International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Aug-2023)

http://dx.doi.org/10.12785/ijcds/140133

Analysis on the Impact of Lombard Effect on Speech Emotions using Machine Learning

Indirapriyadarshini A[1], Mahima S[1], Shahina A[1], Uma Maheswari[1] and A. Nayeemulla Khan[2]

[1]*Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, 603110, India*
[2]*School of Computer Science Engineering, Vellore Institute of Technology, Chennai, 600127, India*

Received 08 Aug. 2022, Revised 09 May. 2023, Accepted 16 May. 2023, Published 01 Aug. 2023

Abstract: The speech signal is a one-dimensional function of time that originates from the mouth, nose, and cheeks of the speaker. Lombard Speech, on the other hand, is the speech spoken under the influence of background noise. Speakers automatically alter their method of speaking to improve voice clarity when speaking in a noisy setting. It is anticipated that the Lombard effect, which arises as a result of this adaptation, will significantly affect the effectiveness of automatic speech recognition systems that were not designed to account for it. In this study, we contrast the emotions in speech under regular circumstances with the emotions in Lombard speech and look at how those two types of communication differ from one another. Since there haven't been many theories that discuss employing Machine Learning or Deep Learning to account for the Lombard effect in speech emotion recognition systems, we aim to study how Lombard effect impacts Speech Emotion Recognition systems that use different Machine Learning or Deep Learning models. This study will help services adapt to the emotional state of customers accordingly. People commonly alter their speech production in response to their surroundings (i.e., Lombard Effect). In this paper, we review and compare a number of speech emotion recognition algorithms for both regular audio recordings and those recorded with an induced Lombard effect. The whole speech dataset, comprising audios from multiple different speakers was created and populated by the authors, and the audio features were extracted using the MelFrequency Cepstral Coefficients feature extraction approach. The machine learning models were trained on a speech dataset recorded in a laboratory and then tested using a different dataset with Lombard speech. The model is also trained using a Convolution Neural Network, as part of the experiment. In comparison to other traditional Machine Learning models, the Convolutional Neural Network model produces higher accuracy for Speech Emotion Recognition, as seen in the results. The accuracies achieved using Lombard speech data are also found to be much lower than those obtained using normal speech data. This can happen because, when speakers perceive external noise, they tend to speak louder to convey the same message with more emotion than they would in a typical, undisturbed setting.

Keywords: Speech emotion recognition, Lombard effect, Convolution Neural Network, Machine Learning

1. INTRODUCTION

The task of recognizing emotions from speech signals is known as speech emotion recognition(SER), and it is critical in the advancement of human-computer interaction [1]. SER has begun to benefit from the capabilities made available by deep learning in order to solve all main challenges in machine learning.

Many machine learning algorithms have been developed and improved throughout the last two decades of research focusing on automatic emotion recognition. Automatic SER aids smart speakers and virtual assistants with a better understanding of their users, particularly when they recognize words with ambiguous meanings. Emotion recognition is applied in numerous applications like public address systems that usually work in real time [2]. Voice portals or contact centres may use anger detection as a quality indicator. It will help existing services adapt to the emotional state of customers accordingly. In civil aviation, monitoring the stress of aircraft pilots can help reduce the rate of possible aircraft accidents. A chatbot is another good application to improve the quality of conversation.

Prior to utilizing deep learning extensively, SER relied on approaches such as Hidden Markov models (HMM), Gaussian mixture models (GMM), and Support vector machines (SVM), as well as substantial pre-processing and accurate feature engineering. In controlled situations however, the findings are increasing from roughly 70% accuracy to 90%, thanks to deep learning, which makes up most of the new research[3]. Speech signals are frequently affected by many types of noise in real-world situations.

E-mail address: indirapriyadarshinia18038@it.ssn.edu.in, mahimamahima18056@it.ssn.edu.in, shahinaa@ssn.edu.in, UmamaheswariS@ssn.edu.in, nayeemulla.khan@vit.ac.in

To communicate effectively in noisy surroundings, speakers frequently modify their speaking styles involuntarily. This type of adaptation is termed as Lombard effect. To the best of our knowledge, there hasn't been a published study that evaluates the influence of the Lombard effect on speech-emotion recognition using Machine Learning techniques. In this paper, we study how the Lombard effect influences the emotional speech of the speaker by comparing the results of statistical classifications and pitch and intensity analysis. For this purpose, we recorded an English emotional database that contains speeches recorded in four emotions: neutral, happy, sad and angry, by using the same corpus for each speaker. The database was recorded in a quiet laboratory environment and the Lombard effect was simulated by making the speakers listen to babble noise in a cafe environment through a pair of headphones.

Models such as KNN, MLP, Random Forest and other classifiers are used to evaluate the dataset. As the next step, Convolutional Neural Network was used to improve the accuracy of the classification of emotions. A significant difference was found between the accuracy of normal speech and Lombard speech to depict the influence of Lombard effect on speech emotions[4].

In this paper, we begin by outlining the existing research done on Lombard speech and Speech Emotion Recognition(SER) in section 2(A). We then talk about how SER systems work in section 2(B). The next subsection 2(C) defines some of the Emotional Speech Features, and the final part 2(D) explains all the classification models used in this project. Section 3(A) consists of the Proposed system design, 3(B) demonstrates the dataset collection process used by the authors, and the analysis of the said dataset's features in terms of pitch and intensity contours is discussed in Section 3(C). The last subsection 3(D) breaks down the implementation of the workflow for this experiment, and how various ML models and CNN have been used for the same. The results obtained by using different algorithms are then consolidated and compared in section 4 and this research is concluded in section 5 while also mentioning its possible uses in the future.

