Detecting Human Emotions Through Voice Signals: A Research Proposal.

Contributor: (Asra Aijaz)

AUGUST 2, 2020
SUKKUR IBA UNIVERSITY

Contents

1 Introduction.

In recent years, technological experts are trying to make the interaction of computer with humans the best according to their ability in HCI. Moreover, in the field of robotics manufacturers are making them to recognize the emotions of human and act according to the behavior of human (Chennoor, Madhur et al. 2020). because they want to make their robots as much intelligent as human and act on condition as human mostly do. In addition, they want the same decision-making ability as human. Normally, human emotion can be recognized by the Facial Expression like Face Detection, Feature Extraction, and Emotion recognition from the image (Chennoor, Madhur et al. 2020). but with the advancement of the technology requirements keep change and technological experts tried to recognize the emotions of human through speech signals (Kwon, Chan et al. 2003). It means if we will talk to any expert HCI system, it will recognize our emotions like Fear, Anger, Sadness, Joy, Disgust, surprise, and trust by using the signals of our speech (Kwon, Chan et al. 2003). The following methods are commonly used in recognizing the emotions of human through speech signals (Kwon, Chan et al. 2003).

- Feature Extraction
- Feature selection
- Classification

From the last decades, too much work has been done by the researchers and they do their best to improve the accuracy of emotion recognition through speech singles [1,2] [4-6]. Studies which we have selected, every researcher have tried to improve the quality of speech by removing the noise and the recognition of emotion through the given the speech signals, but no researcher have talked about the missing signals. The main purpose of this study is to propose a solution for the missing signals, how system will act if there are some missing signals. In addition, Human emotion recognition accuracy will be improved, and we will be able to discover more human emotions. Our latest expert systems are working on only 7 human emotions (Saxena, Khanna et al. 2020).

2 Literature Review.

In the recent past, many human recognition techniques and methods were given as human emotion recognition through image (Chennoor, Madhur et al. 2020), and human emotion recognition through audio and video signals (Chennoor, Madhur et al. 2020). Emotion Recognition through image procedure will start from capturing the picture and then feature extraction and classification algorithm will be applied then emotion will be recognized (Chennoor, Madhur et al. 2020). Moreover, audio and video signals will be used separately by video division into frames and audio extractor then simultaneously emotions will be recognized from video frames and audio signals then results will be carried out and combined (Chennoor, Madhur et al. 2020). In this whole process face gestures are playing a major role. In addition, this model was developed for specific purpose and for each particular purpose different model is prepared. Sound signals are main medium for recognizing the speaker or emotions after processing it (Davletcharova, Sugathan et al. 2015). When a speaker is talking, every word which is said by the said the speaker possessed some signals. Their acoustics difference is providing help in recognizing the emotions when speaker utters different same thing under the different emotional situations (Davletcharova, Sugathan et al. 2015). Many studies have stated that the emotions have direct influence on the nervous system, the heart is directly affected by them (Davletcharova, Sugathan et al. 2015). Moreover, through the heart beat we also can get the information about the emotional state of person and vice versa (Davletcharova, Sugathan et al. 2015). In the hospitals there are many patients whom doctors can not touch because of COVID-19 or any other spreading disease. In the field of medical sciences their many applications widely use the speech signal processing to detect the heartbeat of the patient and emotional information's (Davletcharova, Sugathan et al. 2015). In (Davletcharova, Sugathan et al. 2015) study this process is done through RA analysis and AUC analysis and it only recognize the 4 emotions such as neutral, anger, joy, and sadness. Moreover, in (Davletcharova, Sugathan et al. 2015) paper the result is analyzed but we can apply this to individual group of peoples to get best and accurate result and the development of a software-based agent for emotion

