Psychological Stress Detection Through Speech Using Machine Learning

Nirasha Kulasooriya[1*], Shan Menaka[1], Isuri Prabuddhini[1] and Chamila Karunatilake[1]
[1]Department of Information and Communication Technology, Faculty of Technology,
University of Sri Jayewardenepura, Sri Lanka
Corresponding Author E-mail: nirashakulasooriya@sjp.ac.lk

***Abstract*—This Research focuses on identifying individuals' psychological or mental stress based on their speech signals using machine learning. The speech signals are restricted to the Sinhala Language hence, to utilize in the training and testing of machine learning models, a new dataset has been created with 144 speech files. Two machine learning models were developed with the intention of stress identification and classification using Support Vector Machine (SVM) and Artificial Neural Networks (ANN). Speech signals were converted into frequency domain with Fast Fourier Transform (FFT) and the Mel Frequency Cepstral Coefficients (MFCC) were generated as feature vectors for the machine learning models. The experiments were conducted with both machine learning models where SVM provided 59% accuracy and ANN provided 63% accuracy in identifying psychological stress from speech signals.**

Keywords—artificial neural networks, fast fourier transform, psychological stress detection, speech signal analysis, support vector machine

I. Introduction

Speech is the verbal way of exchanging ideas among humans which is claimed to be derived about 50000 years ago in the Upper Paleolithic age. This form of communication is absent in other closely related species including Neanderthals and earlier humans. Vocal folds of the larynx are mainly responsible for generating voice signals. Nonetheless, the motor cortex, basal ganglia, arcuate fasciculus, Wernicke's area, Broca's area, and auditory cortex are all linked to a complex neurological process. [1]. Therefore, the emotional state of the brain has a direct correlation with speech and it could be reflected in the speech signals [2][3].

Mental stress is a physiological response to perceived challenges in the external or internal environment. In those situations, hormones such as cortisol and adrenaline are released, and hence, sugar levels in the blood are increased. Also, blood flows to the lungs and muscles are raised and heart rate is boosted. This process is usually considered as a positive reaction that prepares the person physically and mentally to confront a threat. These responses died out as the challenging circumstances disappeared. However, there are cases that they do not switch off which could lead to more serious and long-term negative effects [4][5].

People in modern society with complicated lifestyles are frequently confronted with competitive situations that could trigger stress, and hence suffer from prolonged stress-related problems. When a person is mentally stressed, physical and emotional changes could affect the voice of the person and it could be detected in the speech [6][7]. For effective communication with people especially when it is taken place via voice media such as telephone or online voice chat, it is essential to detect the emotional status of the other parties. If an appropriate mechanism can be generated for detecting persons' condition of stress using speech signals, there are many applications in various domains. Also, several studies have been conducted in this area to address the same problem, and reasonable outcomes have been achieved.

This study specifically focuses on the detecting psychological stress in speech signals of the Sinhala Language, using machine learning. However, a dataset of speech related to stress is not available for Sinhala Language, hence, a dataset of 144 audio files was created in the early stage of the research. This dataset includes recorded speech files of different people in stressful and non-stressful situations.

Speech signals originally in the time domain are transformed into the frequency domain using Fast Fourier Transform (FFT) and Mel Frequency Cepstral Coefficients (MFCC). These features are then utilized as the input vector for stress detection models. According to the previous experiments that have been conducted, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have been frequently utilized for detecting stress in speech signals. Hence, those two machine learning techniques were tested and compared with the created dataset and the details of the experiments and the analysis of the results are presented in the later sections.