2. SPEECH EMOTION RECOGNITION IN NOISY ENVIRONMENT: OVERVIEW

A. *Related Work*

Lin et al. [5] focus on classifying 5 emotions using SVM and HMM. MFCC was used for feature selection and the best feature vector was then determined by using the Sequential forward selection (SFS) method. Gender-dependent and independent classification methods were conducted using the DES dataset. The HMM classifier showed a recognition rate of 99.5% and the SVM classifier along with the feature vector selected gave classification rates of 88.9%. In their paper, F. Burkhadt et al. [6] reports on the performance of pitch, energy and duration in a voice portal and the detection of anger. From their research, they discovered that clear anger expression rarely appears in real-world data compared to laboratory data. Parameters such as number of no-matches (NNMs), number of anger turns, number of turns, last turn anger, and last turn NM were used to classify the emotion using Gaussian Mixture models. S. Casale et al. [7] studied the performance of results of a system for emotion classification using the architecture of a Distributed Speech Recognition System (DSR). They tried to improve the performance of the classification system using feature discretization and normalization techniques. Their study showed that the highest performance was obtained using a Support Vector Machine (SVM) trained with the Sequential Minimal Optimization (SMO) algorithm.

Ashish Tawari et al. [8] reviewed the effects of acoustic conditions on the recognition of emotion using speech, in controlled and noisy environments. They also presented a framework with adaptive noise cancellation as a front-end to the speech emotion recognizer. They used the Berlin Database of Emotional Speech (EMO-DB) in which extraneous noise sources can be eliminated and depicts a more realistic situation. The database was collected in a stationary and moving car environment and they concluded that the highest percentage of accuracy can be achieved by using a fusion of GMM and HMM-based classifiers. Their research suggested that the auditory environment has a significant impact on the identification of emotions in speech signals. The review survey conducted by K.S Rao et.al emphasises the normalisation of signals, a preprocessing step that is carried out before the extraction of the features, in addition to a comparative analysis of the databases, classifiers, and features in the SER systems. Additionally, they offers SER system application areas that are not SER technologies but have an impact on them in terms of needs and design. Assessment and comparison of various solutions is a challenge because there isn't currently a sufficiently extensive labelled public corpus of emotional speech. [9] The impact of white noise on an emotion recognition system was studied by Chengwei Huang et al. [10] in 2013. The emotion recognition system is evaluated using both the emotion class model and the dimension space model. GMM is used for speaker and language identification. Clean speech was mixed with Additive White Gaussian Noise (AWGN) and two speech enhancement techniques were used to improve emotion classification. They concluded that the classification rate varies inversely with noise levels, and positive emotions have a higher chance of getting classified as negative emotions as white noise increases.

Huang et al. [11] proposed methods to recognize speech emotions using deep learning techniques such as CNN. The CNN consisted of two steps that would learn both the candidate features and the salient features. After comparing the accuracy and standard deviation with speaker variation and with environmental distortions they suggested that a semi-CNN achieves the highest and most stable accuracy. Farah Chenchah et al. [12] worked on recognizing speech emotions in a noisy real-world environment. They enhanced

the corrupted signal by reducing noise using speech enhancement and performed feature extraction by adopting MFCC (Mel frequency cepstral coefficients). HMM was used to predict the results. Spectral Subtraction Method, Wiener Filter and MMSE are the three speech enhancement methods used. They pointed out that in noisy surroundings, the recognition rate is always lower than in a clean one.

In their paper, Lim et al. [13] proposed using deep learning algorithms such as RNN, LSTM and CNN without using any traditional hand-crafted features. The authors used SFTF for 2D representation of the audio signals. The precision, recall and f1-score of the models were compared and they verified that CNN-based time distributed networks showed better results. Fatemeh et. al. [14] studied and weighed up three classifiers, that included Decision tree, Random Forest Classifier, Adaboost and multi-class SVM, against each other, in 2017. The study was conducted using two major databases, SAVEE and the Polish Database, to perform SER, taking fourteen features into consideration. They obtained the maximum recognition rate for SAVEE as 75.71% using Random Forest classifier and 87.5% for the Polish database using Adaboost Classifier. They also concluded that a 14-dimensional vector with the features pitch, intensity, mean autocorrelation and a few others used by them,will be very effective for SER in other languages.

Maëva Garnier et al. [15], in their study, look into the speculation that speakers alter their speech when surrounded by noise, and the changes lead to better audibility. They conducted this study by recording ten French speakers in different noisy environments. The SPSS software was used for comparing the results by the ANOVA method and they thus concluded that there is no systematic active adaptation to the ambient noise's spectrum properties.In 2019 Yi Zhao et al. [16] studied how the Lombard effect affects emotional speech from both speaker and listener perspectives based on confusion matrix and acoustic analysis. They calculated the average SNR to be around -8.7dB and found that the ratios of the correctly perceived emotional voices under the noise condition are lower than those in the quiet condition. They discovered that speakers can accurately portray their emotions in both quiet and noisy conditions. The recognition accuracy, on the other hand, was significantly determined by the listeners' age and had a relationship with the speakers' and listeners' gender collocation.

Kerkeni et al. [17] uses Deep Neural Networks (DNN) and k-nearest neighbor (k-NN) to recognize emotion from speech, particularly when it comes to fearful emotions. Mel spectrogram, harmonic percussive, chromagram, mel frequency cepstral, beat tracking, and beat-synchronous features aggregation are some of the factors of voice that were taken into consideration. Their research was conducted to play a vital role in the medical and technical fields. In 2019, Ruhul Amin Khalil et al. [18] presented an outline of deep learning techniques used for emotion recognition based on speech. Their study shows that CNN and RCNN are among the most commonly used deep learning techniques. They also discuss what different deep learning models consist of and the problems associated with implementing models like RNN and DNN. MFCCs, PLPs, and FBANKs were utilized to model the system using various sizes of deep hidden layers. When these three acoustic features were combined, they performed better and delivered an identification rate of 92.3%, compared to 92.1% for single acoustic features when employing MFCCs. This research concluded that emotion detection systems based on deep learning architecture outperform systems that use GMMs as classifiers.

