IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Stress Detection through Speech Analysis using Machine Learning

Dr. S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas

Associate Professor, Student, Student, Student, Student
Department of Computer Engineering,
Datta Meghe College of Engineering, Navi-Mumbai, India

Abstract: Voice stress analysis (VSA) is collectively a pseudoscientific technology that aims to infer deception from stress measured in the voice. The technology aims to differentiate between stressed and non-stressed outputs in response to stimuli (e.g., questions posed), with high stress seen as an indication of deception. In this work, we propose a deep learning-based psychological stress detection model using speech signals. With increasing demands for communication between humans and intelligent systems, automatic stress detection is becoming an interesting research topic. Stress can be reliably detected by measuring the level of specific hormones (e.g., cortisol), but this is not a convenient method for the detection of stress in human- machine interactions. The proposed algorithm first extracts Mel-filter bank coefficients using pre-processed speech data and then predicts the status of stress output using a binary decision criterion (i.e., stressed or unstressed) using CNN (Convolutional Neural Network) and dense fully connected layer networks.

***Keywords* - Stress Detection, CNN, MFCC, RAVDESS, MFC, Cortisol.**

I. Introduction

In recent years, interest in artificial intelligence-aided health monitoring or psychological counselling systems has increased due to the convenience and efficiency of machine learning- based algorithms. To provide appropriate services in these areas, the mental state of the user must be detectable. Among various emotional states, we focus on methodology to detect the user's stress status based only on speech signals. Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion plays in shaping human social interaction. Emotional displays convey considerable information about the mental state of an individual.

The use of speech signals for detecting stress has advantages and disadvantages. Unlike bio-signal based approaches, speech signals can be easily detected using microphones, without directly attaching them to the body. This feature is not only convenient for users but also helpful for building a large database for use in a stress-detection system. However, the accuracy of speech-based stress-detection systems is generally worse than that of bio-signal based algorithms. Despite this performance hurdle faced by speech- based stress-detection systems, developing neural network- based techniques by collecting a large amount of data makes such systems more promising.

There are several emotions in human speech, depending on the various situations. All these emotions are primarily classified into seven basic archetypal emotions i.e., anger, happiness, surprise, disgust, fear, sadness, neutral and clam [1], [2]. The extraction of features from a speech signal to depict the emotional state of a speaker is a significant issue that is to be taken into consideration in a speech recognition system. The use of speech signals for detecting stress has some specified advantages and disadvantages. Unlike bio-signal based approaches, speech signals can be easily detected using microphones, without directly attaching them to the body. This feature is not only convenient for users but also helpful for building a large database for use in a stress-detection system. However, the accuracy of speech-based stress-detection systems is generally worse than that of bio-signal based systems algorithms. Despite this performance hurdle faced by speech- based stress-detection systems, developing neural network-based techniques by collecting a large amount of data makes such systems more promising.

Speech is a type of signal containing temporal contextual information and dependency between neighboring frames. As a CNN structure exploits a hidden state that acts as a memory storing information over time, it can handle contextual information, such as speech, that shows specific windowed frame variation [7]. We introduce two types of many-to-one structures with a CNN structure to capture the long-term temporal features of speech. The outputs are then delivered to fully connected layers to make a final decision using a softmax layer.

II. Background

The basic Speech Emotion Detection (SED) system consists of the pre-processing system, feature extraction and classifier block as shown in figure. The raw voice/speech quantities in speech signal after pre-processing stage are given to the feature extraction block. Here, F1, F2... Fn are the features extracted and these are given to the classifier section. Finally, a particular emotion is detected using this classifier.

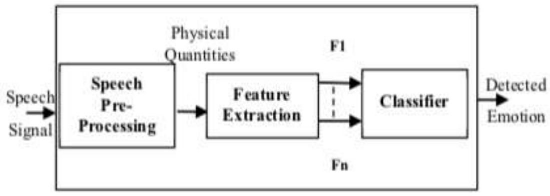


Figure 2.1: Speech Emotion Detection System

Speech is a type of signal containing temporal contextual information and dependency between neighboring frames. The speech signal is pre-processed before giving it to the feature extraction module so as to improve the efficiency and accuracy of the feature extraction process. The pre-processing stages are Filtering, Framing and Windowing. The raw voice quantities like pitch, energy and formants are obtained from the speech signal after pre-processing stage [4]. Filtering is the process used to reduce the noise in a speech signal that occurs due to the disturbances in the environment or during the recording of the speech sample. The purpose of a pre- emphasis filter is to boost the energy of the speech signal in the higher frequencies which are attenuated during the speech signal production from vocal tract.

