

MFCC-based Vocal Emotion Recognition Using ANN

Mandar Gilke , Pramod Kachare , Rohit Kothalikar , Varun Pius Rodrigues
and Madhavi Pednekar

Department of Electronics and Telecommunication,
Don Bosco Institute of Technology, University of Mumbai

Abstract. Emotion recognition and verification is the automated determination of the psychological state of the speaker. This paper discusses the method to extract features from a recorded speech sample, and using those features, to detect the emotion of the subject. Mel-Frequency Cepstrum Coefficient (MFCC) method was used to extract these features. Every emotion comprises different vocal parameters exhibiting diverse characteristics of speech. These features result in different MFCC coefficients that are input to the trained Artificial Neural Network (ANN), which will analyze them with the stored database and compare the same to recognize the emotion.

Keywords: Back-propagation neural network, valence, hidden-layer, confusion-matrix, mel-frequency cepstral co-efficient, graphical user interface.

1. Introduction

Human Machine Interface (HMI) recognition systems incorporate the principles of corporal interaction that deduce perfunctory characteristic extraction methods. The speech characteristics include pitch, formant, prosody and timbre. The emotion verification task designed for such recognition systems uses a-priori information to determine whether the outcome of a speech sample is efficiently construed in a manner in which the sentence is spoken. In practice, a-priori information would normally be available in a real system, instinctively captured when candidate users are registered with that system. Within such constraints, there are two further main branches to this research area; one in which the material being spoken is fixed and the other in which the material being spoken is unrestricted. In the unrestricted case the problem is more difficult, and accuracy may be more closely related to the amount of captured data that can be analyzed than upon the accuracy of the system employed[1].

Automatic Emotion Recognition is a recent research topic which is primarily formulated for the Human Computer Interaction (HCI) field. As computers have become an integral part of our lives, the need has risen for more natural communication interface between human beings. To make HCI more natural, it would be favorable if modeled systems have the ability to recognize emotional situations the same way as humans do. Many researchers have explored several classification methods for recognition systems such as the Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes Classifier (MLC), K-nearest Neighbors (KNN) and Artificial Neural Networks (ANN) [2],[3].

2. Methodology of Emotion Recognition

Emotional speech recognition systems perform two fundamental operations namely, signal modeling and pattern matching. Signal modeling represents the process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter sets from memory which closely matches the parameter set obtained from the input speech signal. After performing these two operations, one can perpetually categorize the emotions with respect to the arousal rate of the dialogue spoken with its valence i.e. the behavioral aspect of the person displaying the emotion.

¹ Mandar Gilke Tel.: +919920260999
E-mail address: gilke.mandar@gmail.com

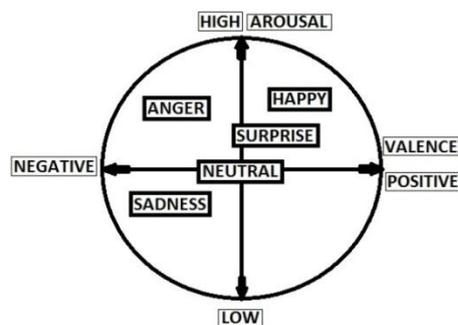


Fig. 1: Arousal-valence Matrix of Emotion recognition

2.1. Mel-Frequency Cepstrum

All Psychophysical studies have shown that human perception of the frequency contents of sound for speech signals does not follow a linear scale. The actual frequency (f) is measured in Hertz (Hz) and a subjective pitch is measured on a scale called the mel scale. The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point the pitch of a 1 kHz tone, 40 db above the perceptual hearing threshold is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz [4].

$$\text{mel}(F) = 2595 \times \log_{10}(1 + f/700)$$

In the next step, we convert the log mel-spectrum back to time. The result is called the mel frequency-cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis because the mel spectrum coefficients are real number. We can convert them to the time domain using the Discrete Cosine Transform (DCT) [5].

2.2. Back-Propagation Neural Network

Artificial Neural Network (ANN) is an efficient pattern recognition mechanism which simulates the neural information processing of human brain. The ANN processes information in parallel with a large number of processing elements called neurons and uses large interconnected networks of simple and non-linear units [6].