detection and heart rate analysis can greatly improve telemedicine-based systems. (Archana and Sahayadhas 2018) study stated that microphone sensors with quantifiable emotion recognition are contributing to the emerging area of research in HCI, it also increases the accuracy of speech emotions recognition compare to state of art. This (Archana and Sahayadhas 2018) has also contributed to reducing the computational complexity of the presented Speech Emotions Recognition (SER) model furthermore, it proposed deep stride convolutional neural network (DSCNN) which learn from spectrogram of speech signals and convolutional neural network (CNN) architecture which uses a dynamic adaptive threshold technique to remove noise and silent signals from speech signals. the proposed (Archana and Sahayadhas 2018) study architecture only learns from spectrogram and it contributes to reducing the noise. In another study (Hegde, Manoj et al. 2006), keeping the idea of voice calls in mind at emergency situations over wireless mess networks it proposed QoS approach. This approach uses speech signal for detecting the emotion from an incoming voice call. The main purpose behind this to detect whether the caller is in a state of extreme panic, moderate panic, or in a normal state of behavior (Hegde, Manoj et al. 2006). This (Hegde, Manoj et al. 2006) study, use several features extracted from the speech signal like the range of pitch variation, energy in the critical bark band, range of the first three formant variations, and speaking rate among others to discriminate between the three emotional states. To extract a smaller collection of features based on the SFS methodology with Bhattacharya distance as a discriminability measure, feature selection is also implemented on the full collection of functions (Hegde, Manoj et al. 2006). In addition, there formulated scheme stated that it provides end-to-end delay performance improvement for panicked call as high as 60% compared to normal calls (Hegde, Manoj et al. 2006).

This (Kwon, Chan et al. 2003) study, used basic three techniques (Feature Extraction, Feature Selection, Classification) for detecting the emotion recognition from speech signals. using these techniques, they make feature streams in one-dimension and extracted statistics used for discriminative classifiers and analyzed by quadratic discriminative analyses (QDA) and support vector machine (SVM). They assumed that stream is one-dimensional signal. In this study pitch and energy are playing major role. in this study is stated that it achieved 42.3% accuracy for 5-class emotion recognition but Furthermore, study is needed to explore new features better representing prosody and timbre. In addition, improve

the pitch and formant tracking algorithms and develop a new sophisticated approach to model dynamics of feature streams and classifier was limited.

During the selection of the studies we found a survey study which was comprised of 86 studies related to human emotions recognition.

This (Saxena, Khanna et al. 2020) study, worked on facial expression recognition, physiological signals recognition, speech signal variations and text semantics on different standard databases. Through these methods seven different basic emotions are recognized and six elementary emotions which human display (sadness, surprise, disgust, happiness, fear, and anger). In this proposed survey study, two major methods were used for recognition Gabor Wavelets and Facial Landmark. Moreover, the complexity of pre-processing the physiological signals is a big challenge for emotion detection through physiological signals and Emotion detection through ECG signals and features such as skin temperatures, Electromyography signals (EMG) which use muscle movement signals are still emerging.

3 Problem Statement

To overcome the problem that solve the problem of limited emotion classification to recognized emotions that is research gap of existing one research, it also proved solution that recognized emotions from signal processing as well as facial recognitions. In this we also classify the voice and facial of male and female accord to that we recognized emotions that are classified in different more than fifteen emotions and provide solution that would accurate with its accuracy more that 70%. Our provided solutions are all for to provide solution to identify research gap and analyze their results and provide solution accord to their future research directions and limitation that we identify after reviewing more than 6 research articles.

4 Research Objectives

In this era of emotion recognition there is a lot of work of already performed many researchers show their expertise in this field of emotion recognition many

researchers doing their job by emotion recognition by text classification and many articles performed there to identify emotions by using synthesis analysis but in proposal our main concerned is to emotion recognition by using signal processing as well as face recognition.

In our idea that we select, we go through many researchers works but during this review we found many limitations such as out of 3 articles from our selected work there is limitation of emotions. In these articles, article (Acharyya, Neehar et al. 2015, Davletcharova, Sugathan et al. 2015) classify only 5 emotions such as neutral, anger, joy and sadness. In article (Saxena, Khanna et al. 2020) there is only classification of only seven emotions. In article **[p4]** there is also limitation of emotion of only 5 classification but in this there is also issue with its accuracy. In article (Chennoor, Madhur et al. 2020) emotion detection system is developed for particular system for every different system in this article emotion is recognized through facial recognition only. In article **[p5]** there is limitation of accuracy of recognition of emotions some missing voice that is not clear or if there is issue in signals are not recognized accurately. There is significant difference in their accuracy. Finally, our objective is to provide neutral solution that easily solve all these issues and behave accurate for all these limitations.

We found our all objectives to provide accurate and beneficial solution to provide better work that would be beneficial for our future researchers and will be helpful for those who want to find future work of our selected reviews. we describe our objectives below.