A. *Collection of Data*

It is not easy to find a database that can be classified into stressed or unstressed conditions from the same speaker. In this work, we used a multi-modal database for use in stress-related research using audio-visual (speech and video) named as RAVDESS.

B. *Database Description*

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 females, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio- only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Note, there are no song files for Actor_18.

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest inter-rater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli.

Audio-only files

Audio-only files of all actors (01-24) are available as two separate zip files (~200 MB each):

- Speech file (Audio_Speech_Actors_01-24.zip, 215 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440.
- Song file (Audio_Song_Actors_01-24.zip, 198 MB) contains 1012 files: 44 trials per actor x 23 actors = 1012.

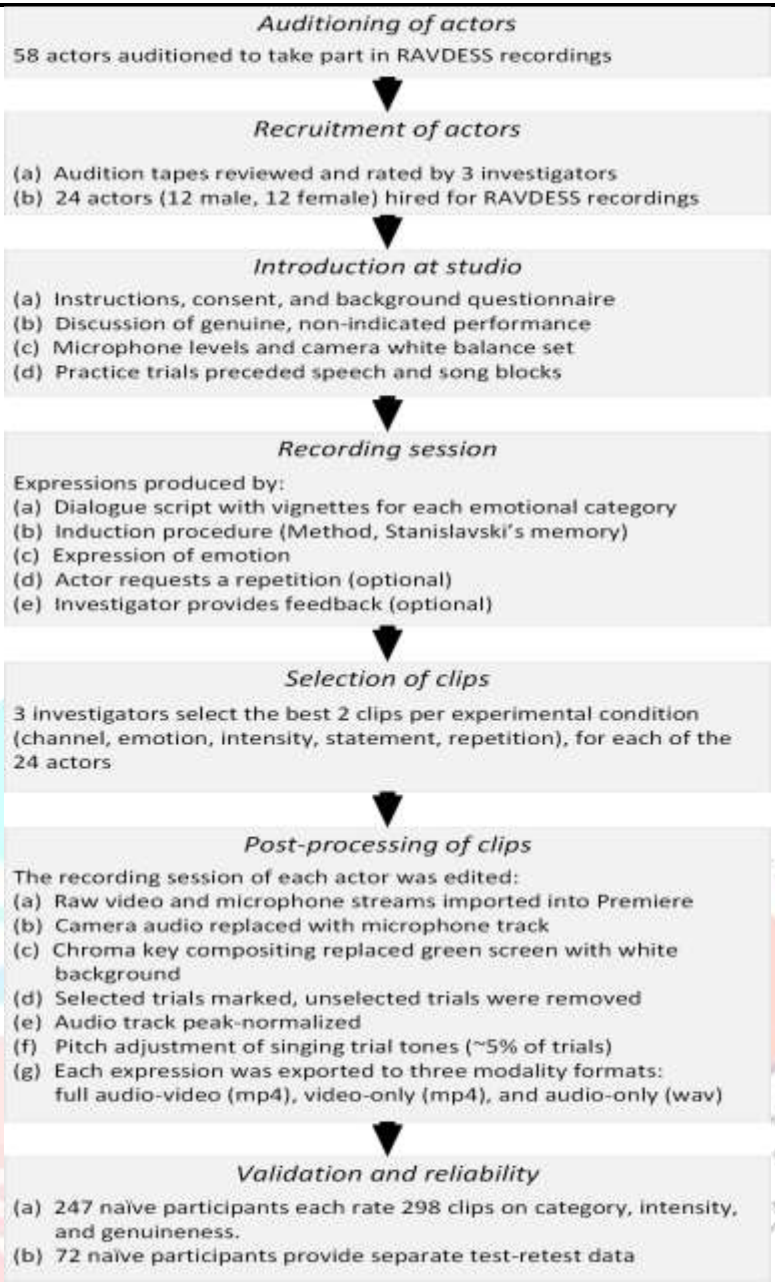
Audio-Visual and Video-only files

Video files are provided as separate zip downloads for each actor (01-24, ~500 MB each), and are split into separate speech and song downloads.

