RESEARCH ARTICLE OPEN ACCESS

Automated Speech Emotion Recognition App Development on Smart Phones using Cloud Computing

Humaid Alshamsi[1] , Veton Kepuska[2] and Hazza Alshamisi[3]
(*Electrical & Computer Engineering Department, Florida Institute of Technology, Melbourne FL, USA*)[1,2,3]

ABSTRACT
Speech Emotion Recognition (SER) has become a major endeavor in Human-Computer Interaction (HCI) and speech processing. Accurate SER is essential for many applications, such as assessingcustomer satisfaction with the quality of services and detecting/assessing the emotional state of children in care. The large number of studiespublished on SER reflects the demand for its use. In thispaper, the proposed system presents a novelmethod of speech recognition based on the cloud model, in combination with the traditional speech emotion system. The process of predictingemotionsfrom speech emotion audio files containsseveral stages. The first stage of this system is the pre-processing stage, whichisapplied by detecting the speech in an audio file and thenreducing the noise. The second stage involvesextractingfeaturesfrom speech emotion files using the Mel-frequency cepstral coefficient (MFCC) feature extraction algorithms. This generates the training and testingdatasetsthatcontain the emotions of Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. Support Vector Machine (SVM) classifiers are thenused for the classification stage in order to predict the emotion. In addition, a Confusion Matrix (CM) technique isused to evaluate the performance of theseclassifiers. The proposed system istested on SAVEEand RMLdatabases and achieved a prediction rate of 95.3%

Keywords-Machine Learning ; Speech EmotionRecognition ; Mel-frequency cepstral coefficient ; Speech Processing; MobileComputing.

I. INTRODUCTION

There is an extensive list of applications that harness the use of Speech Emotion Recognition, and this makes it both an interesting and relevant topic of research. To a degree, emotional recognition is a simple task because humans have a natural ability with respect to speech information analysis. However, using these speech signals to determine emotion is a challenge for any machine because it does not have the same level of intellect and cannot identify the speaker's emotional state [1].

Both speaker identification and speech recognition will help the computer to identify who and what is being said. However, if that machine is then equipped with a system that can detect speech emotion, it could potentially detect how the words are being said [2] and pick up any emotional messaging or signals as well. The definition of Speech Emotion Recognition can be explained as the extraction of the present emotional state of the individual who is speaking, by using that persons' speech signals.

Human Machine interaction remains as the most significant application for Speech Emotion Recognition, specifically for the purpose of making that system more effective for that purpose. An alternative application of the SER system can be found in lie detection systems, intelligent toys and games, psychiatric diagnostics, and call centers [3]. There are additionally other systems that have been purposed for the extraction of emotion from speech, all of which were using alternative classifiers and features. Prosodic and spectral features are also ideal for the recognition of emotion in speech, this is due to the fact that these specific features also have emotional data contained within them. Mel-Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients, also known hereon as MCFF and LPCC are just a few of the relevant spectral features. Loudness, Speech Intensity, Glottal Parameters, and Fundamental Frequency are the prosodic features that are used to model the various emotions [4].

There are a number of classifiers that are available for SER. Namely, these are Artificial Neural Network (ANN), Hidden Markov Model (HMM), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The emotional classification was undertaken using HMM, in Schuller et al., this achieved a recognition rate of 86%. However, this study was limited to features that related to energy and pitch only [5]. A further study was conducted in Shen et al., and an experiment was performed on the Berlin Emotional Database. Emotional classification through speech was rated just using the SVM classifier and resulting in an overall recognition rate of approximately 83% [6].

In this paper, the SVM classifier was utilized in order to classify the different emotion states (Sadness, Fear, Disgust Anger, Surprise, Happiness, and Neutral). These seven states are widely regarded as the seven basic emotions. Mel-frequency cepstral coefficients (MFCC), features that relate to energy are some of those which were used for the SER system. The classification rates for both were observed and documented.

