.17% | 22.50% |
| SVM (Random State 2, gamma auto, RBF) | 47.48% | 48.61% |
| Gradient Boosting classifier | 51.11% | 52.78% |
| MLP | 56.39% | 45% |

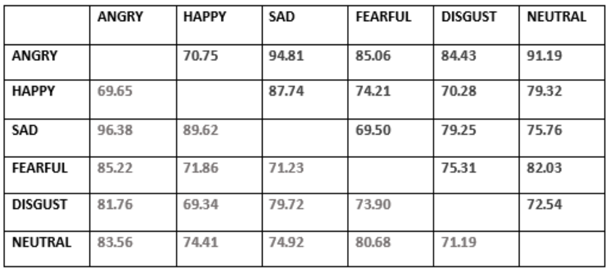
The overall accuracy for EmoDB decreased from 71.64% to 68.66% while using spectral contrast.

**Table 4.** Accuracies between pairs of emotions in SAVEE dataset with spectral contrast (upper-half triangle) and without spectral contrast (lower-half triangle)



| | ANGRY | DISGUST | FEAR | HAPPY | SADNESS | SURPRISED | NEUTRAL |
|---|---|---|---|---|---|---|---|
| ANGRY | | 93.33 | 90 | 70 | 100 | 83.33 | 97.77 |
| DISGUST | 93.33 | | 73.33 | 100 | 86.67 | 93.33 | 95.56 |
| FEAR | 90 | 76.67 | | 80 | 93.33 | 83.33 | 88.89 |
| HAPPY | 86.67 | 96.67 | 83.33 | | 100 | 60 | 97.78 |
| SADNESS | 100 | 86.67 | 90 | 100 | | 96.67 | 88.37 |
| SURPRISED | 93.33 | 96.67 | 86.67 | 63.33 | 96.67 | | 100 |
| NEUTRAL | 95.56 | 93.33 | 88.89 | 97.78 | 91.11 | 100 | |

**Table 5.** Accuracies between pairs of emotions in Crema-d dataset with spectral contrast (upper-half triangle) and without spectral contrast (lower-half triangle)



| | ANGRY | HAPPY | SAD | FEARFUL | DISGUST | NEUTRAL |
|---|---|---|---|---|---|---|
| ANGRY | | 70.75 | 94.81 | 85.06 | 84.43 | 91.19 |
| HAPPY | 69.65 | | 87.74 | 74.21 | 70.28 | 79.32 |
| SAD | 96.38 | 89.62 | | 69.50 | 79.25 | 75.76 |
| FEARFUL | 85.22 | 71.86 | 71.23 | | 75.31 | 82.03 |
| DISGUST | 81.76 | 69.34 | 79.72 | 73.90 | | 72.54 |
| NEUTRAL | 83.56 | 74.41 | 74.92 | 80.68 | 71.19 | |

## 6 Conclusion

The impact of the Spectral contrast feature on Speech Emotion Recognition has provided significant results. Spectral contrast, when combined with Mel, MFCC, and chroma decreased the overall accuracy in the RAVDESS dataset. However, when we tried to classify emotions with relatively different arousal levels it proved significant. Higher accuracies were obtained in detecting certain pairs of emotions after including the Spectral contrast feature in the system. The results on different datasets of the English language suggest that it works across different accents of the same language.

This can also be further combined with language features and phonological representations to improve the accuracy. The modified Feed Forward Neural Network with Particle Swarm Optimization via using Euclidean Distance (FNNPSOED) [16] can be used to handle the classification problem instead of the multilayer perceptron model. The impact of spectral contrast can be studied in the gender identification systems to be used for health-related communications [17]. The Speech Emotion Recognition system which uses the spectral contrast feature can also be extended to identify stress and depression in individuals as the level of arousal between a normal speech and a speech of a depressed person will have a prominent difference and can be used to provide recommendations for metacognition to control the stress levels [18].

## 7 References

[1] W. Wang et al. (2020) ‘Significance of Phonological Features in Speech Emotion Recognition’, *International Journal of Speech Technology*.

[2] ]F. Burkhardt and W. F. Sendlmeier (2000) ‘Verification of acoustical correlates of emotional speech using formant-synthesis’, *Proc. ISCA Tutorial and Research Workshop Speech and Emotion*, Belfast, Ireland, pp. 151–156.

[3] C. F. Huang and M. Akagi (2008) ‘*A three-layered model for expressive speech perception’, Speech Commun.*, vol. 50, pp. 810–828. https://doi.org/10.1016/j.specom.2008.05.017

[4] K. R. Scherer et al., (1991) ‘Vocal cues in emotion encoding and decoding’, *Motiv. Emotion*, vol. 15, no. 2, pp.123–148. https://doi.org/10.1007/bf00995674

[5] M. Schröder (2003) ‘Speech and emotion research’, Ph.D. disservation, Universität des Saarlandes, Saarbrücken, Germany.