- Speech files (Video_Speech_Actor_01.zip to Video_Speech_Actor_24.zip) collectively contains 2880 files: 60 trials per actor x 2 modalities (AV, VO) x 24 actors = 2880.
- Song files (Video_Song_Actor_01.zip to Video_Song_Actor_24.zip) collectively contains 2024 files: 44 trials per actor x 2 modalities (AV, VO) x 23 actors = 2024.

File Summary

- In total, the RAVDESS collection includes 7356 files (2880+2024+1440+1012 files).



Auditioning of actors

58 actors auditioned to take part in RAVDESS recordings

Recruitment of actors

(a) Audition tapes reviewed and rated by 3 investigators
(b) 24 actors (12 male, 12 female) hired for RAVDESS recordings

Introduction at studio

(a) Instructions, consent, and background questionnaire
(b) Discussion of genuine, non-indicated performance
(c) Microphone levels and camera white balance set
(d) Practice trials preceded speech and song blocks

Recording session

Expressions produced by:
(a) Dialogue script with vignettes for each emotional category
(b) Induction procedure (Method, Stanislavski's memory)
(c) Expression of emotion
(d) Actor requests a repetition (optional)
(e) Investigator provides feedback (optional)

Selection of clips

3 investigators select the best 2 clips per experimental condition (channel, emotion, intensity, statement, repetition), for each of the 24 actors

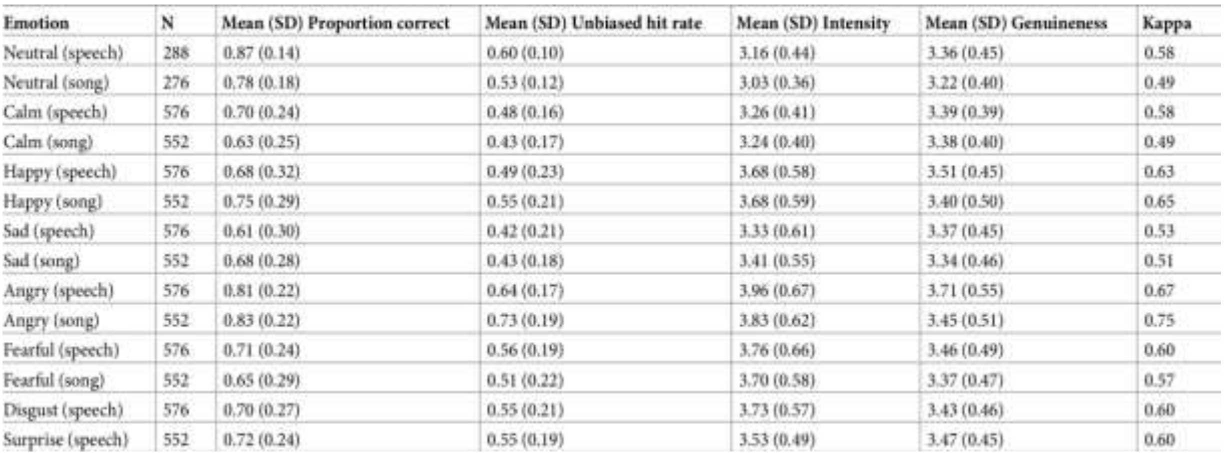
Post-processing of clips

The recording session of each actor was edited:
(a) Raw video and microphone streams imported into Premiere
(b) Camera audio replaced with microphone track
(c) Chroma key compositing replaced green screen with white background
(d) Selected trials marked, unselected trials were removed
(e) Audio track peak-normalized
(f) Pitch adjustment of singing trial tones (~5% of trials)
(g) Each expression was exported to three modality formats: full audio-video (mp4), video-only (mp4), and audio-only (wav)