The remainder of the paper is set out as follows: The second section gives details regarding the SER system. The third section provides information about the feature extraction used within emotional classification process. Section four contains in-depth information about the use of the Support Vector Machine for emotional classification. The fifth section discusses the datasets that were used in this paper. The sixth section discusses results of the experiments that were obtained throughout this study. In the final section, we conclude the paper with future work.

II. THE SPEECH EMOTION RECOGNITION SYSTEM

Figure 1 illustrates the overall structure of the SER system that is considered in this particular study. The core elements of the SER are identical to any typical pattern recognition system. The input is the emotional speech, there is feature extraction, and the classifier of the emotions is the Support Vector Machine with the output being the recognized emotion.

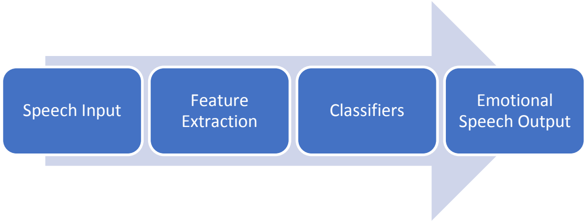


Fig. 1. Overview structure of the Speech emotion recognition.

The overall effectiveness of the SER system is dependent on the databases naturalness that is used within the system. The input of speech into the system could be either from the real world or from the acted data. Following the database collection, also regarded as the training samples, the required features were extracted from the speech signal. The values of these features were then given to the SVM for the classifier training to take place. The test sample that was recorded was then passed to the classifier, which then classified it according to the rules, giving an output as one of the seven recognized emotions.

III. FEATURE EXTRACTION

One of the most important stages in the Speech Emotion Recognition System is the selection of appropriate features that carry the data about emotions from the speech signals. Many researchers have demonstrated that speech energy, formant frequency, fundamental frequency, and Mel frequency cepstrum coefficients are viable parameters for identifying and distinguishing particular emotional states. The feature extraction is founded on segregating speech into frames or smaller intervals [6]. Information regarding speech emotion is held within the pitch signal due to the fact it depends on the vocal fold tension. The vocal fold vibrations are also known as the fundamental frequency. For emotional estimating, the next significant feature is energy. This is due to the fact that there is a change in the speech signal energy when emotions differ. In automated speech recognition and speech emotion recognition, Mel-frequency Cepstral Coefficient (MFCC) is one of the most used spectral features available [7][8], [9]. It has numerous advantages like simple calculation, better ability of distinction, and high robustness to noise. Here MFCC features are extracted from the Praat software [10] with window length 20ms and time step 10ms. Generally, the Hamming window is preferred because of its high-frequency resolution and good sidelobe suppression properties. First, the silence regions present in the database was removed based on the zero–crossing rate and also by thresholding the energy. The silence region does not contain any useful information and is hence removed. Human perception of hearing does not follow a linear scale,and hence MFCC follows the Mel scale [11] which is a frequency scaling having linear spacing below 1000Hz and logarithmic spacing above 1000Hz. The formula to compute the Mel frequency for any given frequency f in Hz is given below [12],

$$Mel\ (f) = 2595\ \times \log\left(1 + \frac{f}{700}\right) \quad (1)$$

The Mel scale filter bank has a triangular series of uniform overlapping filters with constant bandwidth equal to 100 and their center frequencies at 50. This is what is believed to occur in the human auditory system [13]. This corresponds to the spacing on the Mel frequency scale.

IV. SVM TRAINING AND CLASSIFICATION

This is a simplistic and effective computation of an algorithm of machine learning that is utilized for classification and pattern recognition purposes. The Support Vector Machine has the advantage of having a very good level of performant training data. The main idea of SVM [14],[15] is to transform the original input set to a high dimensional feature space by using a kernel function, in which input space consisting of input samples is converted into high dimensional feature space,and therefore the input samples become linearly separable [16] [6] . It is clearly explained by using an optimal separation hyperplane in Figure 2. The main advantage of SVM is that it has limited training data and hence has very good classification performance. For linearly separable data points, classification is done by using the following formula [17],

$$\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1 \qquad (2)$$

$$\langle w \cdot x \rangle + b_0 \geq -1, \forall y = -1 (3)$$

Where, (x, y) is the pair of the training set. Here, $x \in R^n$ and $y \in \{+1, -1\}$.