II. Literature Review

Speech-based stress detection has become a substantial field of exploration in recent years. In this section, the literature on stress detection using speech is presented. Dhole and Kale have used machine learning (ML) models and artificial intelligence (AI) together with MFCC feature extraction methods to develop an algorithm for stress detection and classification [13]. In the study, SVM has achieved 97.52% of accuracy among various classification models such as SVM, Recurrent Neural Networks (RNN), Radial Basis Functions (RBF), Multilayer Perceptron (MLP), Adaptive Neuro Fuzzy Inference System (ANFIS), Feed-Forward Back Propagation Neural Networks (FFBP NN) where the features are extracted using the MFCC [13], [14]. Also, this model is capable of real-time speech detection and it is language and word-independent [13]. A speech analysis has been done to detect candidates' psychological stress while

passing through them in HR (Human Resources) screening interviews with the help of Machine learning. For this experiment, they have used few classification features such as mean energy, mean intensity, and Mel-Frequency Cepstral Coefficients (MFCCs) to detect stress in speech which have been completed on three different datasets: i) EmoDB (German), ii) KeioESD (Japanese) and iii) RAVDESS (English) [14]. Both ANN and SVM algorithms have shown satisfactory results, however, with ANNs have provided slightly better accuracies than SVMs. Hyewon and Han et al, have proposed a deep learning-based algorithm for psychological stress detection using speech signals. For the detection of the stressed condition of speech signals, audio clips have been collected by running through a specific process with three steps and established stress-detection models using deep learning algorithms with long short-term memory (LSTM) [15]. The condition of psychological stress (i.e., stress vs. unstressed) was detected as a binary classification task, and a 66.4% of accuracy has been obtained. A study conducted by Firoz Shah et al focuses on automatic recognition of stress from spoken words in the Malayalam language. It explores stress detection from speech to assess the exact stress level of human voice signals with the aid of various machine learning algorithms [21].

III. Methodology

One of the initial steps of our methodology is extracting the MFCC features from the voice signals. Each step of the MFCC feature extraction process is carefully executed. Following this, the machine learning model training is initiated using the generated datasets. Those steps are detailed below.

A. Feature Extraction

In human speech, various types of features can be identified. However, in alignment with the primary objective of the study, only the most relevant features of the voice signals are concerned. With the careful analysis of previous research on stress detection in speech signals, MFCC was selected as the most suitable set of features. At initial stage of performing feature extraction, the data pre-processing was performed and followed the steps of the MFCC feature extraction method.

B. *Pre-processing with MFCC: Mel-frequency Cepstral Coefficients (MFCC) can be considered the standard method for feature extraction in speech recognition. [14]*

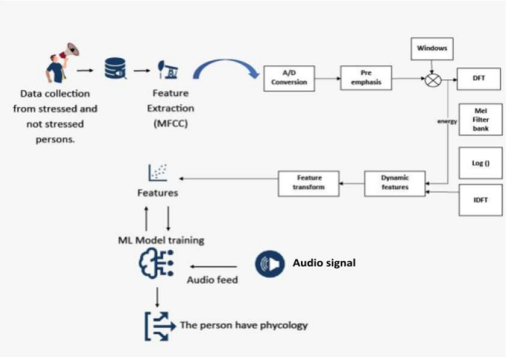

Fig. 1: The overall architecture of the proposed methodology

This method consists of several steps such as pre-emphasizing, framing, windowing, converting the signal into frequency domain using the Fast Fourier Transform (FFT), and passing through the Mel Filter Banks and Discrete Cosine Transform.

Pre-emphasis: This is the first tep of the MFCC feature extraction. Here, a single word sample is passed through a high-pass filter which emphasizes higher frequencies by boosting the energy of the signal at those frequencies. This step assist to retain critical informations in the higher frequency components.

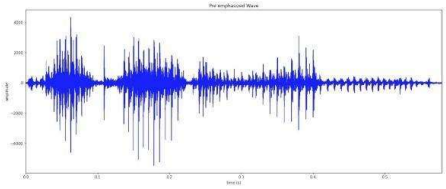

Fig. 2: Result of the Pre-emphasis step for amplifying the high-frequency components of a signal.