Validation and reliability

(a) 247 naïve participants each rate 298 clips on category, intensity, and genuineness.
(b) 72 naïve participants provide separate test-retest data

Figure 2.2: Flowchart of RAVDESS Creation & Validation



Emotion	N	Mean (SD) Proportion correct	Mean (SD) Unbiased hit rate	Mean (SD) Intensity	Mean (SD) Genuineness	Kappa
Neutral (speech)	288	0.87 (0.14)	0.60 (0.10)	3.16 (0.44)	3.36 (0.45)	0.58
Neutral (song)	276	0.78 (0.18)	0.53 (0.12)	3.03 (0.36)	3.22 (0.40)	0.49
Calm (speech)	576	0.70 (0.24)	0.48 (0.16)	3.26 (0.41)	3.39 (0.39)	0.58
Calm (song)	552	0.63 (0.25)	0.43 (0.17)	3.24 (0.40)	3.38 (0.40)	0.49
Happy (speech)	576	0.68 (0.32)	0.49 (0.23)	3.68 (0.58)	3.51 (0.45)	0.63
Happy (song)	552	0.75 (0.29)	0.55 (0.21)	3.68 (0.59)	3.40 (0.50)	0.65
Sad (speech)	576	0.61 (0.30)	0.42 (0.21)	3.33 (0.61)	3.37 (0.45)	0.53
Sad (song)	552	0.68 (0.28)	0.43 (0.18)	3.41 (0.55)	3.34 (0.46)	0.51
Angry (speech)	576	0.81 (0.22)	0.64 (0.17)	3.96 (0.67)	3.71 (0.55)	0.67
Angry (song)	552	0.83 (0.22)	0.73 (0.19)	3.83 (0.62)	3.45 (0.51)	0.75
Fearful (speech)	576	0.71 (0.24)	0.56 (0.19)	3.76 (0.66)	3.46 (0.49)	0.60
Fearful (song)	552	0.65 (0.29)	0.51 (0.22)	3.70 (0.58)	3.37 (0.47)	0.57
Disgust (speech)	576	0.70 (0.27)	0.55 (0.21)	3.73 (0.57)	3.43 (0.46)	0.60
Surprise (speech)	552	0.72 (0.24)	0.55 (0.19)	3.53 (0.49)	3.47 (0.45)	0.60

Figure 2.3: Validity Ratings and Iterator Reliability

C. Mel-Frequency Cepstral Coefficients (MFCC)

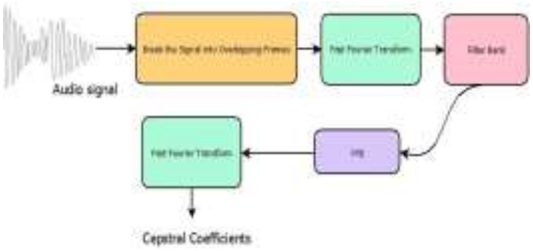


Figure 2.4: MFCC Features Extraction

In sound processing, the **mel-frequency cepstrum** (**MFC**) 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 nonlinear mel scale of frequency. **Mel-frequency cepstral coefficients** (**MFCCs**) are coefficients that collectively make up an MFC.[1] They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

III. PROPOSED STRESS DETECTION SYSTEM

The speech features (i.e., mel-filterbank coefficients) obtained in the feature extraction module are passed to a deep-learning based stress detection model. The CNN structure determines the user's stress state by a hard decision process. At this time, we decided to make a label-based decision criterion to predict the stress status by putting the stressed state as sad, fearful emotion and the unstressed state as happy, calm. These labels were employed for the training model using a one-hot-encoding approach.

The proposed module is composed of eight CNN layers and fully connected layers. The neural network layers capture the temporal information of the extracted features and calculate the frame-level output at every time sequence f = (f1, f2, ..., f_T). The frame-level output is converted to a sentence-level feature that embraces the entire characteristics before feeding it into the fully-connected layer. Two kinds of features are extracted from the layer's: the average value of the output sequence f_{avg} and the last frame-level output of the layer f_T . These vectors are used as sentence-level features $f_{sent.}$, as they are assumed to contain overall information of the input features such as pitch, sample rate, formant, and energy etc.