$\langle w \cdot x \rangle$ represent the inner product of w and x whereas b_0 refers to the bias condition.

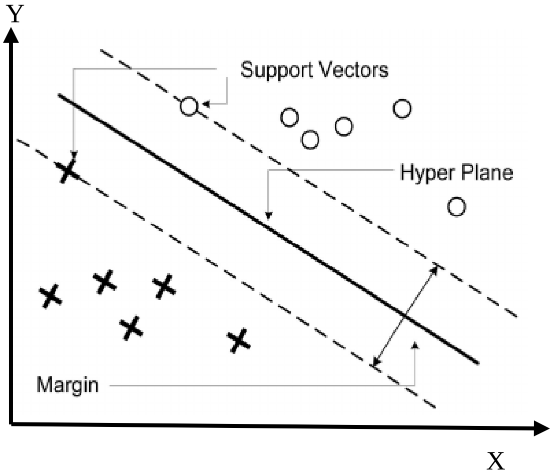


Fig. 2. Support Vector Machine structure

SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function [18] is used here. The linear Kernel function is given by the formula below,

$$\text{Kernel } (x, y) = (x \cdot y) \qquad (4)$$

The radial basis kernel function is given by the following formula,

$$\text{Kernel } (x, y) = e^{\frac{-\|x-y\|^2}{2\sigma^2}} \qquad (5)$$

V. DATASETS

Typically, the emotional database within emotional recognition is applied towards to the study of phonetics and acoustics, along with research and development in the area of emotion speech recognition systems. For the purpose of this research, both RML and SAVEE DB have been studied.

1. *SAVEE - Surrey Audio-Visual Expressed Emotion Database*

Consisting of four male actors, aged between 27-31, in seven different emotions, and 480 different British speeches that were selected from the TIMIT DB. The SAVEE [19] databases have recorded video, audio, and audio-video. The database samples are 60 fps for video and 44.1 kHz for audio. The classification of the audio files for the seven different emotions is listed here. Disgust – 60. Happy – 60. Angry – 60. Neutral – 120. Sadness – 60. Fear – 60. Surprise – 60.

Only audio was used in our experiments. The data from all speakers was randomly split into training (70%) and test (30%) sets.

2. *Ryerson Multimedia Research Laboratory (RML) Database*

Ryerson Multimedia Research Laboratory (RML) also makes ongoing efforts to build multimodal databases related to emotion recognition. The RML emotion database is language and cultural background independent audiovisual emotion database [20]. The video samples were collected from eight human subjects, speaking six different languages and six basic human emotions are expressed. It contains 720 audiovisual emotional expression samples.