Framing: In framing, the pre-emphasized speech signal is divided into small consistent blocks known as frames, typically duration of 20ms, with in the range of 5ms to50ms. In this research, a speech signal is segmented into 25ms frames with an overlap of 20ms.

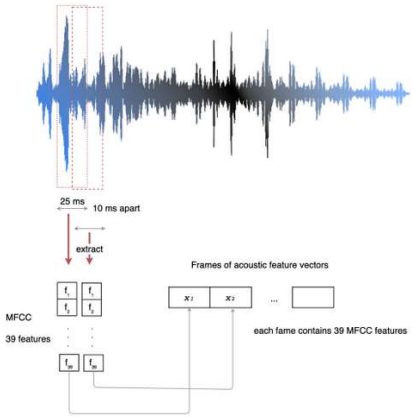

Fig. 3: The signal is segmented into Frames

Windowing: After segmenting the signal into frames, the Hamming window is applied on each frame as a windowing function. In this process, for maintaining the continuity of the first and last points in each frame, they should be multiplied with a window function. Hamming window, used in this process, acts as a smoothing function for the signal and it reduces spectral leakage ensuring a more accurate representation of the frequency components.

Fast Fourier Transform: Fast Fourier Transformation (FFT) is performed to convert each frame which consists of N samples from the time domain into the frequency domain. After applying FFT, the frequency content of the signal can be visualized and analysed.

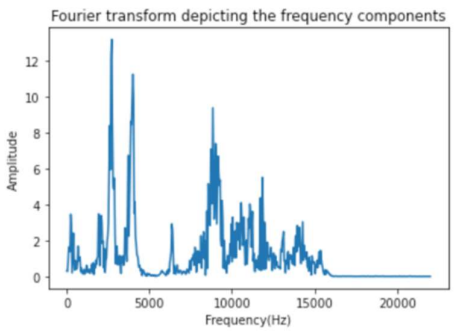

Fig. 4: The frequency spectrum of a frame generated using the Fast Fourier Transform (FFT)

Mel Filter Banks: This can be considered the final step of the feature extraction process. Mel can be defined as the unit of measuring actual frequency for human ear perception of sound. In Mel-filter banks, a set of triangular filters are utilized, typically the number of filters ranges from 20-to 40 can be seen. However, in this study, 13 band-pass filters were adopted to pass the Fourier-transformed signal. Here, the filters are placed based on Mel-scale are applied to each frame. The main objective of using a Mel-filter bank is, to extract frequency bands by converting the Mel-scale into a power spectrum.

The conversion is done by using the following equations,

$$10\ (1+f\ 700) \qquad (1)$$

Equation (1) shows the way of converting to Mel-scale where f is the frequency I Hertz and m is for Mel scale,

$$F=700\ (10m/2595-1) \qquad (2)$$

Equation (2) shows the way of converting to the Frequency Spectrum.

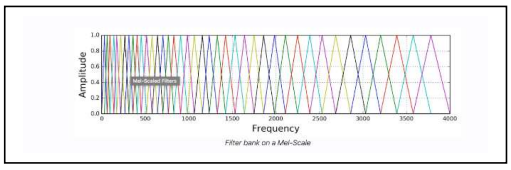

Fig. 5: Mel Filter Banks

For experimental purposes, the datasets were generated for 15ms, 25ms, and 35ms window sizes separately. Inside of a single feature vector, there were 13 parameters such as sample_rate, win_length, win_step, nfft, low_freq, high_freq, etc.

C. Dataset

In the domain of mental stress analysis in Sinhala speech, one of the main challenges was to find a standard dataset to be utilized as the benchmark. Since the lack of availability of a public dataset, it was essential to create a dataset with speech recordings in Sinhala Language. For speech recordings, 24 people were utilized and each subject was asked to speak three emotional-neural sentences.

The dataset should include speech recordings when the subjects are in stressful conditions as well as when they are in a relaxed state. To record speech in stressful conditions, two types of environments were used. voice clips from the students on the day before the GCE (Advanced Level) Examination and voice recordings from the people who are playing an intense video game were recorded. For unstressed conditions, the same group of people was utilized when they were free and in a comfortable mode. Before recording, a piece of relaxed music was played in the background.