In 2021 Babak Joze Abbaschian et al. reviewed deep learning approaches for speech emotion recognition and provided a comparison between practical neural network approaches [19]. By implementing algorithms such as LSTM and CNN with different layers, they have increased the accuracy by around 20% compared to traditional ML algorithms. Recently Eva Lieskovská et al. focused on the implementation of DNN architecture for Speech Emotion Recognition [20]. They concluded that deep convolutional neural networks with recurrent networks provide good results. Labied et al. [21] focused on feature extraction to identify robust features in the audio. They derived the proper criteria that are needed to be applied when performing the feature extraction technique using the weighted scoring method. Various feature extraction techniques such as MFCC, PCA (Principal Component Analysis), LPC (Linear Predictive Coding), LPCC (Linear Predictive Cepstral Coefficient), PLP (Perceptual Linear Prediction), DWT (Discrete Wavelet Transforms were compared by multiple standards and based on the criteria needed, the extraction technique must be chosen. It also highlights the importance of applying hybrid feature extraction.

B. Emotion Recognition

Systems with a speech signal as input and estimated emotion as output are referred to as SER systems. These systems, like many other pattern recognition systems, use signal features and classification to estimate the sentiment of the input signal. Preprocessing, feature extraction, dimension reduction (optional), and feature classification are the four parts of a typical SER system [22]. The preprocessing of the voice signal is considered first before feature extraction. It is a crucial stage in the development of an effective speech emotion recognition system.

C. Emotional Speech Features

Mel-frequency cepstral coefficients (MFCCs) technique for feature extraction comprises windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by using the inverse DCT. MFCCs are frequently employed in speech recognition systems, such as those that can automatically recognize digits spoken into a phone. When compared to other techniques, the MFCC provides higher recognition accuracy in speech recognition systems.

Pitch Contour of an audio is a function or curve that records the sound's perceived pitch through time. A contour's essential relative characteristics, such as sharp pitch shifts or a pitch that increases or decreases over time, can be transposed without losing them. Because the pitch of the signal is determined by the tension of the vocal folds, it provides information about the emotional state.

Intensity Contour plots the intensity of an extracted audio object. Intensity is a key feature that reveals changes in speech energy. The intensity of acoustic signals, as well as their level and variations, reveal a lot about emotional states and how they evolve over time.

D. Speech Emotion Recognition Methods

A classifier can be built using a number of pattern recognition techniques for modelling emotional states [23]. The K-nearest neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It is a classifier that implements the k-nearest neighbors' vote. The Multi-layer Perceptron classifier (MLPClassifier) relies on an underlying Neural Network to perform the task of classification. Since partial derivatives of the loss function with respect to the model parameters are computed at each step to update the parameters, the model trains iteratively. By learning simple decision rules derived from prior data, Decision Tree produces a training model that predicts the class or value of the target variable (training data). It starts at the root of the tree when predicting a class label for a record. The root attribute's values are compared to the record's attribute. The branch corresponding to that value is followed based on the comparison, and then it jumps to the next node.

Random forest, as the name implies, is made up of a huge number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model. The Support Vector Machine (SVM) algorithm's identifies the best decision boundary or line for classifying n-dimensional space into groups so that future data points can be quickly assigned to the appropriate category. The extreme points/vectors also known as support vectors, that help create the hyperplane are chosen via SVM. Gradient Boosting creates an additive model in a progressive stage-by-stage manner; each predictor strives to improve on its previous by minimizing errors. It fits a new predictor to the former predictor's residual errors.

Gradient Boosted decision trees are implemented in XGBoost. Instead of a single decision in each "leaf" node, it uses CART trees (Classification and Regression trees), which hold real-value scores indicating whether an object belongs to a group. After the tree has reached its maximum depth, the decision can be made by transforming the scores into categories using a specific threshold. By integrating a number of ineffective classifiers, the AdaBoost classifier produces a powerful classifier that is very accurate. Adaboost's core principle is to establish the weights of classifiers and train the data sample in each iteration so that appropriate predictions of uncommon observations may be made. Bagging classifiers are ensemble meta-estimators that fit base classifiers to random subsets of the original dataset and then aggregate their individual predictions (either by voting or average) to generate a final prediction. Training in parallel is done for every base classifier with a training set that is constructed by randomly selecting N examples (or data) from the original training dataset and replacing them with new ones.

Conv1D model A 1D convolution layer (for example, temporal convolution) generates a tensor of outputs by convolving the input layer with a convolution kernel across a single spatial (or temporal) dimension. When the convolution kernels are slid along one dimension, conv1d is employed (i.e., reusing the same weights, sliding them along 1 dimensions). Since audio is a one-dimensional data, Conv1D model has been used in this undertaking.

3. PROPOSED METHODOLOGY

A. Workflow

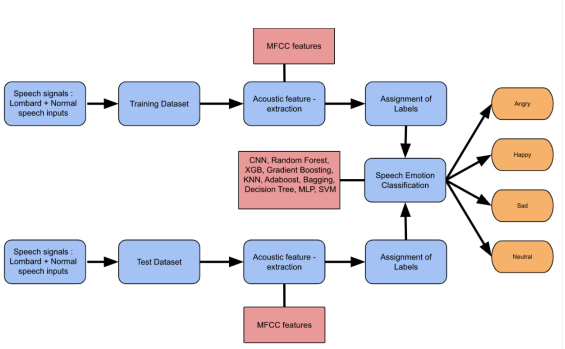


Figure 1. Workflow Diagram of proposed system design

Figure 1 illustrates the overall workflow diagram for the proposed system design. Normal speech and Lombard speech inputs are obtained from the speakers in two sets. Lombard speech input is collected by simulating the Lombard effect in a quiet environment. The simulation is done with the help of an audio consisting of babble noise. The speech inputs are separated into training and test datasets by splitting the corpus into 70%-30% respectively. The raw audio signals are then preprocessed using techniques that include pre-emphasizing, framing, windowing, and vocal activity detection. The signal will then be supplied to the feature extraction module after it has been preprocessed. The next stage consists of extracting the acoustic features. Here MFCC features were used, since according to previous research [24], they prove to show the highest accuracy for SER. Pitch and Intensity features have also been analyzed using pitch and intensity contour graphs according to each emotion for both Lombard and normal speech inputs. The

essential and emotion-relevant features will be retrieved from the signal at this stage. In the next stage, the processed speech features are assigned the labels “Happy”, “Sad”, “Angry” and “Neutral”. The decision, i.e. emotion recognition is made using the different algorithms (CNN, Random Forest, XGB, Gradient Boosting, KNN, Adaboost, Bagging, Decision Tree, MLP and SVM). This entire process is repeated using the test dataset. It is then pre-processed and the features are selected and extracted. The labels are assigned respectively and the processed data is fed into the different classifiers that were trained previously. Finally, a classification approach such as a neural network, support vector machine, Gaussian mixture model, hidden Markov model, and others will categorize the dimensionally reduced features and estimate the emotional class of the input signal in the system’s final section. The changes in emotion recognition are then analyzed after the classification.