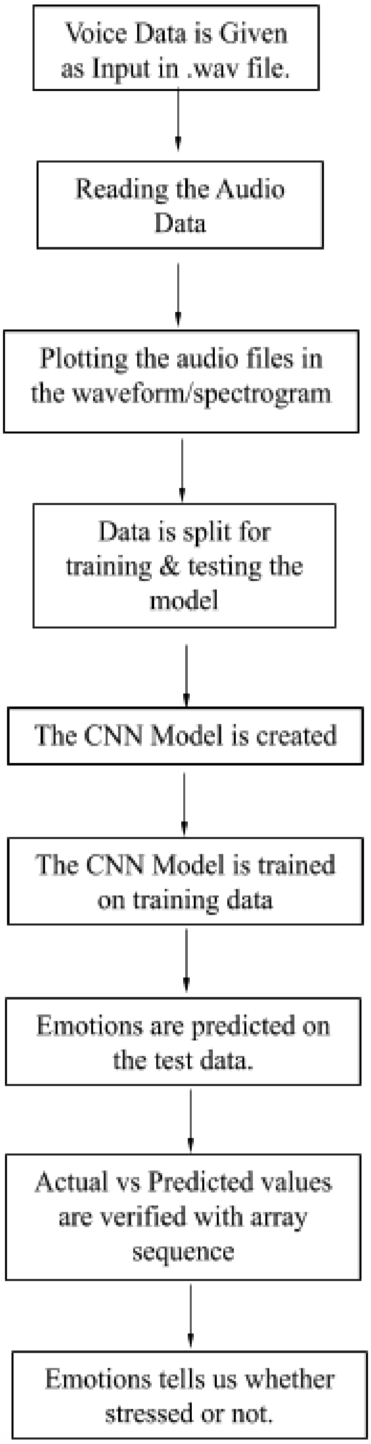

Voice Data is Given as Input in .wav file.
Reading the Audio Data
Plotting the audio files in the waveform/spectrogram
Data is split for training & testing the model
The CNN Model is created
The CNN Model is trained on training data
Emotions are predicted on the test data.
Actual vs Predicted values are verified with array sequence
Emotions tells us whether stressed or not.

