EMOTION DETECTION FROM VOICE BASED CLASSIFIED FRAME-ENERGY SIGNAL USING K-MEANS CLUSTERING

Nazia Hossain[1], Rifat Jahan[2,] and Tanjila Tabasum Tunka[3]

[1]Senior Lecturer, Department of Computer Science & Engineering, Stamford University Bangladesh, Dhaka, Bangladesh
[2&3]Undergraduate Student, Department of Computer Science & Engineering, Stamford University Bangladesh, Dhaka, Bangladesh

ABSTRACT

Emotion detection is a new research era in health informatics and forensic technology. Besides having some challenges, voice based emotion recognition is getting popular, as the situation where the facial image is not available, the voice is the only way to detect the emotional or psychiatric condition of a person. However, the voice signal is so dynamic even in a short-time frame so that, a voice of the same person can differ within a very subtle period of time. Therefore, in this research basically two key criterion have been considered; firstly, this is clear that there is a necessity to partition the training data according to the emotional stage of each individual speaker. Secondly, rather than using the entire voice signal, short time significant frames can be used, which would be enough to identify the emotional condition of the speaker. In this research, Cepstral Coefficient (CC) has been used as voice feature and a fixed valued k-means clustered method has been used for feature classification. The value of k will depend on the number of emotional situations in human physiology is being an evaluation. Consequently, the value of k does not necessarily consider the volume of experimental dataset. In this experiment, three emotional conditions: happy, angry and sad have been detected from eight female and seven male voice signals. This methodology has increased the emotion detection accuracy rate significantly comparing to some recent works and also reduced the CPU time of cluster formation and matching.

KEYWORDS

Cepstral Coefficient (CC), Emotion Detection Accuracy, Mel Frequency Cepstral Coefficient (MFCC), Vector Quantization (VQ), k-means Clustering

1. INTRODUCTION

Automatic recognition of an emotional state from human speech is an important research topic with a wide range. In different applications in real life, e.g. to measure psychiatric condition in the development of mental health system, customer reaction analysis in social call-center etc. nowadays emotion detection opens a wide range of research area. Emotion can be identified based on facial orientation, gesture detection, voice recognition etc. However, in a different scenario where only audio signals are available, emotion detection via voice signal is the way alone. With respect to this, here an attempt has been made to recognize and classify the speech emotion from a voice database. Voice signal processing consists of different stages; signal pre-processing, noise elimination, farming and windowing, feature extraction and modeling. Along with different signal pre-processing techniques, speech feature extraction consisting of fundamental frequency, energy, linear predictive coding (LPC) [11], and Mel Frequency Cepstral Coefficient (MFCC) [12] from the short-time framed signal are comprehensively investigated to

find the speaker. Moreover, PNCC (Power Normalized Cepstral Coefficient) [12], Formant Frequency Calculation [3], pitch, loudness are the features to identify anyone's emotional stage. In the field of emotion detection, the things are slightly changed. When a person becomes emotional, his voice gets modified on the basis of the condition of the emotion. For example, if a person gets angry then some voice features like stress, power, loudness increase. On the other hand, if a man is sad, his voice tone gets lower and therefore the mentioned features get decreased on average. Also, the intensity, jitter, shimmer are different features which are important in emotion detection. Basically, speaker identification and emotion detection features are similar, however, in emotion detection the classification and feature modeling part get priority along with the extracted features.

The process of emotion detection almost follows the stages of speaker identification. In this regard, after extracting the voice features, the feature storing and matching techniques have been followed are VQ-LBG (Vector Quantization using Linde, Buzo, and Gray), Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) etc. Empirical experimentation shows that the combined features of Cepstral Coefficient (CC) on the significant frame, shows the highest accuracy on this database. As some measure of a feature coefficient can represent a different emotional stage of a different person, therefore, this is a necessity to categorize the training data based on each user. Moreover, short-time framing also has come into account as emotional detection does not require to examine the entire voice signal. In this experiment total fifteen (eight female and seven male) users have given their voices in happy, angry and sad mode of emotional states. Naturally, when people are happy their rate of word utterance increases with respect to the time and voices gets louder. On the other hand, when they are sad, the voice gets lower in the perspective of the energy. Lower energy implies higher frequency. And, then people gets angry, the stress level of voice changes noticeably and energy of voice gets higher.