More recent studies have shifted towards the use of Deep learning techniques for speech emotion recognition. The most used method amongst these is CNN, which is employed in this study. Convolutional neural networks (CNNs, or ConvNets) are a type of artificial neural network (ANN) used to assess visual imagery. In comparison to other classification methods, CNNs require very little pre-processing. This implies that the network learns to optimize the filters (or kernels) by automatic learning, as opposed to hand-engineered filters in traditional techniques. This lack of reliance on prior information or human intervention in feature extraction is a significant benefit. In this paper, we propose to use a Conv1D CNN model to achieve better speech emotion classification results.

B. *Dataset Collection*

In 2018, Swain et.al.[25] did a comparative study of 59 researches and the databases that were used and if the speech was simulated, natural or elicited by the authors. Those databases were all in different languages and none of them were open-sourced or readily available for our usage. Hence, an English emotional database was recorded, by making participant speakers read appropriate prompts in four emotions: neutral, happy, sad and angry. The sample rate was set at 16 KHz and the encoding was set to Lin16, for recording all the audio files. The dataset was thus formed by collecting the voices of 25 college students (aged 18-22). Eight samples(4 Normal + 4 Lombard speech inputs) were collected from each of these students, thus 200 speech samples were trained and tested in total. None of the speakers had any prior knowledge about Lombard effect. Each speaker recorded their voice in a quiet laboratory environment. Lombard effect was simulated by making the speakers listen to babble noise in a cafe environment through a set of noise canceling headphones [26] . Some of the participating students were part of a film club and hence their emotions were captured better than the others, but none of them reported any difficulty in the recording procedure. The audio was recorded in a speech recording sound-proof room with an omni-directional mic. All the speakers recorded emotional speech in two sets, each set with a duration of almost 1 minute. The first set was emotional speech recorded in a quiet laboratory environment. The second set of emotional speech was recorded after inducing Lombard effect. The speakers were required to listen to noise played through a pair of noise-canceling headphones while reading out the lines this time. In total, eight recordings of each speaker were collected. Every speaker was given the same corpus for the respective emotion.

C. *Analysis of Dataset Features*

The sound waves from each frame were extracted as objects. The pitch and intensity contour plots for the frames were drawn for each emotion, for both normal and Lombard. The intensity contours were plotted between the range of 0-100dB whereas the pitch contour plots were between the range 0-500Hz.

Pitch Contour plots

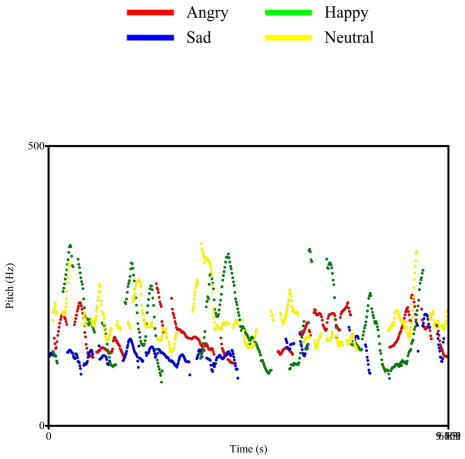


Figure 2. Normal Speech Emotion

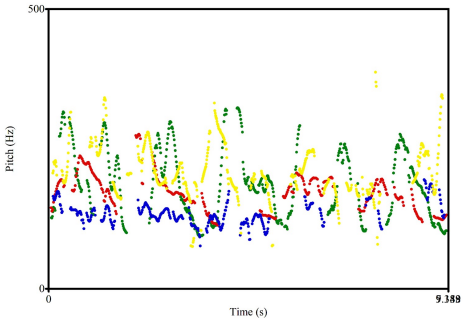


Figure 3. Lombard Speech Emotion

From the contour plots shown in Figure 2 and Figure 3 we can see that the emotions Happy and Neutral have higher pitches compared to the Angry and Sad emotions that show a significantly lower pitch. In the pitch contours extracted,

we observe that happy and neutral emotions have notably higher peaks considering that the contour plot for sad and angry emotions do not show visible variations.