VI. EXPERIMENTAL EVALUATION

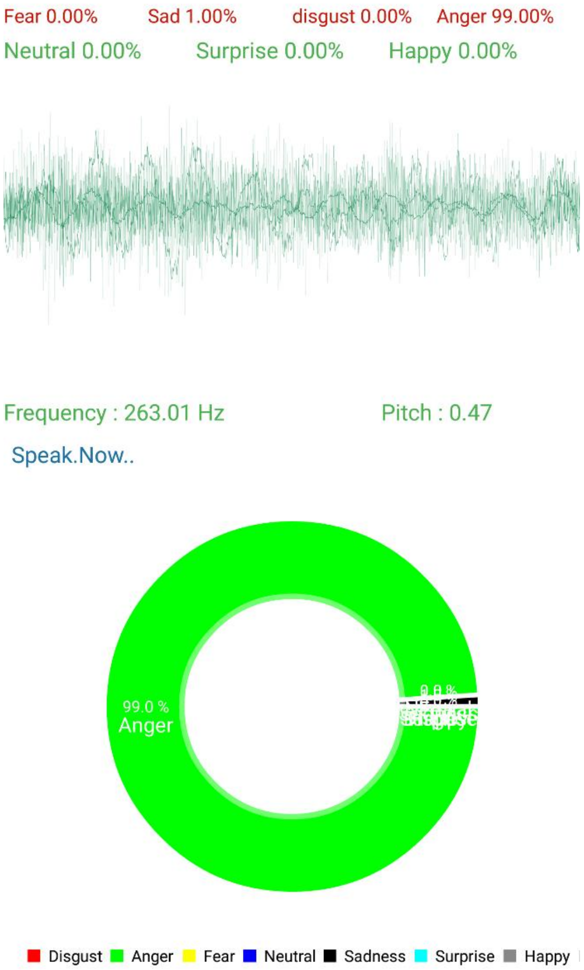
A. *Emotion Classification*

In this paper we recognize the mood of a user through their voice, on the basis of their mood and classify this into the classifier from a smartphone; using cloud computing to compare the actual result that is taken in real time with the dataset that has been stored in the cloud:

Anger – Fear – Disgust – Happiness – Sadness – Neutral – Surprise.

1) *Anger Emotion*

Anger requires high energy to be expressed. The definition and meaning of anger are simple and extreme displeasure [21]. In the case of anger, aggression increases in which the control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, happiness, and sadness. The widest observed pitch range and highest observed rate of pitch change are other findings of the emotion label anger when compared with other emotions [21]. Besides this, a faster speech rate is observed in angry speeches.



Fear 0.00% Sad 1.00% disgust 0.00% Anger 99.00%

Neutral 0.00% Surprise 0.00% Happy 0.00%

Frequency : 263.01 Hz Pitch : 0.47

Speak.Now..

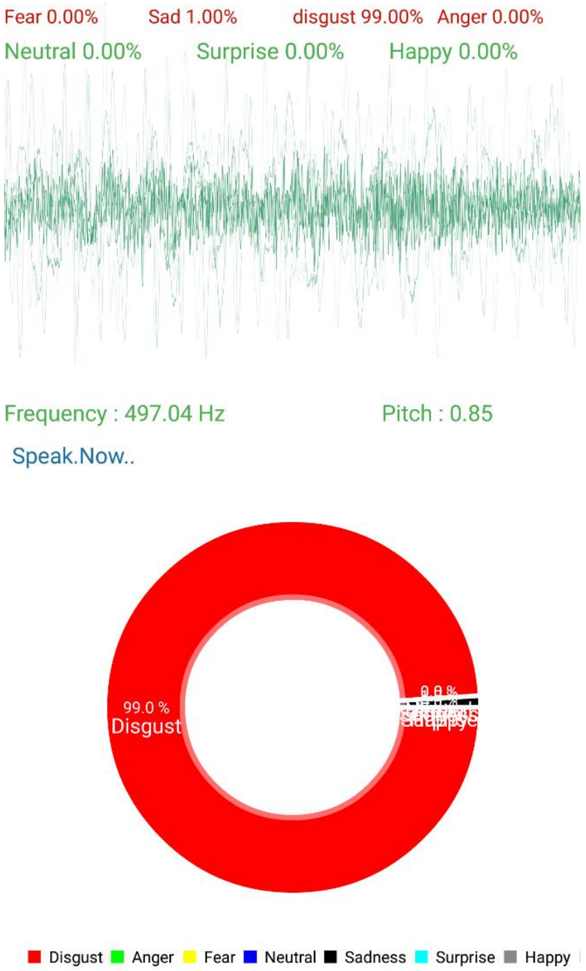
Fig. 3. Anger Emotion

2) Disgust Emotion

Increased articulation precision at stressed content words was noted. The pitch contour showed downward inflections at the phrase endings, and also downward pitch inflections at word endings. A rise in pitch was noted at the beginning of stressed content words. The speech rate was low, with a large number of introduced pauses, increased phonation time, and lengthening of the stressed syllables in stressed content words.