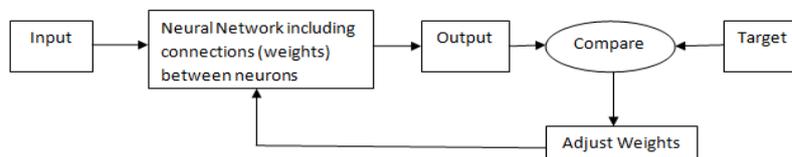


Fig. 2: Neural Network [7]

The quantitative modeling and processing of data using neural networks is effectively performed using the Supervised Learning Neural Network Back-Propagation Algorithm. For a given set of training input-output pair, this algorithm provides a procedure for changing the weights in a back-propagation network (BPN) to classify the input patterns correctly. The aim of this neural network is to train the network to achieve a balance between the network's ability to respond (memorization) and its ability to give reasonable responses to the input that is similar but not identical to the one that is used in training (generalization) [7]. A BPNN is a multi-layer, feed-forward neural network consisting of an input layer, a hidden layer and an output layer. The hidden layers are used to classify successfully the patterns into different classes. The inputs are fully connected to the first hidden layer, each hidden layer is fully connected to the next, and the last hidden layer is fully connected to the outputs.

2.3. Front-End Display using Graphical User Interface

The ultimate stage of a recognition model is its capacity to display the computed results in an efficient and intelligent mode. The graphical user interface (GUI) aids in this aspect to describe the response of the system using figures, text-box and plots. One such example of this technique is shown below. The two-dimensional acoustic plot is generated by the system once the user pushes the 'display plot' button. Consequently, a number of features can be obtained in the built-in GUI Toolbox of MATLAB that provides a suitable technique for human computer interaction.

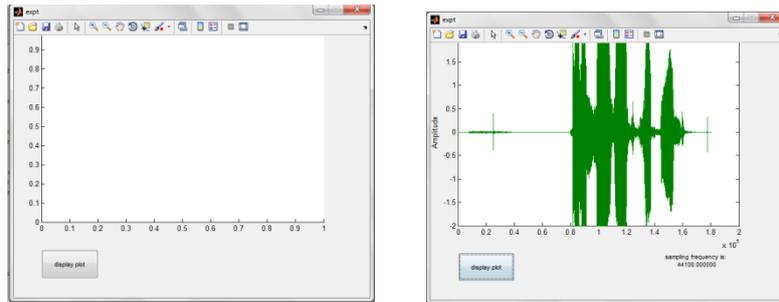


Fig. 3: Illustration of the Graphical User Interface

3. System Implementation

The following steps of data preparation and computation, for training the system, provide a succinct approach to the recognition model.