1) Our objective is to provide solution that classify more than fifteen emotion classification such as (happy, anger, neural, drossy, sleepy, amazed, annoyed, anxious, furious, ashamed, hopeful, proud, hurt, confused, inspired, jealous, lonely, eager).
2) Our next objective is to provide solution is recognized emotion by signal processing as well as facial recognitions.
3) Our objective is to provide solution that prove its accuracy better than existing one developed solution. We are anxious to develop solution that has least one accuracy greater than 70%.
4) Our finally objective is to provide solution that has specialized features for recognition of voice of female and male but identify their frequency and pitch clearly.

Our provided objectives are significantly different than existing one and would be developed solution of future research directions of reviewed articles. It would be unique and best than existing solutions that has features that detect emotions accurately by using defined characteristics.

5 Methodology

Despite the great progress made in artificial intelligence, we are still far from having a natural interaction between man and machine, because the machine does not understand the emotional state of the speaker. Speech emotion detection has been drawing increasing attention, which aims to recognize emotion states from speech signal. The task of speech emotion recognition is very challenging, because it is not clear which speech features are most powerful in distinguishing between emotions.

The fields of Human Computer Interaction (HCI) and Affective Computing are being extensively used to sense human emotions. Humans generally use a lot of indirect and non-verbal means to convey their emotions. The presented exposition aims to provide an overall overview with the analysis of all the noteworthy emotion detection methods at a single location.

5.1 Related Work:

There are multiple ways through which we can determine the emotions and that are given below.

5.1.1 Hybrid PSO assisted Biogeography-based Optimization:

For studying the basic nature of features in speech under different emotional situations, we used data from three subjects. As part of the data collection, we recorded the voice of three different female subject and that are also available there means their faces also detected. The subjects were asked to express certain emotions when their speech was recorded. A mobile phone was used to record the speech and was kept at a distance about 15 CMS away from the mouth.

The proposed algorithm used three databases to test its accuracy. The databases used are BES, SAVEE and SUSAS. BES database has samples of Anxiety, Angry, Happiness, Disgust, Sad, Boredom and Neutral emotions from ten German speakers. SAVEE database has samples of the states- Happiness, Fear, Neutral, Surprise, Sadness, Anxiety and Disgust of four English speakers. SUSAS database has samples of simulated stressful and multi- style dialogue. The silent portions were eliminated before the feature extraction process by establishing a threshold value where each database had its separate threshold value. Linear predictive analysis method along with inverse filtering was applied to extract glottal waveforms. To attend the glottal and speech waveforms spectrally, first order pre-emphasis filter was employed. The resulting waveforms were divided into frames with a 50 percent overlap. Finally, windowing of each frame was done by hamming window method, which helped in reducing the discontinuity and distortion in the signal.

5.1.2 MFFC algorithm:

The analysis of the recorded speech signals was done in a MATLAB environment which provides several graphical visualizations for analyzing a signal.

The Mel-frequency cepstral coefficients (MFCC) are widely used in audio classification experiments due to its good performance. It extracts and represents features of speech signal. The Mel-cepstral takes short-time spectral shape with important data about the quality of voice and production effects. For emotion classification, we have used a bigger dataset for 30 subjects in the age group of 20-45. The subjects consist of an equal proportion of males and females. We recorded the voice of each subject for 30 times. To calculate these coefficients the cosine, transform of real logarithm of the short-term spectrum of energy must be done. Then it is performed in Mel frequency scale. Further, after pre-emphasizing the speech segments are windowed ... All the recorded data was labelled into three categories/classes of emotions: neutral, anger and joy. Instead of including all the different human emotions, we have used only a limited number of emotions as it can clearly reveal the fact that there are certain distinguishing emotion related elements in human speech.

5.1.3 Convolutional Neural Network (CNN)-Based SER

In this section, we present a CNN-based framework, the proposed framework utilizes a discriminative CNN for feature learning scheme using spectrograms to specify the controversial state of the speaker. The proposed stride CNN architecture has input layers, convolutional layers, and fully connected layers followed by a SoftMax classifier. A spectrogram of the speech signal is a 2D representation of the frequencies with respect to time, that have more information than text transcription words for recognizing the emotions of a speaker. Spectrograms hold rich information and such information cannot be extracted and applied when we transform the audio speech signal to text or phonemes. Due to this capability, spectrogram improve the speech emotion recognition. The main idea is to learn high-level discriminative features from speech signals, for this purpose we utilized a CNN architecture to learn high-level features, the spectrogram is well suited for this task.