2. LITERATURE REVIEW

In the recent years, excessive investigations have been completed to recognize emotions by using speech statistics. Some researchers have focused on the classification methods [1][2][5][6] and some have focused on the extracted features [3][4].

2.1. CLASSIFICATION BASED METHODS

In the regard of the classification methods; Cao *et. al.* [1] proposed a ranking SVM method for synthesizing information about emotion recognition using binary classification. This ranking method, instruct SVM algorithms for particular emotions, treating data from every utterer as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM achieves two advantages, firstly, for training and testing steps in speaker- independent it obtains speaker specific data. Secondly, it considers the intuition that each speaker may express mixed of emotion to recognize the dominant emotion. Chen *et. al.* [2] aimed to improve emotion recognition in speaker-independent with three level voice based emotion recognition method. This method classifies different emotions from coarse to fine then select an appropriate feature by using Fisher rate. The output of Fisher rate is an input parameter for multi-level SVM based classifier. Furthermore, the principal component analysis (PCA) along with artificial neural network (ANN) is being employed to minimize the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments such as Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence indicates in dimension reduction Fisher is better than PCA and for classification, SVM is more extensible than ANN for emotion recognition in speaker independent. On the other hand, Albornoz *et. al.* [5] investigated a new spectral feature in order to determine emotions and to characterize groups. This study was based

on acoustic features and a novel hierarchical classifier, emotions are grouped. Different classifier such as HMM, GMM and MLP have been evaluated with distinct configuration and input features to design novel hierarchical techniques for classification of emotions. The innovation of the proposed method consisted of two phases, first the election of foremost performing features and second is employing of foremost class-wise classification performance of total features same as the classifier. Also, Lee *et. al.* [6] proposed a hierarchical structure for binary decision tree in emotion recognition fields. This method concentrated on the simpler classification obstacle at the top level of the tree to diminish agglomeration of error. This structural method also maps input speech data into one the emotion classes via following layer of binary classification.

2.2. FEATURE EXTRACTION BASED METHODS

On the other hand, Wu *et. al.* [3] proposed a new modulation spectral features (MSFs) human speech emotion recognition. Appropriate feature extracted from an auditory-inspired long-term spectral-temporal features are extracted by utilizing a modulation filterbank and an auditory filterbank for speech decomposition. This method obtained acoustic frequency and temporal modulation frequency components for carrying out the important data which is missing from traditional short-term spectral features. For classification process, SVM with radial basis function (RBF) is adopted. Yang *et. al.* [4] presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory. Firstly, they began from the predicted pitch of a speech signals, then computing the spherical autocorrelation of pitch histogram. It calculates the incidence of dissimilar two-pitch duration, which has caused a harmonic or in-harmonic impression. In the classification step, the Bayesian classifier plays an important rule with a Gaussian class-conditional likelihood. The experimental result is evaluated on Berlin emotion database by using harmony features indicate an improvement in recognition performance. The recognition rate of this experiment has been improved by 2% on average than the previous work.

Therefore, from the previous studies, this is clear that some papers have focused on the classifier methods. Depending on the dataset size, the classifiers are being changed. Though, this is sometimes not feasible in real life scenario. Again, some methodologies are categorized into male and female users. So, this type of method has a prerequisite that, the system has to differentiate the gender first and then goes to the emotion detection part. Moreover, both genres of researches have used a huge dimension of feature coefficient to store and match. Hence, to address these limitations of the emotion detection research methodologies, a novel scheme of categorizing user with short signification framing process has been discussed in this paper.

3. PROPOSED METHODOLOGY

In the proposed methodology two key points have been considered. First of all, the previous researches have not considered about the individuals, rather they have taken the gender (male or female person) into consideration. In this research, each person has been pursued significant as the emotional stage is not a regular and usual state. Therefore, the regular findings of a voice signal cannot be matched or compared with the excited ones. Hence, the emotional stages of each individual have been gathered into one group in the entire dataset. And secondly, significant voice signal frames have been taken into account rather than the entire voice signal. Because the whole input voice signal is not equally important in emotion detection like speaker identification or speech recognition systems. The proposed method has been named as ‘F-1 CC method based on k-means clustering’ method when the value of k is three (3) in this experiment for three type of emotion detection. Based on the emotion detection variety, the value of k can be increased. The entire emotion detection system model is shown in Figure 3 in both training and testing period

where, the training and testing phases are represented by the left part and right part respectively. The details of each step are described below:

3.1. DATA COLLECTION

In Figure 3, first phase of the system model of both training and testing environment is, input signal acquisition from the user. In this research work, each user dataset seems like a group. Different input voices of the same individual have been categorized in one group of that person. In this system, three emotions have been detected from total 15 users, where there is 8 female and 7 male voices. Figure 1 shows an example of the user-data input group scenario.