1. ඒක අරගන්න (Eeka Araganna). (Get it)
2. මෙතනින් ඉද ගන්න. (Methanin Inda Ganna) (Sit here)
3. මේ වැඩේ ඉවර කරන්න. (Mee wede iwara karanna) (Finish this job)

Also, when voice clips were recorded, the subjects' blood pressures were measured using a digital blood pressure monitor to verify the stress levels. In general, the heart rate is increased based on the person's psychological stress level. The main reason for that is, that the ventricles in the heart contract more forcefully, causing an increase in blood pressure [12].

Six (06) numbers of separate and differentiated datasets were generated based on the differentiation of window sizes in two conditions of stress. In the first version, 34 samples of voice clips were taken by impelling to read three sentences under both conditions for 15ms, 25ms, and 35ms. In the second version, the dataset had been increased to 144 samples of voice clips, with the same sentences differentiating window sizes as 15ms, 25ms, and 35ms. Those were taken under both stressed and non-stressed conditions.

Also, the duration of these audio files varies from 2 to 4 seconds. These sentences were kept short to make the audio files short in length.

D. Model Training Process

In the model training process, three machine learning models were used with the dataset, Support Vector Machine (SVM), and Artificial Neural Network (ANN). SVM and ANN were selected and tested initially.

E. Support Vector Machine (SVM)

SVMs belong to supervised learning algorithms commonly used to deliver classification and regression tasks. [20] SVM's main objective is to divide the datasets having any number of features into groups to find the maximum marginal hyperplane (MMH) that uniquely classifies the data points. First, SVM will iteratively generate hyperplanes with the best isolation of the groups. Then it will choose the hyperplane that correctly divides the groups. [19] Since the SVM works in a high-dimensional feature space, kernel functions are included to perform computations without mapping the data point more explicitly. The kernel types such as linear, polynomial, and radial basis function (RBF) can be considered as general types.

The basic mathematical theory behind SVM is to find the hyperplane which helps to maximize the margin between different classes. Given a dataset with features, label xi and label yi, where i= 1,2,3,4…, n. and y can be -1 or 1.

$$w^T*x +b =0 \qquad (3)$$

where in the equation (3) the weight vector is denoted by w, and the bias term is denoted by b. The decision rule for classifying a new data point.

$$\min 1/2||w||2 \quad (4)$$

Here, the weights are denoted by w, b is the bias term, and ||w|| denotes the Euclidean norm of as equation (4).

F. Artificial Neural Network (ANN):

Models of computing based on the architecture and operation of the human brain are called artificial neural networks (ANNs). [21] A specialized cell that is an essential component of the nervous system is the biological neuron. Information transmission and processing are carried out via neurons. Neurons facilitate communication through electrical and chemical signals. The architecture of an ANN consists of a vast amount of artificial neurons, which are considered nodes and are made up of interconnected nodes grouped into layers such as input, hidden, and output. The fundamental unit of an ANN is the neuron, which determines a weighted sum of all its inputs and an activation function is applied to produce an output. With the last number of neurons, this type of model can handle more than one task simultaneously. ANNs are capable of performing tasks including classification, regression, and pattern recognition. Here the output of a neuron is denoted from Oj where in a single neuron,For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

$$O_j = f\left(\sum_{i=1}^{n} w_{ij} \cdot a_i + b_j\right) \quad (5)$$

Here in equation (5), xi represents the input to the neuron j, Wij is the weight associated sum of the input i with respect to the neuron j, bj denotes the bias term and f denotes the activation function in this equation.