Intensity was quite loud, which is not a typical characteristic for disgust [23], though it decreased towards the end of the utterance. For all of the disgust utterances, an increase in articulation precision was noted. Again, there was an emphasis on the pitch contour changes, with much accenting and use of high intensity. Large dynamic changes within the intensity contour were noted, and variations in contour between utterances.

In [22], low mean pitch level, a low-intensity level, and a slower speech rate are observed when disgust is compared with the neutral state. Disgust is stated the lowest observed speech rate and increased pause length [21].



Fear 0.00% Sad 1.00% disgust 99.00% Anger 0.00%

Neutral 0.00% Surprise 0.00% Happy 0.00%

Frequency : 497.04 Hz Pitch : 0.85

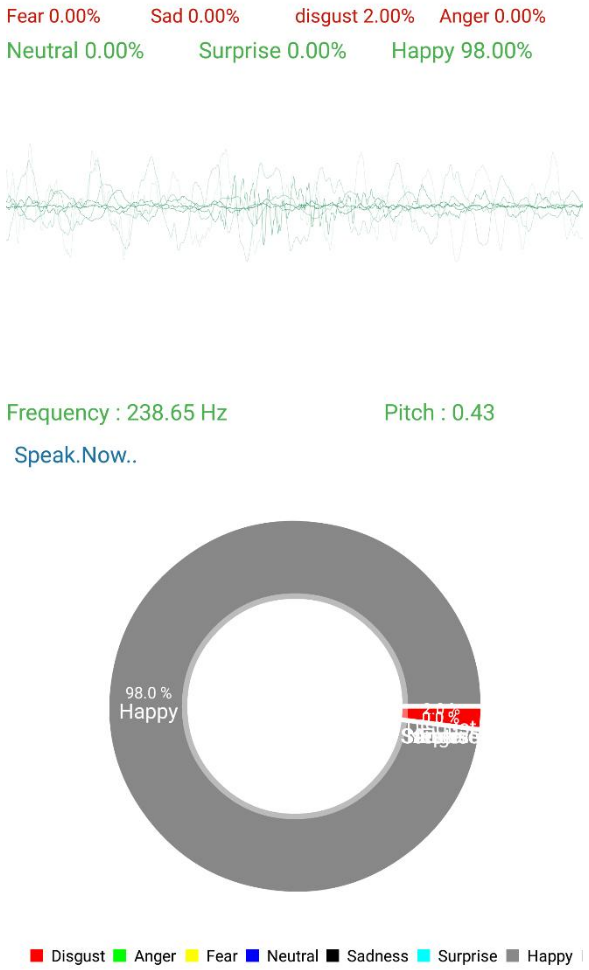
Speak.Now..

Fig. 4. Disgust Emotion

3) Happiness Emotion

An increase in articulation precision was noted for content words, and the voice generally soundedbreathy. The general form of the pitch contour was with a second rising part towards the end of theutterance, with a terminal fall. It was noted that the line of the pitch contour was not smooth; it had sharpsmall oscillations at the primary stressed syllables and local downward pitch changes which seem to berhythmic (stressedphonemes occurring at regular intervals).

Happiness exhibit a pattern with a high activation energy, and positive valence. The strength of the happiness emotion may vary. In the emotional state happiness or joy, pitch mean, range, and variance increases [22]. In [21], it is stated that fundamental and formant frequencies increase in the case of a smile. Moreover, amplitude and duration also increase for some speakers.


Fear 0.00% Sad 0.00% disgust 2.00% Anger 0.00%
Neutral 0.00% Surprise 0.00% Happy 98.00%

Frequency : 238.65 Hz Pitch : 0.43

Speak.Now..