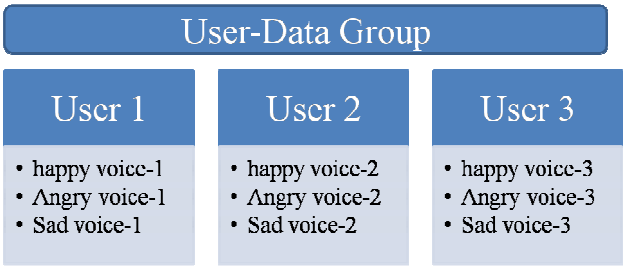

Figure 1. User-Data Input Group

3.2. FRAMING AND FEATURE EXTRACTION

In Figure 3, the second and third phases are Framing and Cepstral Coefficient (CC) calculation respectively. In this stage, the entire input voice signal has been framed into a different frame. The unvoiced and voiced signal portions have been differentiated using Vocoder. Among the voiced frames, we have taken the first signal frame. Each frame consists of 250ms of the overall voice signal. From the voice frame, cepstral coefficient (CC) has been calculated. In this experiment, thirty (30) cepstral coefficients have retrieved in each frame. The Cepstrum is defined as the Inverse Discrete Fourier Transform (IDFT) of the log magnitude of the discrete Fourier transform (DFT) of the signal shown in the equation 1 [8]:

$$c[n] = f^{-1}\{\log \mid f\{x[n]\mid\}\} \qquad (1)$$

3.3. FEATURE CLUSTERING

The fourth phase of Figure 3 is Formation of Cluster. In this experiment, vector quantization (Linde, Buzo, Gray algorithm) has been used for k-means clustering. So, the extracted voice features have been stored in the vector codebook by this algorithm in the clustered way. Here, for the value of k, three (3) has been used, as three type of emotions have been recognized in this research; sadness, happiness, and anger. During the training period, the extracted voice features will be assigned any of the three clusters automatically by the clustering algorithm. Therefore, regardless of the size of user input, the cluster number will be constant always. Figure 2 symbolizes the clusters of the vector codebook.

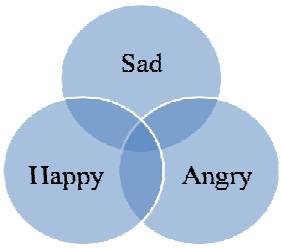

Figure 2. Three clusters symbol in a vector codeboo

3.4. FEATURE MATCHING

From the different training clusters, a test cluster is being matched in this phase of Figure 3. As a feature matching technique, the mean value of the respective Euclidian distance has been used here to compare the distances among the cepstral coefficients during the testing period, from the trained clusters in the codebook.

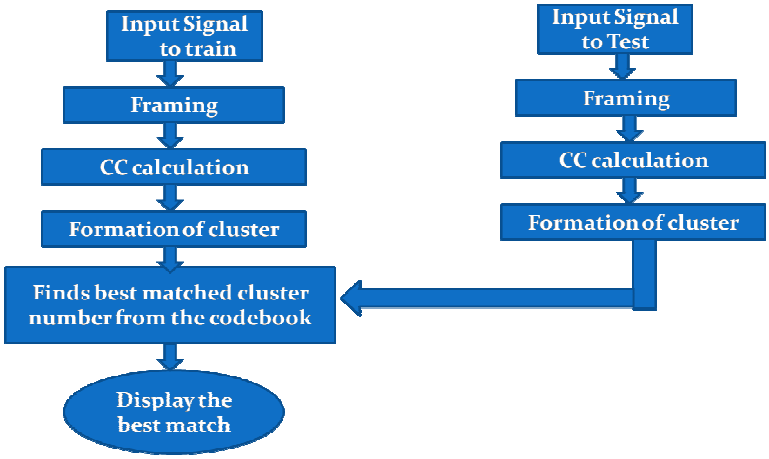

Figure 3. System Model (both Training and Testing)