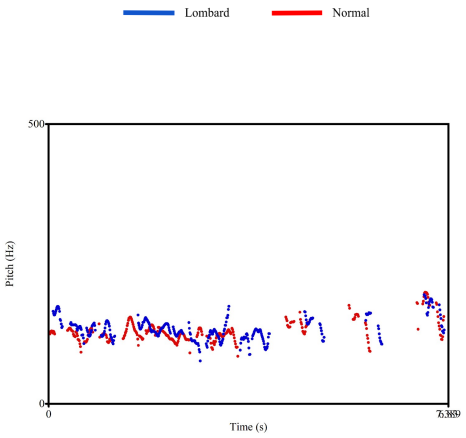


Figure 4. Sad vs Sad Lombard

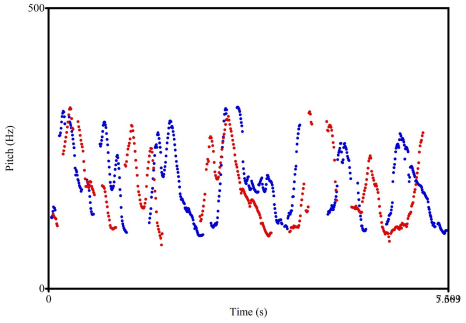


Figure 5. Happy vs Happy Lombard

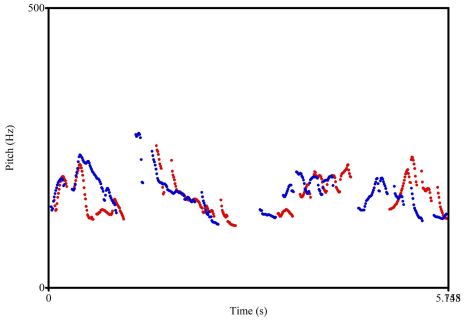


Figure 6. Angry vs Angry Lombard

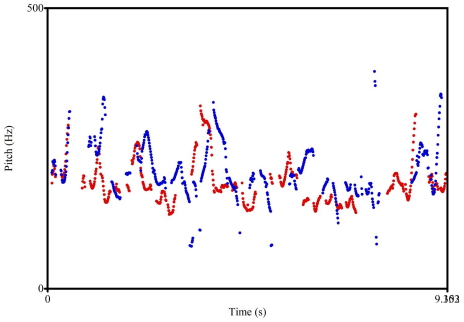


Figure 7. Neutral vs Neutral Lombard

From Figure 4 to Figure 7 we note that the Lombard speech has greater pitch than the normal speech audios. This can happen because the speakers tend to be louder when they hear external noise to convey the same message that they would, with better emotional expression in a normal, quiet environment. All this can cause the pitch to increase more significantly than other features in the audio.

Intensity Contour plots

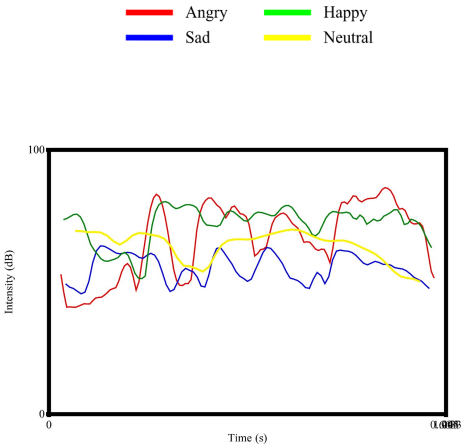


Figure 8. Normal Speech Emotion

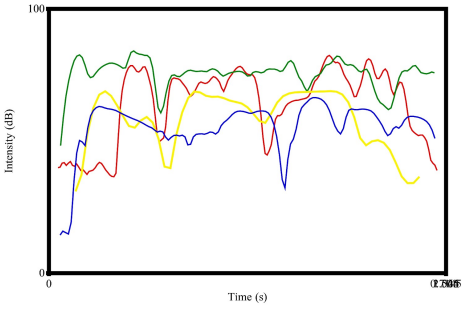


Figure 9. Lombard Speech Emotion

From Figure 8 and Figure 9 we can observe from the intensity contour plots of the four emotions that angry and happy emotions exhibit much higher intensities in contrast to the lower intensities seen in the sad and neutral emotion plots.

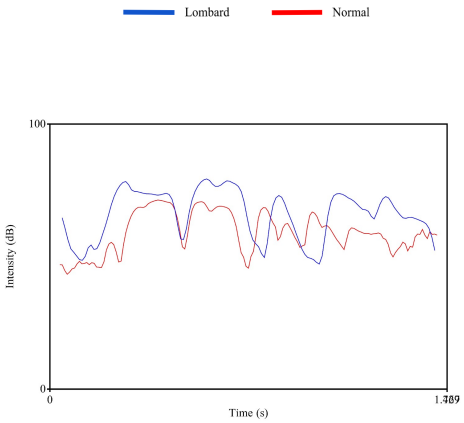


Figure 10. Sad vs Sad Lombard

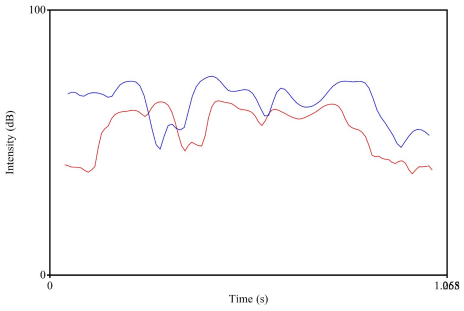


Figure 11. Happy vs Happy Lombard

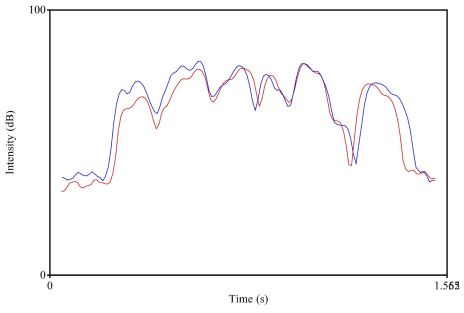


Figure 12. Angry vs Angry Lombard

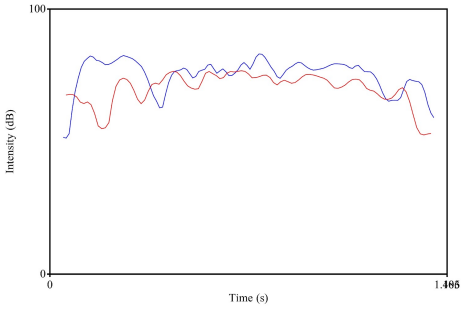


Figure 13. Neutral vs Neutral Lombard

This deduction is consistent with existing studies which state that the happiness curve has more peaks, lower holes, and higher fluctuations, whereas the sadness curve does not have any observable variations, and the neutral curve remains lower than other plots. Similar to the normal vs Lombard speech pitch contour plots, the intensity contour plots of Figure 10 to Figure 13 also show that Lombard speech appears to have higher intensity values than normal speech, for similar reasons associated with the loudness of the speakers, after inducing Lombard effect.