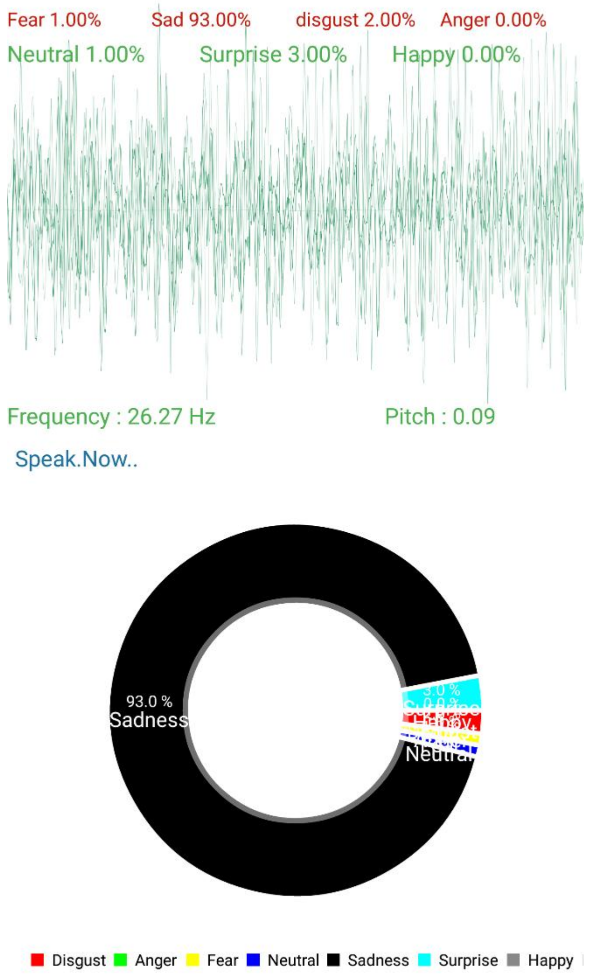
98.0 %
Happy
🟥 Disgust 🟩 Anger 🟨 Fear 🟦 Neutral ⬛ Sadness 🟦 Surprise ⬜ Happy

Fig. 5. Happiness Emotion

4) Sadness Emotion

As the sad voice exhibited an overall decrease in articulation precision. Small downward inflections at phoneme level were noted, and there were regular pauses. Small downward inflections were noted at word and phoneme level. A low-intensity contour was noted for all the utterances, with intensity decreasing towards the ends.

In emotional dimension, sadness requires very low energy. In addition, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo [21]. Speech rate of a sad person is lower than the neutral one [22].


Fear 1.00% Sad 93.00% disgust 2.00% Anger 0.00%
Neutral 1.00% Surprise 3.00% Happy 0.00%

Frequency : 26.27 Hz Pitch : 0.09

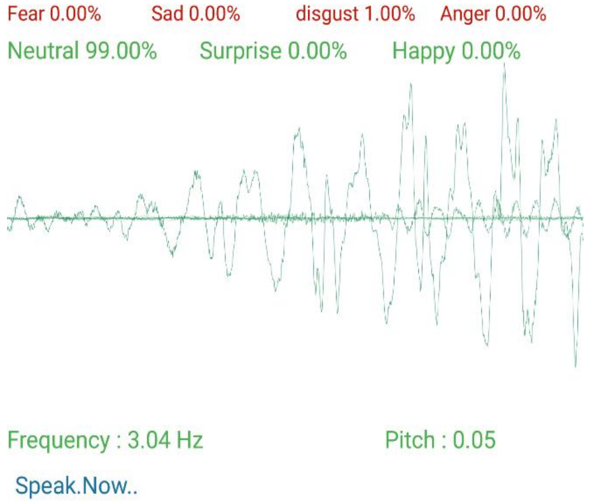
Speak.Now..

