International Journal of Electrical Electronics & Computer Science Engineering
Volume 5, Issue 1 (February, 2018) | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
Available Online at www.ijeecse.com

Modeling and Recognizing Emotions from Audio Signals: A Review

[1]Ritu Tanwar, [2]Deepti Chaudhary
[1] UG Scholar, [2]Assistant Professor, UIET, Kurukshetra University, Kurukshetra, Haryana, India
ritu.tanwar2012@gmail.com, deeptic2015@kuk.ac.in

Abstract: Emotions are very important in human beings life and help to communicate whatever human feels in a certain situation. Emotion recognition (ER) of human beings is one of the significant domains to increase the interaction between human beings and machines. ER can be done by audio signals, facial expressions and EEG (Electroencephalogram) signals. ER from audio signals is one of the active topics in research field as audio signals like music, environmental sounds and speech signals trigger emotions in human beings. The review includes recognition of basic emotions like happiness, sadness, anger, fear, disgust and surprise. In this paper, a system is described in which different features of audio signals like energy features, spectral features, temporal features, rhythm features and harmony features that contain information related to emotional state are identified. The various machine learning models or classifiers like Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Neural Networks are used to recognize the emotion related to the input audio signal. The main objective is to study various features of audio signal and the machine learning models used to classify the emotions and to analyze the models by evaluating various parameters like recognition rate, accuracy of the system and F-measure.

Keywords: Emotion Recognition, Audio Signals, Models and Performance Metrics.

I. INTRODUCTION

Emotions have a very significant role in human beings life and their recognition is important for many applications like security systems, medical applications, e-learning systems, creating smart environment and entertainment [1], [2]. Therefore, ER exhibits an intense research domain academically as well as industrially. The systems developed for ER can be based on facial expressions, audio signals and EEG signals [2] [3].

Human emotions are modeled on the category-based, dimensional-based and appraisal-based methods. Categorical and dimensional based methods are the most common approaches used to model the emotions. Categorical based methods studies primary emotions which are happiness, sadness, anger, fear, surprise and disgust. Dimensional based methods are used to define emotions on dimensional plane which can be two-dimensional and three-dimensional plane. Mostly two-dimensional emotion plane (Arousal-Valence) is considered to model an emotion. Dimensional based methods describes continuous emotion plane whereas categorical describes discrete one [4]. The classification of emotions on a 2-D plane proposed by Thayer is shown in the Fig. 1 given below:

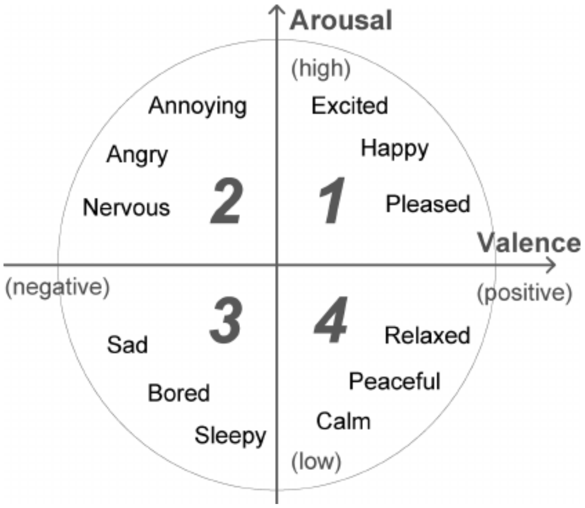

Fig. 1. Arousal-Valence Emotion Space [5]

This paper provides a review of an ER system from audio signals. ER from audio signals can be performed by extracting various features of audio signals and then categorize them to train the classifier. The audio signal features like MFCC, pitch, formants, energy and rhythmic features provide significant help in recognizing the emotional content present in the audio signal [6]. Machine learning techniques are implemented to classify the emotions where audio features are given as input and the machine learning tool predicts the emotion present in audio signal. Various techniques used to classify the machines are SVM, GMM and k-NN [7].

In this paper, an overview of emotion recognition from audio signals is presented in which its two principal factors are discussed: (1) feature extraction and (2) machine learning models. We discussed various types of features associated with audio signals and different machine learning models used to model an emotion recognition system. The paper has three sections: (1) Emotion recognition system (2) feature extraction and (3) machine learning models.