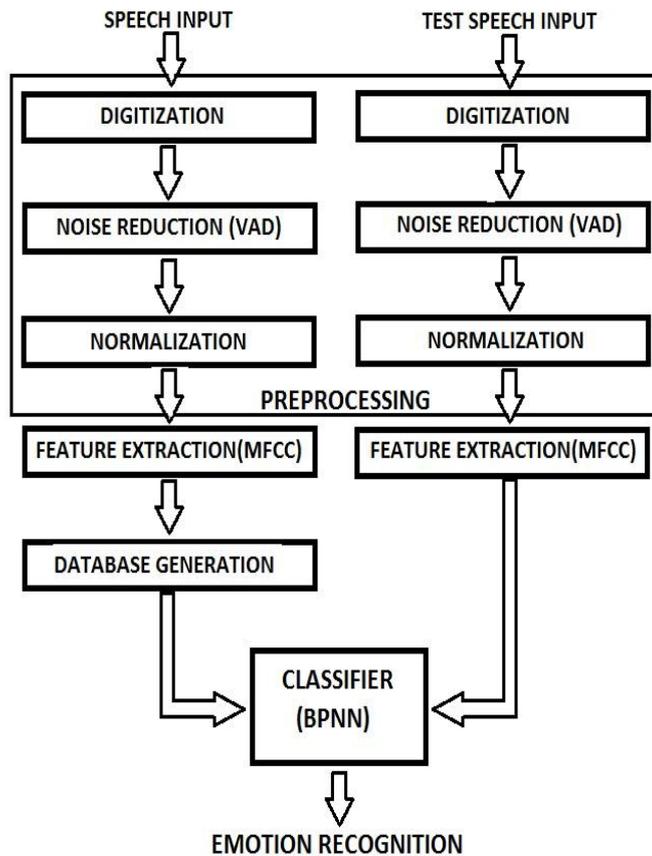


Fig. 4: Vocal Emotion Recognition Flowchart

3.1 . Data Acquisition

A headphone-mic, a computer and GOLDWAVE software were used for single channel recording of emotionally biased utterances of fixed lengths in each emotion from 3 male and 3 female speakers. Each speaker was asked to utter 5 times a fixed set of 3 short sentences, each of a different emotion. The necessary emotional acting was obtained by narrating to the speaker a short emotional story so as to sufficiently arouse the same emotion in the dialogues. Utterances corresponding to sad, surprise, happy, anger and neutral utterances are recorded. MATLAB 7 Software was used for all computations.

3.2 . Preprocessing

After data acquisition, silence periods within the samples were removed. The usage of Voice Activity Detection (VAD) technique was used to delete such silence frames. Then the speech samples were passed through the Low Pass Filter $(1 - 0.97z^{-1})$ which gives a spectral tilt to the speech samples.

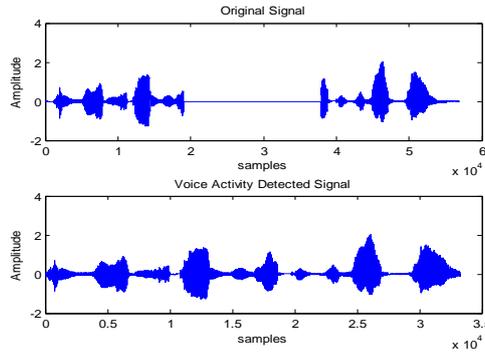


Fig. 5: Voice Activity Detection

3.3 . Windowing

The filtered voice samples were segmented into 23.22 ms frames with each frame having 50% overlap with the adjacent frames. Each frame is then multiplied by a Hamming window of the same length i.e. 246 samples per frame length [8].

3.4 . Feature Extraction

20 MFCC and 1 total Log-Energy features were calculated from each frame using 24 triangular Mel-frequency filter banks. Then each feature vector was normalized by mean and standard deviation, which were computed from the feature vectors extracted from the trained utterances [8].

3.5 . Pattern Matching

The trained Artificial Neural Network (ANN) was required to be tested with features extracted from the test-utterances. The ANN was trained with multiple voice samples taken at different instances uttering the same phrase at all times. The total log-likelihood of these test-vectors of one test utterance with respect to the trained matrix corresponding to each emotion-class was computed. The test utterance is considered to belong to that emotion-class with respect to which the total log-likelihood becomes the largest [8]. An ambiguity may arise when surprise may be expressed along with any other emotion such as anger-surprise, happy-surprise, etc. Also some of the emotions like surprise-anger, surprise-happy, anger-happy and sad-neutral appear to have similar acoustic characteristics. So a Confusion-Matrix was prepared which would take care of these uncertainties up to a certain level.

3.6 . Display of Output

We have to take the input from user. The system processes and extracts its features using the MFCC coefficients. This vector is given alongside as input to the trained matrix and the evaluated output is compared with the available feature model. The GUI for this system is then prepared using MATLAB.