Figure 14. Pitch and Intensity Spectrum

To summarize, happy emotion lies in the high intensity, high pitch spectrum, while the angry emotion lies in the high intensity low pitch spectrum. Similarly, the high pitch, low intensity spectrum contains the neutral emotion whereas the sad emotion is present in the low pitch, low intensity region, as shown in the Figure 14 above.

D. Implementation

This section talks about the experiments carried out for SER in different scenarios. The audio files from the created

dataset, were first downloaded and visualized by plotting a Spectrogram of a few selected frames, shown in Figure 15-

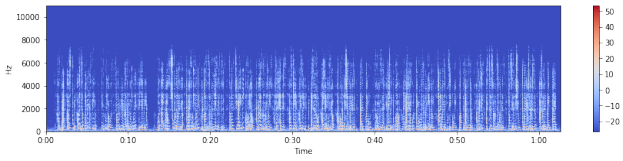


Figure 15. Spectrogram of Audio File

After visualization, the data is preprocessed, wherein it is scaled using the fit_transform method and the string values are then converted to categorical values using the LabelEncoder library. To enhance the performance of the models to be used, the audios were processed using Audacity, a multi-track audio recording and editing software. By doing so, silences or non-voice segments from the audio were removed and filtered out using a threshold value of -20dB. The remaining voice segments were then truncated together. This process was continued for every audio file in the dataset. Feature extraction was then performed on the collected dataset, which was sampled at 16kHz, as the first step. This was done using MFCC (Mel-frequency cepstral coefficients) feature extraction function of Librosa library as seen in Figure 16. These features were extracted for window size of 20ms and overlap stride shift of 5ms. A 39-dimensional vector was extracted for each window.

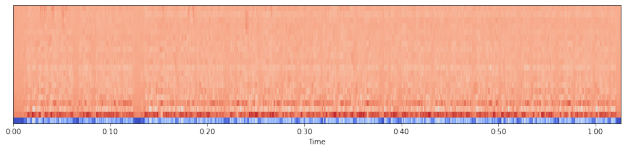


Figure 16. Plot of MFCC features of Audio File

The feature extraction was repeated for each emotion (25 audios each) and stored into 4 different bins with their respective labels. All the extracted feature vectors from normal speech audios were concatenated into one dataframe. To avoid running into type or value errors, the null values were replaced with 0. To ensure equal distribution of emotions during training the models, the data frame was then randomly shuffled internally, using the shuffle function from sklearn library. The same was done for Lombard speech audios. The new dataframe is then split into train and test data, using 70% for training and 30% for testing. Traditional machine learning models like Multi-Layer Perceptron (MLP), Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGboost, Decision Tree, Bagging, Gradient Boosting and AdaBoost were first tested and implemented for normal speech dataset, with the above-mentioned train-test split. To obtain better results with emotion recognition, a one-dimensional Convolutional Neural Network model was applied, since deep learning techniques have been said to give more accurate results compared to traditional ML models.

After assessing the normal speech dataset, emotion recognition was performed on Lombard speech. This experiment was carried out in two sets. In the first case, normal speech was used for training and Lombard speech audios were used for testing. Eight audio files which exhibited the most dominant emotions from the Lombard speech dataset, were used against the twenty five audio files in the normal speech dataset, to preserve the 70:30 ratio. The second set was executed with Lombard speech data for both training and testing, similar to the emotion recognition performed on normal speech.

Both the sets of experiments were performed using different ML models first. An accuracy of 56.29% and 27.39% for the KNN model tested on normal speech and Lombard speech respectively, was received by defining the number of neighbors as four representing the four different emotions studied in this project. The Random Forest Classifier employed 300 estimators and obtained classification results of 73.7% and 34.75% corresponding to the normal and Lombard speech data. The linear kernel was used for the Support Vector Classifier to yield 28.76% and 26.28% as accuracies for normal and Lombard speech data, in the mentioned order. The MLP Classifier with alpha = 0.01, batch size = 256, with one hidden layer consisting of 300 neurons and a learning rate set to adaptive gave 30.79% and 27.43% as the outcome for each normal and Lombard speech emotion recognition. The XGB Classifier with the parameter objective set to multi:softprob displayed recognition rate of 70.44% and 32.38% for normal and Lombard speech data respectively. 100 estimators were used in the Gradient Boosting Classifier to give the prediction results as 67.41% for normal speech and 32.99% for Lombard speech data consequently. The Bagging classifier with KNeighbors classifier and maximum samples set to 0.5 and maximum features set to 0.5 resulted in accuracies of 42.92% and 27.83% for normal and Lombard speech data respectively. The Decision tree classifier from the sklearn library gave 36.96% and 27.42% as the performance rates using normal speech and Lombard speech data, subsequently. Lastly the AdaBoost classifier with 100 estimators was trained and tested to achieve a predictive performance of 48.42% and 28.37% for normal and Lombard speech data respectively.

Finally, a Convolutional Neural Network (CNN) model consisting of six Convolutional layers has been implemented for the dataset in hand; the padding “same” and the activation function RELU is used with each layer. It is a sequential model with a total of 10 layers including softmax as the final layer. To execute this model, the train and test values were expanded into an array of three dimensions. The shape of the training labels was (2083,4) and that of the training features was (2083,16538). A single pooling layer of size 8 was applied after the second convolutional layer. A dropout of 0.1 after the second convolutional layer and a dropout of 0.2 after the fifth layer was applied to prevent

overfitting. Adam optimizer was applied with a learning rate of 0.001 The final softmax output layer is given as "Dense (4)" to detect the four emotions being studied. This model was fit and run for ten epochs with a batch size of 10. An accuracy of 97.63% was achieved with normal speech data and the same model when trained and tested with the Lombard speech data gave an accuracy of 78.60% as the result. The model was then trained on normal speech data and tested with Lombard speech data to study the impact of Lombard effect on SER, and this experiment gave an accuracy of 36.46% with the above-mentioned CNN model.