We clean the audio signals to remove the background noises, silent portion and other irrelevant information from speech signal using the adaptive threshold-based preprocessing .In this method, we find the relationship of energy with amplitude in speech signal using direct relation policy. The energy amplitude relationship is that the amount of energy passed by a wave is correlated to the amplitude of the wave. A high energy wave is considered by a high amplitude; a low energy wave is considered by a low amplitude. The amplitude of a wave mentions the extreme amount of displacement of an element in the middle from its rest location. The logic underlying the energy-amplitude relationship is as follows to remove the silent and unnecessary particle from speech signals. Three steps are included in this process; first, read the audio file step by step with 16,000 sampling rates. In the next step, we find the energy-amplitude relationship in waves and then compute the maximum amplitude in each frame using and passed from a suitable threshold to remove the noises and salient portion and save it in an array. In the last step, we reconstruct a new audio file with the same sample rate without any noise and silent signals.

5.2 Proposed Methodology

5.2.1 Experimental Design

Twenty participants are briefed on the experiment setup and were asked to sign a consent form for taking part in experiments. Then, participants were asked to sit down in a quiet, and controlled room. Prior to the data gathering, each participant is made familiar with the experiment procedure. After that, the sensors are positioned on the head of each participant and sensors are placed on the terminals of the body. Data collection is obtained by a machine called Brain Marker. At first, participants are asked to close their eyes for 1 min and then open for 1 min. afterwards, the picture clips with four basic emotions are displayed to them for 1 min per picture clip. Lastly, the captured brain signals are saved for later analysis.

5.2.2 Electrodes Placement

For ECG electrode placement, 3 electrodes were attached to the body surface of the subject; 1 electrode is attached to the right wrist, 1 electrode is attached to the left wrist and 1 electrode is attached to the left leg. This placement is known as Lead II placement which taken from the standard 12 lead placement. Figure 2demonstrate the block diagram for analyzing the raw ECG signal acquired from the subjects. The signal is pre-processed, and the noise is removed through band pass filter. Then the signal is segmented, and short Fourier transform is applied. KDE and MFCC act as feature extraction. Finally, MLP is employed as classifier.

5.2.3 Feature Extraction

Mel frequency cepstral coefficients (MFCC) and kernel density estimation (KDE) were used for feature extraction in this study. These are normally employed in waveforms dimension minimization applications. A mildest tool is employed to compute the cestrum of the signal. We have employed 12 MFCC coefficients to acquire the appropriate features of the ECG. The MFCC feature extraction method is widely used in speech recognition applications. Mel frequency cepstral

coefficients (MFCC) and kernel density estimation (KDE) were used for feature extraction in this study. These are normally employed in waveforms dimension minimization applications. A mildest tool is employed to compute the cestrum of the signal. We have employed 12 MFCC coefficients to acquire the appropriate features of the ECG. The MFCC feature extraction method is widely used in speech recognition applications.

5.2.4 Classification

The final step in this procedure is the classification of the extracted features having a purposeful and yet effective classifier. Multilayer perceptron (MLP) technique has been selected to categorize the features in order that it can obtain the pre-emotion of the subject which refers to the participant emotion responses. The neural network architecture, MLP with feed forward routes groups the input data onto a set of relevant output. Selecting the right parameters for the number of layers and the number of neurons required for MLP structure is important to make sure a good result come out. Dataset given into the input layer are the 228 features extracted from the MFCC and KDE stage respectively. Each of the input is processed by the MLP by computing the product of input data with its assigned weights in the hidden layers. In this research work, 0.1 was set as the mean square error (mse) having a one hidden layer composed of 10 neurons. In the experiments, eyes close and eyes open data were collected to establish a baseline and determine the pre-emotion or baseline emotion. The data acquired from the individuals are analyzed from the emotion data of calm, happy, sad, and fear.

6 Research Outcomes

After studying lots of articles we came to know to that there are some limitations in their proposed work of determining the emotions of human beings. Now we highlighted the limitations of all proposed methodology and after that discuss the outcomes that we get in our study…

6.1 Hybrid PSO assisted Biogeography-based Optimization

1) In this emotion was detected by face gestures to identify the emotions.