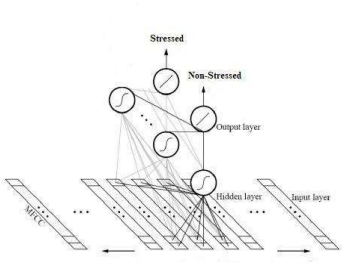

Fig. 6: Abstract view of the ANN model

IV. Experimental Overview

According to the observations in this study, data were collected two times transforming those into the dataset setting three different window sizes and two versions. As mentioned in the above methodology part, six types of datasets were generated under three different kernels in SVM. In data acquisition, three environments were designed. One is for neutral condition and the other is both for stressed condition. With the help of MFCC, features of voices were extracted for all the combinations of datasets in both versions. Here our own dataset was used for the model training process.

Mainly, for the training process of models the Support Vector Machine (SVM) and Artificial Neural Network (ANN) were used. In SVM, support vectors that are the nearest to the hyperplane utilizing Vectors-Data Points, which can be considered as one of the most essential concepts. With the aid of these data points, the separation line or hyperplane can be easily initiated. SVM's main objective is to divide the datasets into groups and find a maximum marginal hyperplane in between them. In training the SVM model different kernels were utilized within the SVC functions like polynomial, Radial Basis Function (RBF), linear, and sigmoid.

Artificial Neural Networks are capable of handling issues with the use of self-learning and self-organizing where they are distinguished by their topologies, activation functions, and weight vectors, which are occupied in their hidden and output layers for processing with the help of basic mathematical operations.

Further during the study, an Artificial Neural Network sequential model was built using Keras for the training of Neural Networks. During the stage of model building, a fully connected network structure (Dense) with three layers was used. Moreover, the ReLu activation function was applied on the first two layers, and in the output layer, the sigmoid activation function was applied.

V. Results and Evaluations

In line with the experiments on both SVM and ANN, all the models have gained over 50% as mentioned in tables I and II. In this case, the window sizes are represented by each column as 15 ms, 25 ms, and 35 ms and for the kernel functions polynomial, radial basis function, and sigmoid are indicated in each row.

To evaluate the performance of the SVM and ANN models more hypothetically, confusion matrices were constructed using the resulting positive and negative counts of the results in actual and predicted ways. For the calculation of accuracy (6), precision (7) and recall (8) following equations were used.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

According to the derived outcomes, among both SVM and ANN, an ANN model with a 35ms Window size and a second version dataset performed well for our dataset which displayed 63% accuracy.

Both models have their limitations. The ANN model's reliance on substantial computational resources makes it less practical for real-time applications on devices with limited hardware capabilities. Additionally, ANN requires significant expertise in selecting appropriate architectures and hyperparameters, which could introduce variability in results. On the other hand, the SVM model's inability to adapt dynamically to varying temporal patterns limits its effectiveness in handling data with high variability.

TABLE I. RESULTS OF SVM

Window size	0.015s		0.025s		0.035s	
Versions	Initial	Second	Initial	Second	Initial	Second
Polynomial	57%	57%	54%	58%	56%	58%
Radial basis Function	57%	57%	55%	58%	**59%**	**59%**
Sigmoid	50%	50%	50%	51%	55%	51%

TABLE II. RESULTS OF ANN

Window size	0.015s		0.025s		0.035s	
Versions	Initial	Second	Initial	Second	Initial	Second
Epoch – 150 Batch size- 10	59%	58%	60%	58%	**61%**	**63%**

Furthermore, SVM's performance heavily depends on the choice of kernel and hyperparameters, and improper tuning can lead to suboptimal results.

VI. Conclusion and Future Work

To detect mental stress in individuals based on speech in the Sinhala language, this research applied Support Vector Machines (SVM) and Artificial Neural Networks (ANN), achieving accuracies of 59% and 63%, respectively. One of the major contributions of this research was to introduce the Sinhala speech audio dataset which consists of voice clips recorded in stressful and non-stressful situations which has been a gap in the domain. The accuracies are concerned, the SVM displayed slightly higher accuracy than ANN since the SVM is robust to the size of the dataset. The maximum number of data here is 144 which could be the reason for the lower accuracy in ANN. Expanding the dataset may increase the accuracy of ANN. Also, the longer audio recordings would provide more data for training.