93.0 %
Sadness
3.0 %
Neutral
🟥 Disgust 🟩 Anger 🟨 Fear 🟦 Neutral ⬛ Sadness 🟦 Surprise ⬜ Happy

Fig. 6. Sadness Emotion

5) Neutral Emotion

For both actors, the utterances spoken with neutral emotion are clearly articulated speech and show some pausing between words.


Fear 0.00% Sad 0.00% disgust 1.00% Anger 0.00%
Neutral 99.00% Surprise 0.00% Happy 0.00%

Frequency : 3.04 Hz Pitch : 0.05

Speak.Now..

Fig. 7. Neutral Emotion

6) Surprised Emotion

[24]Noted that with surprise “the voice suddenly glides up (or up-down), falls to a mid-level (joyful surprise) or to a lower level (stupefaction). [24]Also noted a very wide pitch range for surprise, with tempo and pitch median normal or higher.

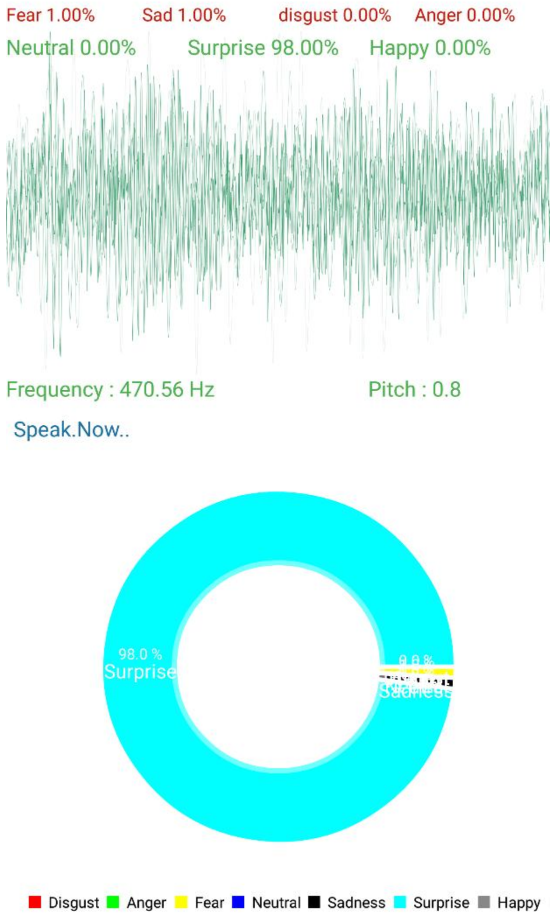


Fig. 8. Surprised Emotion

VII. CONCLUSION AND FUTURE WORK

Speech Emotion Recognition is effective and useful for accentuating the naturalness within the speech that is founded on human/machine interaction. SER systems have expansive applications within everyday life. For instance, the emotional analytics of phone calls between criminals would further aid those investigating the activities of criminals, helping them with the earlier detection of criminal activities. Conversations that may occur with humanoid partners and robotic toys and pets will develop and become more real to life and with that, more enjoyable; more so if they have the capacity to comprehend and express emotions in a human-like manner. There is also a range of useful applications for automated emotional analysis such as speech-to-speech translation whereby speech in any given language can instantly be translated by the machine into another language.

In this paper, we have reviewed the method of classification for the study of Speech Emotion Recognition system;namely Support Vector Machine. There were a number of speech features such as energy, fundamental frequency, and Mel-frequencycepstrum coefficients that were extracted from the sample taken for emotional speech. A relatively similar accuracy of emotion classification was attained for both of these classifiers. It can be therefore determined that the accuracy of the system used is highly dependent the emotional speech database that is utilized. For this reason, it is required that the emotional speech database is recorded correctly.

The method that is proposed has significantly enhanced speech emotion recognition and enable progress to be elevated in Artificial Intelligence (AI) and Human-Computer Interaction (HCI). It is our goal to enhance the method and further test its performance on a multilingual speech emotion database at a point in the future.