[6] E. Navas et al., (2004) ‘Acoustic analysis of emotional speech in standard Basque for emotion recognition’, *Progress in Pattern Recognition, Image Analysis and Applications, ser. Lecture Notes in Computer Science*. Berlin,Germany: Springer, vol. 3287, pp. 386–393. https://doi.org/10.1007/978-3-540-30463-0_48

[7] K. R. Scherer. (2003) ‘Vocal communication of emotion: A review of research paradigms’, *Speech Commun.*, vol. 40, pp. 227–256. https://doi.org/10.1016/s0167-6393(02)00084-5

[8] S. Davis and P. Mermelstein (1980) ‘Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences’, *IEEE Trans. Acoust.*, Speech Signal Process., vol. 28, no. 4, pp. 357–366. https://doi.org/10.1109/tassp.1980.1163420

[9] F. Burkhardt et al., (2005) ‘A database of german emotional speech’, *Interspeech*, pp.1517–1520

[10] Jackson, P. and Haq, S. Surrey (2014) ‘Audio-Visual Expressed Emotion (SAVEE) Database’, University of Surrey: Guildford, UK.

[11] H. Cao et al. (2014) ‘CREMA-D: crowd-sourced emotional multimodal actors dataset’, *IEEE Trans. Affective Computing*, 5(4):377–390. https://doi.org/10.1109/taffc.2014.2336244

[12] Livingstone SR and Russo FA. (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. https://doi.org/10.1371/journal.pone.0196391

[13] T.L. Pao et al. (2006) ‘Mandarin Emotional Speech Recognition based on SVM and NN’, *Proc. International Conference on Pattern Recognition*, vol. 1, pp. 1096-1100. https://doi.org/10.1109/icpr.2006.780

[14] Kunxia Wang et at. (2015) ‘Speech Emotion Recognition Using Fourier Parameters’, *IEEE Transaction on Affective Computing*, Vol 6, No 1.

[15] Bin Yang and M.Lugger (2010) ‘Emotion recognition from speech signals using new harmony features’, Signal Processing, Volume 90, Issue 5. https://doi.org/10.1016/j.sigpro.2009.09.009

[16] Asia L. Jabar, Tarik A. Rashid (2018) ‘A Modified Particle Swarm Optimization with Neural Network via Euclidean Distance’, International Journal of Recent Contributions from Engineering, Science & IT (iJES). https://doi.org/10.3991/ijes.v6i1.8080

[17] Kawther A. Al-Dhlan (2017) ‘Speech Synthesis for Gender Classification’, International Journal of Recent Contributions from Engineering, Science & IT (iJES).

[18] Athanosios Drigas, Eleni Mitsea (2021) ‘Metacognition, Stress - Relaxation Balance & Related Hormones’, International Journal of Recent Contributions from Engineering, Science & IT (iJES). https://doi.org/10.3991/ijes.v9i1.19623

[19] S. Yoon et al. (2018) ‘Multimodal Speech Emotion Recognition using Audio and Text’, 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112-118. https://doi.org/10.1109/slt.2018.8639583

[20] M. Akçay and K. Oğuz (2020) ‘Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers’, Speech Communication, Volume 116, Pages 56-76. https://doi.org/10.1016/j.specom.2019.12.001

[21] Sandhya P. et al. (2020) ‘Spectral Features for Emotional Speaker Recognition’, 2020 Third International Conference on Advances in Electronics, Computers, and Communications (ICAECC), pp. 1-6 https://doi.org/10.1109/icaecc50550.2020.9339502

[22] M. Xu et al., (2021), 'Speech Emotion Recognition with Multi Scale Area Attention and Data Augmentation', arXiv preprint arXiv:2102.01813v1, 2021.

[23] P. Tzirakis et al. (2021), ' Speech Emotion Recognition Using Semantic Information', arXiv preprint arXiv:2103.02993v1, 2021.

## 8 Authors

**Swarnalaxmi Thiruvenkadam** is a Computer Science Engineering student at College of Engineering Guindy, Anna University, India (email: thiruvenkadamswarna@gmail.com).

**Shreya Kumar** is a Computer Science Engineering student at College of Engineering Guindy, Anna University, India (email: shreyakumar603@gmail.com).

Article submitted 2021-03-30. Resubmitted 2021-05-06. Final acceptance 2021-05-08. Final version published as submitted by the authors.