4. EXPERIMENTAL SETUP AND RESULT ANALYSIS

The experiment has been installed in MATLAB R2017a version. To arrange the dataset, the students of Stamford University Bangladesh have given the input voice in different emotional aspects. Here, a total of 15 students (aged around 20-25 years) have recorded their voice in three different emotional states (happy, sad and angry); that is 15*3 total 45 recorded audio files. Among fifteen (15) users, there is eight (8) female and seven (7) male voices. Input voice signal duration covers from 7 to 15 seconds. The experimental result is evaluated on the basis of detection accuracy rate and CPU time comparison. Equation 2 [7] shows the formula to calculate the Emotion Detection Accuracy (EDA) rate. Moreover, to apply the Vocoder, an open source software named "Audacity" has been used.

$$EDA = \frac{TotalNumberofTrueDetection}{TotalNumberofDetectionAttempt} \times 100\% \qquad (2)$$

4.1. PERFORMANCE EVALUATION OF F-1 CC METHOD

Performance evaluation has been performed based on another two methods, there are MFCC [4], Jitter [5], CC[3] based method with the proposed F-1 CC method based on k-means Clustering. Figure 4 shows the Emotion Detection Accuracy (EDA) rate comparison among different methodologies with the growing number of user input. From the figure 4, this is clear that though MFCC and jitter are quite stable with the growing number of the user, although, the proposed method shows a significant improvement for the given dataset. Again, in the following, Figure 5 shows that this proposed technique also performs well, in the aspect of CPU time; as the number of clusters is not being increased with the growing number of input.

Therefore, this is clear that the proposed F-1 CC method performs significantly better (around 9%) than the relatively similar methodologies proposed recently. Also, this algorithm is less complex as in k-means clustering technique, the value of k is fixed. The value of k can be changed on the basis of the detection criteria of emotional stages. Regarding this, the run-time complexity has to be reduced compared to the other methodologies. This is shown that the runtime is not increasing rapidly after enriching the number of users. This is a clear denotation of the reliability and flexibility of this method.

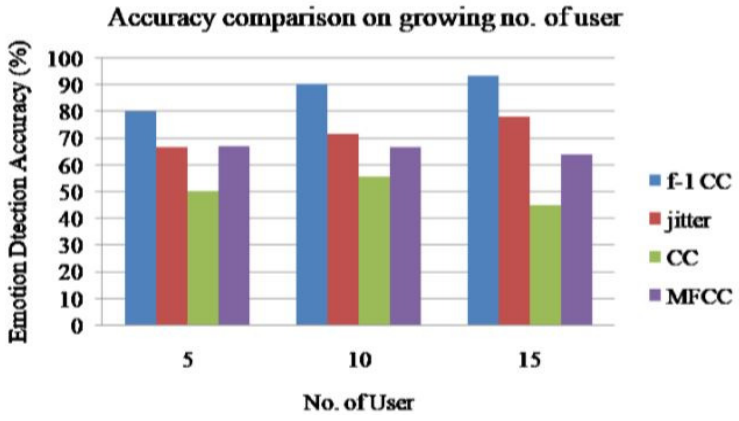

Figure 4. Emotion Detection Accuracy (EDA) rate comparison among different methodologies with the growing number of user

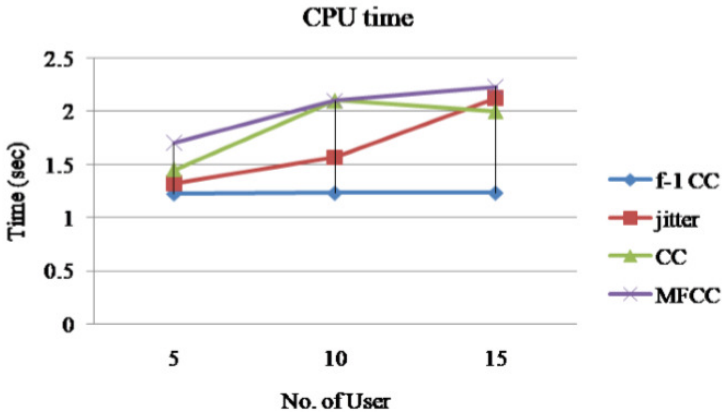

Figure 5. CPU time consumption comparison among the different methodologies with the growing number of user