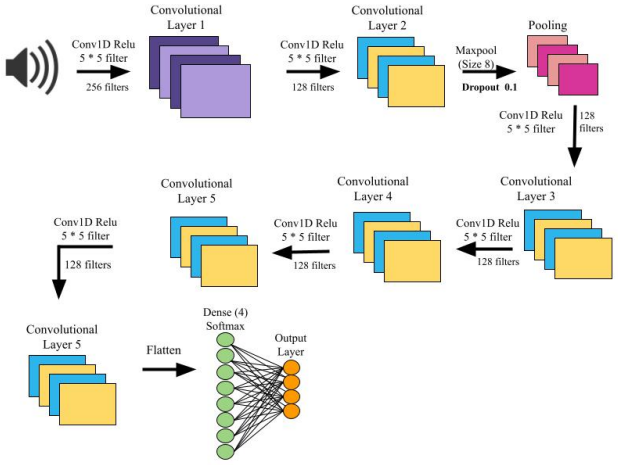


Figure 17. CNN Architecture

4. EXPERIMENTAL RESULTS

After trying the different algorithms mentioned above, the results have been tabulated in Table I, in the decreasing order of accuracy.

TABLE I. Accuracies Obtained from Different Models

Algorithm Name	Testing Data: Normal Speech, Training Data: Normal Speech	Testing Data: Lombard Speech, Training Data: Lombard Speech	Testing Data: Lombard Speech, Training Data: Normal Speech
CNN	97.63%	78.60%	36.46%
Random Forest	73.7%	66.83%	34.75%
XGB	70.44%	60.35%	32.38%
Gradient Boosting	68.20%	62.15%	32.99%
KNN	56.29%	53.30%	27.39%
AdaBoost	48.42%	42.92%	28.37%
Bagging	42.92%	37.13%	27.83%
Decision Tree	36.96%	36.42%	27.42%
MLP	30.79%	28.08%	27.87%
SVM	28.76%	24.82%	26.28%

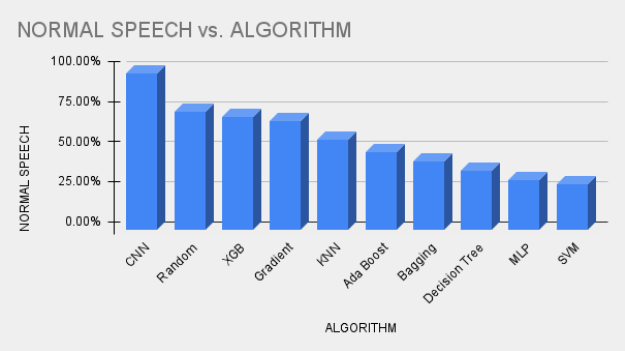


Figure 18. Results after training and testing with Normal Speech

The second column indicates the accuracies obtained after training and testing the algorithm with the normal speech data. As the speakers were not disturbed by babble noise or white noise while speaking, the emotions were clearly visible in their speech and the respective emotional features had a clear difference during feature extraction. From Figure 18, we understand that the CNN algorithm has a much better performance than all the other algorithms.

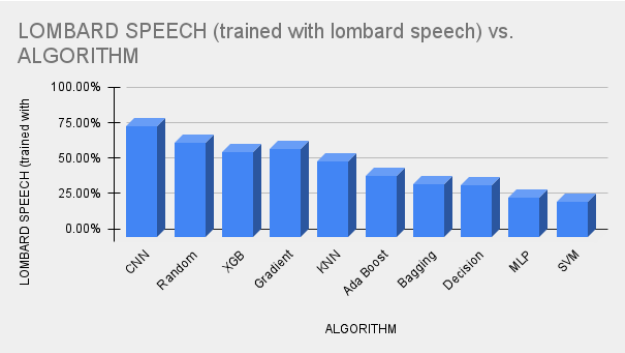


Figure 19. Results after training and testing with Lombard Speech

The results that were achieved while training and testing the model with the Lombard speech data is consolidated in the third column. The accuracies of all the models have a significant decrease when the input is Lombard speech, as the emotional features might not have been clearly identifiable since the speakers might have increased their volume or the manner in which they speak in the simulated noisy environments. These results can be seen in Figure 19.

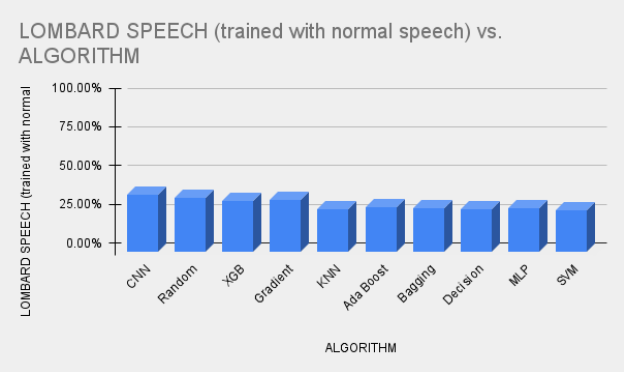


Figure 20. Results after testing with Lombard Speech and Normal Speech as Training Data

The final column consists of the outputs received by training the model with the normal speech data and testing the same model using Lombard speech data. The prediction results are the lowest for this case, as seen in the fourth column, and this is due to the considerable distinction between normal speech audios and the ones recorded with induced lombard speech Figure 20 . These findings demonstrate that the Lombard effect significantly affects speech emotion recognition and should be studied independently because regular SER is not always practical and is not effective in noisy environments.