2) Its developed model is used for specific purpose and for each particular purpose different model is prepared.

6.2 MFFC algorithm

1) It only recognized emotions such as neutral, anger, joy, and sadness.

2) Limited Classifier

6.3 Convolutional Neural Network (CNN)-Based SER

1) In this it learns from only spectrogram and it also does not detect drop or slow voice.

6.4 Emotion recognition using ECG:

Emotion recognition based on ECG signals is an important research fields with promising application future. We present an electrocardiogram (ECG) -based emotion recognition system using self-supervised learning. Our proposed architecture consists of two main networks, a signal transformation recognition network, and an emotion recognition network. First, unlabeled data are used to successfully train the former network to detect specific pre-determined signal transformations in the self-supervised learning step. Next, the weights of the convolutional layers of this network are transferred to the emotion recognition network, and two dense layers are trained to classify arousal and valence scores.

We show that our self-supervised approach helps the model learn the ECG feature manifold required for emotion recognition, performing equal or better than the fully supervised version of the model.

:::::::

1) It is applicable for all sorts of mode means it does not depend upon the any mode.

2) In this we used two classifiers

3) Same model is used for all types of emotions

4) It overcome the issues of if sound is missing due to connection and frequencies issues that's why sound is dropped.

7 Research Significance

As technological experts are trying to make interaction of computer with human the best according to their ability in HCI. so, this research will be beneficial for those technological experts and robots manufactures who are trying their best to make the robot behave according to the behavior of human.

The main purpose of this study is to propose a solution for missing signals, how system will act if there are some missing signals. so, this research proposal will help in to get human recognition accuracy improved and will be able to discover more human emotions.

8 Research Timeline

The timeline of our research proposal is as follow:

- **2 Weeks:** So, first two week we only focus on literature review to identify gaps in the knowledge
- **1 Weeks:** Identify specific aims of projects based on our research vision, plan and literature review results.
- **2 Weeks:** Next two week we write a proposal draft, then all four of us review that draft.
- **1 Weeks:** In this week we write our final research proposal based on all the review and identified gaps

9 References

- ✔ Archana, K., and A. Sahayadhas (2018). "Automatic rice leaf disease segmentation using image processing techniques." International Journal of Engineering & Technology **7**(3.27): 182-185.

- ✔ Chennoor, S. N., et al. (2020). "Human Emotion Detection from Audio and Video Signals." arXiv preprint arXiv:2006.11871.

- ✔ Davletcharova, A., et al. (2015). "Detection and analysis of emotion from speech signals." arXiv preprint arXiv:1506.06832.

- ✔ Hegde, R., et al. (2006). Emotion detection from speech signals and its applications in supporting enhanced QoS in emergency response. Proceedings of the 3rd International ISCRAM Conference.

- ✔ Kwon, O.-W., et al. (2003). Emotion recognition by speech signals. Eighth European Conference on Speech Communication and Technology.

- ✔ Saxena, A., et al. (2020). "Emotion recognition and detection methods: A comprehensive survey." Journal of Artificial Intelligence and Systems **2**(1): 53-79.

- ✔ AM AlzeerAlhouseini et al. “Emotion detection using physiological signalsEEG & ECG”. In: J. Clinical Neurophysiology33.4 (2016), pp. 308–311.

- ✔ Assel Davletcharova et al. “Detection and analysis of emotion from speechsignals”. In: arXiv preprint arXiv:1506.06832(2015)

- ✔ Soonil Kwon et al. “A CNN-assisted enhanced audio signal processing forspeech emotion recognition”. In: Sensors20.1 (2020), p. 183.

✔ Acharyya, A., et al. (2015). An accurate clustering algorithm for fast protein-profiling using SCICA on MALDI-TOF. 2015 IEEE International Symposium on Circuits and Systems (ISCAS).

✔ Chennoor, S. N., et al. (2020). "Human Emotion Detection from Audio and Video Signals." arXiv preprint arXiv:2006.11871.

✔ Davletcharova, A., et al. (2015). "Detection and analysis of emotion from speech signals." arXiv preprint arXiv:1506.06832.

✔ Saxena, A., et al. (2020). "Emotion recognition and detection methods: A comprehensive survey." Journal of Artificial Intelligence and Systems **2**(1): 53-79.