Figure 3.1: Flowgraph of the System

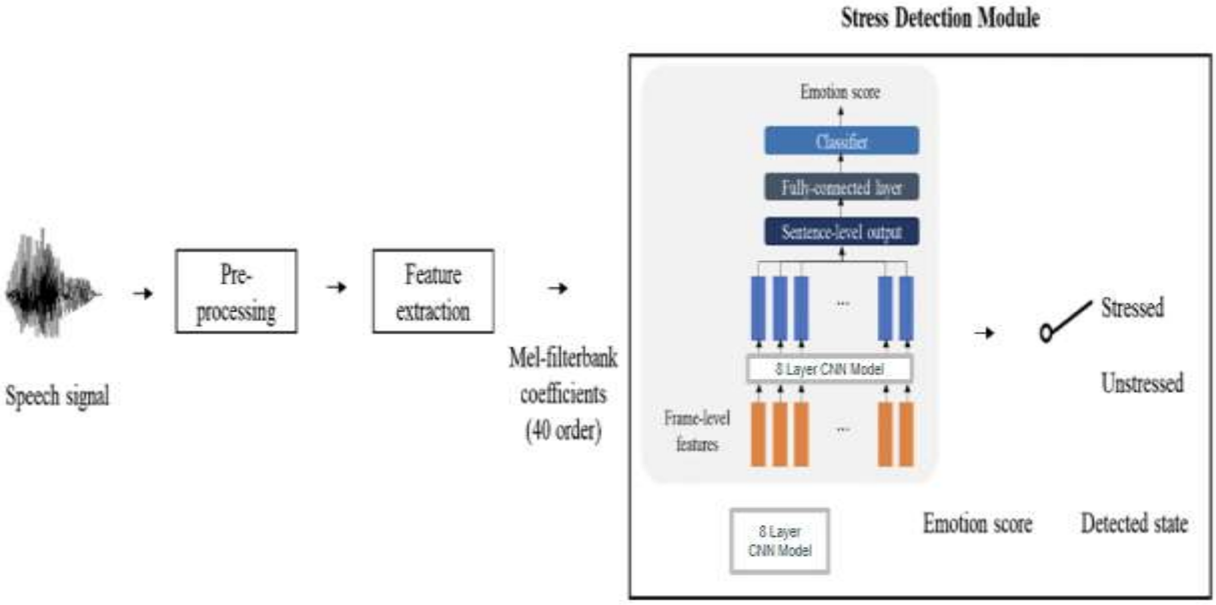


Figure 3.2: Module for Stress Detection Design

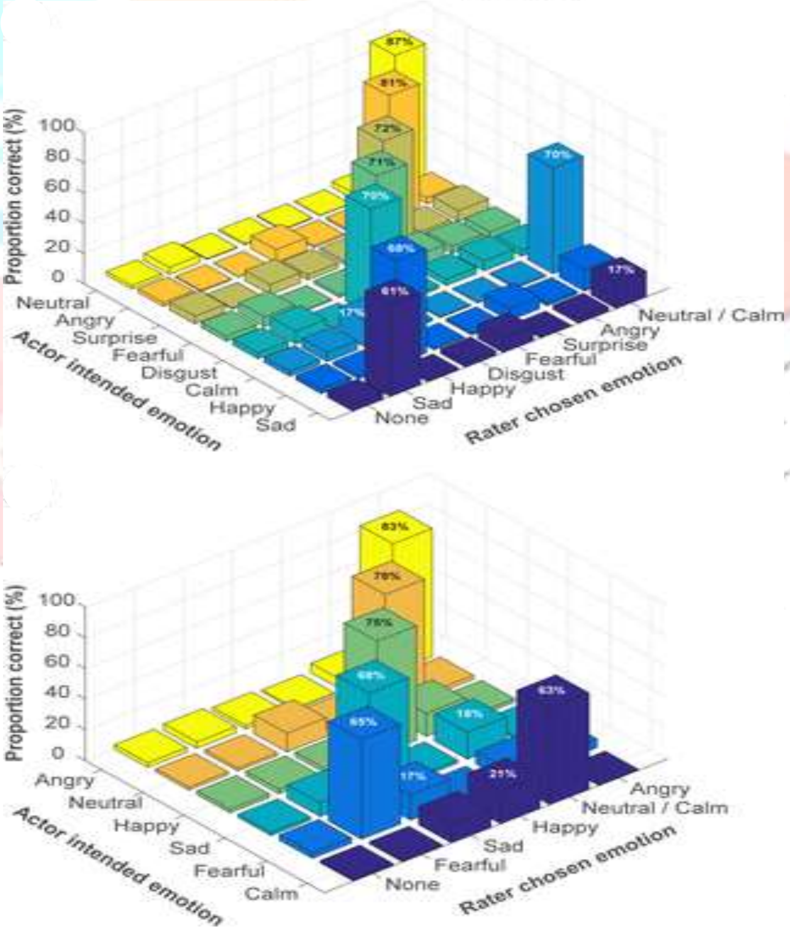


Figure 3.3: Confusion Matrix of Emotional Validity

Table 1: Confusion matrix of proposed system using MFCC Feature Extraction

Classification Accuracy %				
Emotion	**Pitch/Sample Rate**		**MFCC(CNN)**	
	Male	**Female**	**Male**	**Female**
Angry	59.2	62.7	98.75	99.25
Calm	22	36.4	97.5	97
Fearful	45.7	53.1	88.75	89
Happy	62.4	40	98.3	94.56
Sad	57.1	62.1	88.5	91.25
Overall Accuracy	**49.28**	**50.86**	**94.26**	**94.3**

The proposed system is carried out using Ravdess database. Both for the training and testing, 1440 speech utterances of twelve male and twelve female speakers with the emotions namely, anger, anxiety, disgust, neutral and sadness are considered. The confusion matrices of Stressed emotion accuracy using Pitch/ Sample Rate, MFCC is shown in **Table 1** with male and female speakers respectively.