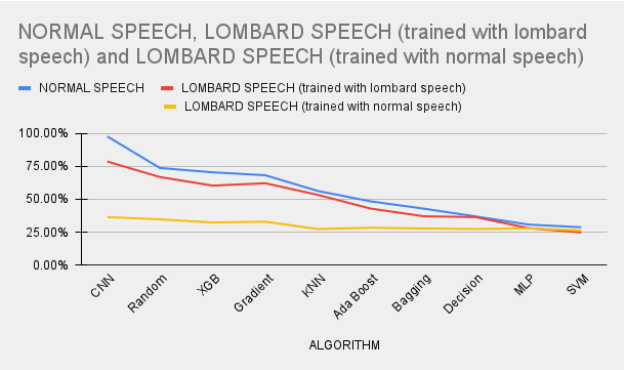


Figure 21. Comparison of Results

Figure 21 is a comparison chart between the accuracies produced by the different models that have been used in this project, in decreasing order. As observed from the figure, CNN Conv1D model is better compared to any other similar model on this dataset by a large margin. CNNs are better than other machine learning models for Lombard speech emotion recognition for several reasons - Lombard speech is a form of speech produced under stress, and as such, it is more variable than normal speech. CNNs have the ability to extract features from raw audio data, making them well-suited for recognizing patterns in the audio that are indicative of different emotions, even in Lombard speech.Lombard speech contains more noise and variability than normal speech, and CNNs are robust to noise. They can learn to identify patterns in the audio despite the presence of noise and variability, which makes them well-suited for Lombard speech emotion recognition. Lombard speech has a lot of information in its prosodic and spectral characteristics, CNNs are good at handling time-series data, and can effectively extract information from prosodic and spectral characteristics, this makes them well-suited for Lombard speech emotion recognition. Lombard speech has different characteristics than normal speech, CNNs can learn these characteristics and generalize them to new samples, this is a good feature for Lombard speech emotion recognition. CNNs have been widely used and have shown to be successful in many emotion recognition tasks, Lombard speech emotion recognition is not an exception. All these points combined make CNNs a good choice for Lombard speech emotion recognition.

5. Conclusions and Future Work

In this paper we review and compare the various algorithms for speech emotion recognition along with induced Lombard effect. The entire speech dataset was created by the authors and around 16538 features were obtained using the MFCC feature extraction method. In this study, the authors analyzed the role of pitch and intensity and how it varies with the different emotions. From table 1, the experimental results conclude that the CNN model produces higher accuracy for SER compared to other traditional Machine Learning models. We also note that the accuracies obtained using Lombard speech data are significantly lower than the ones obtained using normal speech data. The reason for the same is the impact of Lombard effect on the speakers' pitch, intensity, loudness, and other audio features. We can thus say that Lombard effect makes SER less precise. The results in the final column of table 1 show that the audios recorded in normal conditions are very different from the ones recorded after inducing Lombard effect. So, if any new SER tool is to be designed keeping Lombard effect in mind, the training dataset should be recorded with the influence of Lombard effect on the audios. This research can be developed further to make applications or tools for SER more efficient by including the Lombard effect speech recognition as an additional feature. Some ways to achieve better results, such as using customized CNN models or other deep learning models, that are more accurate, and considering other audio features like pitch, energy, voice quality, intensity, etc., that impact emotion recognition, can also be implemented to make the said application more efficient.

References

[1] 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, 2020.

[2] K. Kakol, G. Korvel, G. Tamulevičius, and B. Kostek, "Detecting lombard speech using deep learning approach," *Sensors*, vol. 23, no. 1, p. 315, 2023.

[3] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[4] F. Kelly and J. H. Hansen, "Analysis and calibration of lombard effect and whisper for speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 927–942, 2021.

[5] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 international conference on machine learning and cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.

[6] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs." in *INTERSPEECH*, 2006.

[7] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *2008 IEEE international conference on semantic computing*. IEEE, 2008, pp. 158–165.

[8] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 4605–4608.

[9] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[10] C. Huang, G. Chen, H. Yu, Y. Bao, and L. Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, 12 2013.

[11] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[12] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," in *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 2016, pp. 788–792.

[13] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[14] F. Noroozi, D. Kaminska, T. Sapinski, and G. Anbarjafari, "Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost," *Journal of the Audio Engineering Society*, vol. 65, no. 7/8, pp. 562–572, 2017.

[15] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the lombard effect on audio and visual speech recognition systems," *Speech communication*, vol. 100, pp. 58–68, 2018.

[16] Y. Zhao, A. Ando, S. Takaki, J. Yamagishi, and S. Kobashikawa, "Does the lombard effect improve emotional communication in noise?-analysis of emotional speech acted in noise," *arXiv preprint arXiv:1903.12316*, 2019.

[17] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social media and machine learning*. IntechOpen, 2019.

[18] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[19] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[20] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021.

[21] M. Labied and A. Belangour, "Automatic speech recognition features extraction techniques: A multi-criteria comparison," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, 2021.

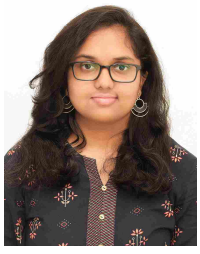
[22] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[24] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45–55, 2020.

[25] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[26] D. I. S. Pigeon, "mynoise bv," 2013. [Online]. Available: https://mynoise.net/NoiseMachines/cafeRestaurantNoiseGenerator.php



Indirapriyadarshini A is a final year undergraduate student at Sri Siva Subramaniya Nadar College of Engineering. She has worked on a funded project based on information security of android users and has also participated in two conferences. Her interests include web development, front end design, machine learning, mathematics and logics.

Mahima S is a final year engineering student doing her B.Tech in Information Technology from Sri Sivasubramaniya Nadar College of Engineering. She has around 9 months of internship experience as a Software engineer and has one journal publication. Her research interest lies in Machine learning, Deep learning and analytics.

Shahina A is a professor in the department of Information Technology at Siva Subramaniya Nadar College of Engineering. She has more than 20 years of teaching and research experience, with over 5 years of research in the field of speech processing. She works on applying machine learning, deep learning and reinforcement learning algorithms in the fields of healthcare, security and surveillance.

Uma Maheswari has around 13 years of teaching and 6 years of industrial experience. She has completed her Ph.D. from Faculty of Electrical Engineering, Anna University, Chennai in the area of speech processing. Her research interests include pattern recognition, machine learning and speech processing.