Future research is conducted based on the dataset expansion, and the model enhancements. Also, it is desired to utilize Deep Learning models such as Recurrent Neural Networks with LSTM. Deep Learning methods tend to demand larger datasets. Based on the enhancements, this stress detection model has potential applications in the areas of virtual interactions, customer care, health care, and others. In addition to that, a complete solution for biofeedback therapy utilizing a multimodal approach will be proposed that contains facial and emotional recognition with voice analysis.

Reference

[1] D. E. Lieberman, "Testing Hypotheses about Recent Human Evolution from Skulls: Integrating Morphology, Function, Development, and Phylogeny," Current Anthropology, vol. 36, no. 2, pp. 159–197, 1995.

[2] K. R. Scherer, et al., "Vocal Cues in Emotion Encoding and Decoding," Motivation and Emotion, vol. 15, no. 2, pp. 123–148, June 1991.

[3] K. R. Scherer, "Vocal Cues to Speaker Affect: Testing Two Models," The Journal of the Acoustical Society of America, vol. 76, no. 5, pp. 1346–1356, Nov. 1984.

[4] A. Abbott, "Stress and the city: Urban decay," Nature News, 2012

[5] M. E. Kemeny, "The Psychobiology of Stress," Current Directions in Psychological Science, vol. 12, no. 4, Aug. 2003. Available:https://journals.sagepub.com/doi/abs/10.1111/1467-8721.01246.

[6] A. R. Subhani, et al., "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," IEEE Access, vol. 5, pp. 13545–13556, Jan. 2017.

[7] N. Souissi and A. Cherif, "Artificial neural networks and support vector machine for voice disorders identification," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016.

[8] M. Stanek and M. Sigmund, "Psychological Stress Detection in Speech Using Return-To-Opening Phase Ratios in Glottis," Elektronika Ir Elektrotechnika, vol. 21, no. 5, Oct. 2015.

[9] K. Masood and M. A. Alghamdi, "Modeling Mental Stress Using a Deep Learning Framework," IEEE Access, vol. 7, pp. 68446–68454, 2019.

[10] N. Ferguson and B. Schneier, Practical Cryptography. Indianapolis, IN: Wiley, 2003.

[11] C. Sampaio, "Implementing SVM and kernel SVM with Python's Scikit-Learn," Stack Abuse, 2023.

[12] "Stress and high blood pressure: What's the connection?," Mayo Clinic, 2022. Available: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/stress-and-high-blood-pressure/art-20044190.

[13] N. P. Dhole and S. N. Kale, "Stress Detection in Speech Signal Using Machine Learning and AI," in Machine Learning and Information Processing: Proceedings of ICMLIP 2019, Springer Singapore, 2020, pp. 11–26.

[14] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, and S. Hawila, "Stress detection through speech analysis," in ICETE (1), 2018, pp. 560–564.

[15] H. Han, K. Byun, and H.-G. Kang, "A deep learning-based stress detection algorithm with speech signal," in Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, 2018, pp. 11–15.

[16] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 209–214.

[17] P. P. Singh and P. Rani, "An approach to extract feature using MFCC," IOSR Journal of Engineering, vol. 4, no. 8, pp. 21–25, 2014.

[18] P. Rajasekaran, G. Doddington, and J. Picone, "Recognition of speech under stress and in noise," in ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 11, 1986, pp. 733–736.

[19] V. Vapnik, Estimation of Dependences Based on Empirical Data. Springer Science & Business Media, 2006.

[20] Sigmund, M., 2007. Spectral Analysis of Speech under Stress. IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4.

[21] Shah A, Firoz &. A, Raji & Anto, Babu. (2009). Automatic Stress Detection from Speech by Using Discrete Wavelet Transforms.