From the results, it can be observed that, the classification accuracy of the Stress Detection system using pitch or sample rate is 52% for both male and female speakers, and that for MFCC is 94.33 % for both male and female speakers. From this analysis it is clear that by introducing the signal raw energy operator the stressed emotions are detected with improved accuracy.

CONCLUSION

Stress is an uncomfortable emotional state that is accompanied by biochemical, physiological, and behavioral changes. To detect the stressed state using only speech signals, we used audio-visual data through a multi-step process and established stress-detection models using deep learning frameworks using an CNN structure. The stress status (i.e., stress vs. unstressed) will be detected by a labelled classification task with emotion labels assigned. Thus, the approach can likely be extended to a better conceptual multi-modal based approach to further increase the detection accuracy. More reliable experimental results could be obtained by professional measurement on the variation in cortisol levels in each raw audio stage. In the future, we will consider all these aspects to build an improved model for stress detection.

REFERENCES

ASME Standard
Journal Paper,

[1] A.Baum. Stress, intrusive imagery, and chronic distress. Health psychology, 9(6): 653, 1990.

[2] N.Sharma and T. Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Computer methods and programs in biomedicine, 108(3):1287–1301, 2012.

[3] Khan.M, Vijay.R, Sondhi,S and A.K.Salhan. Vocal indicators of emotional stress. International Journal of Computer Applications, 122(15), 2015.

[4] I. R. Murray and Arnott J. L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America, 93(2):1097–1108, 1993.

[5] Andrews S., Ellis D., Dobrescu R., Sandulescu V. and O. Martinez-Mozos. Mobile app for stress monitoring using voice features. In E-Health and Bioengineering Conference (EHB), 2015, pages 1–4. IEEE, 2015.

[6] Vignolo L. Schlotthauer G., Colominas M., A.Rufiner, H.L.Sharma,R. and S.R.M. Prasanna. Empirical mode decomposition for adaptive am-FM analysis of speech: a review. Speech Communication, 88:39–64, 2017

[7] J. Lee and I. Tashev. High-level feature representation using recurrent neural networks for speech emotion recognition. 2015.

[8] C.N. Anagnostopoulos and T. Iliou. Towards emotion recognition from speech: definition, problems and the materials of research. In Semantics in Adaptive and Personalized Services, pages 127–143. Springer, 2010.

[9] M. Hashemi. Language stress and anxiety among English language learners. Procedia-Social and Behavioral Sciences, 30:1811–1816, 2011.

[10] L. Woodrow. Anxiety and speaking English as a second language. RELC Journal, 37(3):308–328, 2006.

[11] [11] K.Manley. Comparative study of foreign language anxiety in Korean and Chinese students. 2015

[12] M.Boden.A guide to recurrent neural networks and back propagation .The Dallas project, 2002.

IEEE Standard Papers,

[13] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2):157–166, 1994.

[14] S.Hochreiter and J.Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[15] Mohamed A.R.Graves,A.and G.Hinton. Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 IEEE International Conference on, pages 6645–6649. IEEE, 2013.

[16] KumstaR. von Dawans B. Monakhov M. Ebstein R.P.Chen, F.S.and M. Heinrichs. Common oxytocin receptor gene (oxtr) polymorphism and social support interact to reduce stress in humans. Proceedings of the National Academy of Sciences, 108 (50):19937–19942, 2011.

[17] Yamaguchi M. Aragaki T. Eto K. Uchihashi K. Takai, N. and Y. Nishikawa. Effect of psychological stress on the salivary cortisol and amylase levels in healthy young adults. Archives of oral biology, 49(12):963–968, 2004.

[18] L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008

[19] Muray I.R, Baber C, South A. Towards a Definition and Working Model of Stress and its Effects on Speech Communication, Speech Communication Journal, volume 20, Issue 1-2, Nov 1996, pp 3-12.

Dataset Refer Link:
https://zenodo.org/record/1188976#.Xru65GgzaUk