4. Results and Discussions

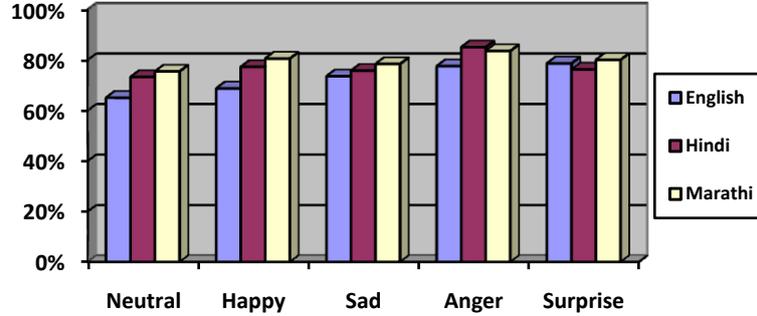
The use of three different language databases for emotion recognition has resulted in the following observations. The recognition accuracies were obtained using the same feature extraction method i.e. MFCC and BPNN classification techniques. Correspondingly, the confusion-matrix shows that the confusion between anger and surprise is high in comparison with any other pair of emotion. This is due to the approximation of near formant and pitch acoustic features of these two emotions. The back-propagation algorithm proves to be an efficient method for emotion recognition with reference to the graphical result.

TABLE 1: Confusion-Matrix obtained as an average of the three language databases

Emotional Class	Neutral	Happy	Sad	Anger	Surprise
Neutral	79%	10.2%	15.4%	0%	4%
Happy	13.2%	75.2%	12.7%	11.76%	35.2%
Sad	13.59%	14.5%	71.35%	0%	0%
Anger	0%	0%	0%	87.43%	54.3%
Surprise	3.1%	36.8%	5.5%	49.28%	85.9%

TABLE 2: The Recognition Accuracies for the three languages corresponding to the five emotions

Emotional Class	Neutral	Happy	Sad	Anger	Surprise
English	65%	68.7%	73.5%	77.5%	78.6%
Hindi	73.2%	77.2%	75.7%	85 %	76.2%
Marathi	75.4%	80.5%	78.35%	83.5%	80%



GRAPH 1: The recognition accuracies obtained for the three language databases

5. Conclusion

The parameterized recognition framework built on MATLAB interprets the emotion of the speaker using Back-Propagation Algorithm from the emotional database. The recognition system strictly adheres to the computed results of the database, independent of age, gender and linguistic background. This system can apply its features to sophisticated recognition mechanisms.

6. Acknowledgements

The authors gratefully acknowledge the co-operation of the faculty of Don Bosco Institute of Technology who offered immeasurable guidance throughout the implementation of the proposed system. The valuable inputs by Prof. J.H. Nirmal from K.J. Somaiya College of Engineering, affiliated to the University of Mumbai, have enabled the successful implementation of the system and we thereby express our sincere gratitude. We also appreciate the help of all those who volunteered to make available the emotion samples for the successful execution of the system design.

7. References

- [1] Ian McLoughlin, "Applied Speech and Audio Processing with MATLAB Examples", Cambridge University Press, 2009.
- [2] Beth Logan, "Mel Frequency Cepstral Co-efficients for Music Modeling", Proceedings of the International Symposium on Music Information Retrieval (ISIMIR), Plymouth, MA 2000.
- [3] Yashpalsing D. Chavhan and M.L. Dhore, "Speech Emotion Recognition using SVM" IEEE International Conference on 'Computer Vision and Information Technology, Advances and Applications', ACVIT-09, December 2009, pp. 799-804.
- [4] L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Pearson Education, 2009.
- [5] Madhavi S. Pednekar, Kavita Tiwari and Sachin Bhagwat, "Continuous Speech Recognition for Marathi Language Using Statistical Method", IEEE International Conference on 'Computer Vision and Information Technology, Advances and Applications', ACVIT-09, December 2009, pp. 810-816.
- [6] Firoz Shah. A, Raji Sukumar. A, and Babu Anto. P, "Discreet Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, June 2010, pp.319-322.
- [7] S.N. Sivanandam, S.N. Deepa, "Principles of Soft Computing", WILEY India, 2009.
- [8] A.B. Kandali, A.B. Routray, Basu T.K., "Emotion Recognition From Assamese Speeches Using MFCC And GMM Classifier", IEEE Region 10 Conference TENCON 2008, Hyderabad, India, pp. 1-5.