5. CONCLUSION AND FUTURE WORKS

Emotion detection system is new and promising in the machine learning and artificial intelligence area of computer science. Although, this technique is not fully optimized enough, however, this f-1 CC method with k-means clustering has shown an indicator of success in its experimental scenario. Though the experimental dataset is not large enough and the noisy environment is not considered here. The setup has assumed that any state-of-art noise elimination technique can be applied in the noisy data and after that, the rest of the experiment would outperform as well.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, especially those who have helped this research by giving their voice as input in different emotional states. Therefore, the authors are grateful to all the faculty members and students of Department of Computer Science and Engineering in Stamford University Bangladesh.

REFERENCES

[1] H. Cao, R. Verma, and A. Nenkova, “Speaker-Sensitive Emotion Recognition via Ranking: Studies on Acted and Spontaneous Speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[2] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech Emotion Recognition: Features and Classification Models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[3] S. Wu, T. H. Falk, and W.Y. Chan, “Automatic Speech Emotion Recognition Using Modulation Spectral Features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.

[4] B. Yang and M. Lugger, “Emotion Recognition from Speech Signals using New Harmony Features,” Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010.

[5] E. M. Albornoz, D. H. Milone and H. L. Rufiner, “Spoken Emotion Recognition using Hierarchical Classifiers,” Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.

[6] C.C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion Recognition Using a Hierarchical Binary Decision Tree Approach,” Speech Commun., vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.

[7] Singal-to-Noise Ratio, Available on: https://en.wikipedia.org/wiki/Signal-to-noise_ ratio [Accessed on: 08.04.2018]

[8]F. Orsag, M. Darhansky, "Biometric Security Systems: Fingerprint and Speech Technology," Security Systems: Fingerprint and Speech Technology, 2003.

[9] I.Bisio, A. Delfino, F. Lavagetto, M. Marchese, A. Sciarrone, “Gender-driven emotion recognition through speech signals for ambient intelligence applications”, IEEE Transaction on Emerging Topics in Computing, vol. 1, no. 2, pp. 244-257, Dec. 2013.

[10] E. Kim, K. Hyun, S. Kim, Y. Kwak, “Improved Emotion Recognition with a Novel Speaker-Independent Feature”, IEEE/ASME Transaction on Mechatronics, vol. 14, no. 3, Jun. 2009.

[11] “Autoregressive Modeling: the linear predictive coding, difference between power spectrum and linear prediction”, [Online]. Available: http://www.clsp.jhu.edu/~s sriram/ research/ autoreg/ ar_ model.html [Accessed: 16.02.2018]

[12] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, J. A. Chamber, “Study on Fusion Strategies and Exploiting the Combination of MFCC and PNCC Features for Robust Biometric Speaker Identification”, 4th International Workshop on Biometric and Forensics (IWBF), Mar. 2016.

[13] S. Thomas, S. Ganapathy, H. Hermansky, “Spectro-Temporal Features for Automatic Speech Recognition Using Linear Prediction in Spectral Domain”, 16th European Signal Processing Conference, 2008.

[14] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition", in Proc. ISCA Speaker Odyssey, 2012.

AUTHORS

Nazia Hossain is a Senior Lecturer in Stamford University Bangladesh. She has complete her M.Sc. Engg Degree in Computer Science & Engineering from Bangladesh University of Engineering and Technology (BUET) and B.Sc. Engg. in Computer Science & Engineering from Khulna University of Engineering and Technology (KUET). Currently, her research area is speech signal processing. She is also a member of Bangladesh Engineering Institute (IEB).

Rifat Jahan is a B.Sc. Engg. In Computer Science and Engineering Student in Stamford University Bangladesh. She has completed her Higher Secondary Certification in 2012 from Bangladesh Ideal School & College. Now, she is a 4th year student of the four year Bachelor degree and doing her thesis on Speech Signal.

Tanjila Tabasum Tunka is a B.Sc. Engg. In Computer Science and Engineering Student in Stamford University Bangladesh. She has completed her Higher Secondary Certification in 2012 from Barguna Govt. College. She is a 4th year student of the four year Bachelor degree and doing her thesis on Speech Signal.