REFERENCES

[1]. Chiriacescu I., ‘Automatic Emotion Analysis Based On Speech,’ M.Sc. Thesis, Department of Electrical Engineering, Delft University of Technology, 2009.

[2]. Vogt T., Andre E. and Wagner J., ‘Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization,’ Proceedings of LNCS 4868, 75-91, 2008.

[3]. Ayadi M. E., Kamel M. S., andKarray F., ‘Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases’, Pattern Recognition, 44 (16), 572-587, 2011.

[4]. Zhou y., Sun Y., Zhang J, Yan Y., ‘Speech Emotion Recognition using Both Spectral and Prosodic Features,’ IEEE, 23(5), 545-549, 2009.

[5]. Schuller B., Rigoll G., Lang M., ‘Hidden Markov Model Based Speech Emotion Recognition,’ IEEE ICASSP, 1-3, 2003.

[6]. Shen P., Changjun Z. and Chen X., ‘Automatic Speech Emotion Recognition Using Support Vector Machine,’ Proceedings of International Conference On Electronic And Mechanical Engineering And Information Technology, 621-625, 2011.

[7]. Satyanand Singh, Dr. E.GRajan, “MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors” in International Journal of Computer Applications, may 2011.

[8]. Vimala.C, Dr.V.Radha, 2011, “Speaker-Independent Isolated Speech Recognition System for Tamil Language using HMM,” International Conference on Communication Technology and System Design, Speech Communication 46.

[9]. Emily Mower, Maja J Mataric, and Shrikanth Narayanan, 2011, “A Framework for Automatic Human Emotion Classification Using Emotion Profiles,” IEEE Transactions on Audio, Speech

and Language Processing, Vol. 19, No. 5, pp. 1057 - 1070.

[10]. http://www.fon.hum.uva.nl/praat/, Last accessed on 12.11.2012.

[11]. Simon Haykin, 1999, “Neural networks: A Comprehensive Foundation,” Pearson Education.

[12]. Vibha Tiwari, 2010, “MFCC and its applications in speaker recognition,” International Journal on Emerging Technologies, ISSN: 0975-8364.

[13]. Rabiner, L.R., &Juang, B.H., 1993, “Fundamentals of Speech Recognition,” Englewood Cliffs, Prentice-Hall.

[14]. Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, 2004, “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture,” IEEE, ICASSP, pp. I – 577 - I – 580.

[15]. Hsu C.W, Chang.C, Lin C.J, “A Practical Guide to Support Vector Classification,” Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

[16]. Simon Haykin, 1999, “Neural networks: A Comprehensive Foundation,” Pearson Education.

[17]. Bhoomika Panda, DebanandaPadhi, KshamamayeeDash,and Prof. Sanghamitra Mohanty, 2012, “Use of SVM Classifier & MFCC in Speech Emotion Recognition System,” International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 3.

[18]. Vaishali M. Chavan, V.V. Gohokar, 2012, “Speech Emotion Recognition by using SVM-Classifier,” International Journal of Engineering and Advanced Technology, IJEAT, Vol. 1, Issue 5.

[19]. S. Haq and P. J. B. Jackson, "Machine Audition: Principles, Algorithms,and Systems," Hershey PA, 2010, pp. 398-423.

[20]. Yongjin Wang, Ling Guan, “Recognizing Human Emotional State from Audiovisual Signals,” IEEE TRANSACTIONS ON MULTIMEDIA, VOL.10, NO.5, AUGUST 2008

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

[22]. Ververidis, D., &Koropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48, 1162–1181.

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

[24]. Murray, I. R. & Arnott, J. L. (1993): Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature of Human Vocal Emotion. Journal of Acoustic Society of America 93 (2), 1097-1198.

Humaid Alshamsi "Automated Speech Emotion Recognition App Development on Smart Phones using Cloud Computing ""International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.5, 2018, pp. 71-77