II. RELATED WORK

Since the last decade, a lot of research has been done for recognizing emotions through audio signals. Murty and Yegnanarayana [8] proposed an emotion recognition system by integrating the commendatory characteristic of residual phase with Mel-frequency cepstrum coefficients (MFCC) features of audio signal. This approach provided an Equal error rate (EER) of 10.5% which is much less than EER obtained by implementing these features individually [8]. Tao Li and M. Ogihara [9] presented the benefits of extracting Daubechies Wavelet Coefficient Histograms (DWCH) feature so as to retrieve the information of given music signal. By integrating DWCH features with Fast fourier transform (FFT) and MFCC, an accuracy upto 80% is obtained [9].Yang et al. [5] estimated the AV values and their accuracy determines the capability of music emotion recognition. AV values are computed by taking Music emotion recognition (MER) as a regression problem [5]. Jothilakshmi et al. [10] proposed an approach for recognition of emotions by integrating residual phase and MFCC features. This approach is implemented by using SVM and provided an accuracy upto 85.97% [10]. Muller et al. [11] discussed different models to analyze the music signals which are capable of addressing different music signal features like harmony and melody. The impacts of music signal features on different techniques are examined [11]. Yu at al. [12] proposed an optimized SVM method for recognition of emotions from audio signals by properly choosing the kernel function of SVM. SVM is optimized to obtain the better accuracy results and improving the efficiency of model [12]. S .Koolagudi and K .Rao [13] presented the various kinds of features associated with the speech signals and different machine learning tools used to recognize the emotions [13]. Bisio et al. [2] proposed an emotion recognition system to recognize the emotional state so as to enhance the interaction between human and machines. A system which is composed of two sections named gender and emotion recognition is proposed to increase the recognition rate of the system [2]. Lanjewar et al. [14] proposed an emotion recognition system by fusing MFCC, wavelet and pitch features of the audio signals. GMM and KNN models are implemented to recognize the emotions and recognition rates are determined for different emotions [14]. N .Nalini and S . Palanivel [15] presented the emotion recognition from music signals by fusing MFCC and residual phase features and this combination provided an accuracy rate of 99% by using SVM model [15]. H .Palo and M . Mohanty [16] combined MFCC, LPCC and wavelet coefficients features of audio signals and tested on Radial basis function neural network (RBFNN) machine learning model to recognize the emotions namely angry, fear, happy, disgust and neutral [16].

The paper is described in a proper way where firstly the emotion recognition system, secondly feature extraction and finally the emotions are recognized by using machine learning models.

III. Emotion Recognition System

An emotion recognition system from audio signals is composed of three main parts which are (1) data collection and pre-processing (2) feature extraction and (3) emotion classification [2]. The basic block diagram for emotion recognition system is shown in the Fig. 2 given below:

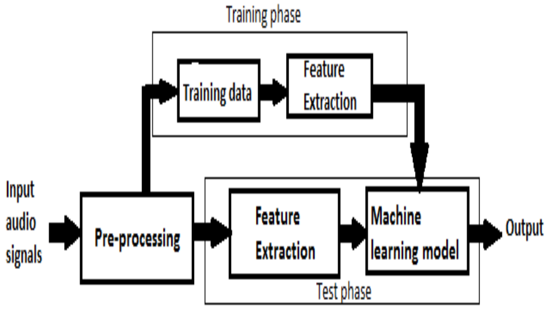

Fig. 2. Block Diagram of Emotion Recognition System

A. Data Collection: The data set consisting of audio signals can be collected manually and are also freely available online for research purpose like Beihang University Database of Emotional Speech (BHUDES), Berlin Emotional Speech database (BES), FAU Aibo Emotion Corpus (FAU AEC), and Speech Under Simulated and Actual Stress (SUSAS) database, etc. [14], [17], [18], [20].

B. Pre-Processing: The signal is pre-processed by passing it through the pre-emphasis filter so that the signal which is transmitted through air can be standardized. Each and every frame of audio signal spectrum is expanded by using a window function so as to minimize the irregularities in audio signal spectrum [19].

C. Feature Extraction: The different types of features are extracted from the audio signals for representation of various perceiving characteristics like rhythm and melody of listening audio signals [20]. The various methods which include algorithms like Spectral contrast and DWCH algorithms and software programs like PsySound, MIRToolbox and Marsyas are implemented on the pre-processed signal to extract the audio signal features like pitch, MFCC, loudness, and spectral characteristics, etc. [5].

D. Emotion Classification: This involves prediction of emotions from audio signal by implementing machine learning techniques like SVM, HMM, k-NN and GMM etc. Therefore, a model is prepared from test data to estimate the human emotional state [2].

IV. Audio Features Extraction

Audio features extraction is a process in which large amount of audio signals are represented in a very compact form which contains important information of the audio signals [21]. The features associated with audio signals are described in the Fig. 3 shown below:

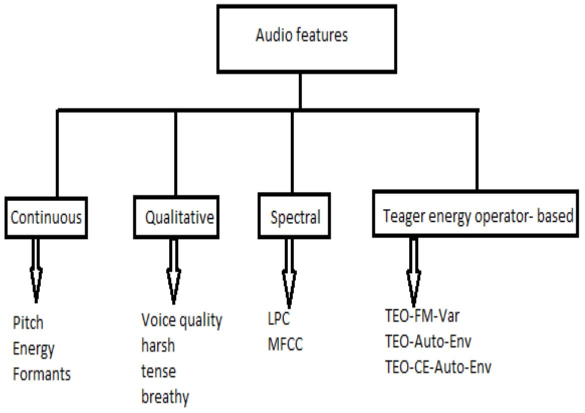

Fig. 3. Audio Features Classification [4]

A. Continuous: Continuous features are the features that are significant in conveying emotions associated to the audio signal and are therefore used to recognize emotions from audio signals [6]. Continuous features include pitch, energy and formants. Pitch is a pervasive characteristic of audio signals [11]. Pitch can be described as momentum of oscillations produced by vocal tract when input audio signal is speech signal [2]. Energy feature of audio signal is related to the intensity of the signal [6]. The voice of the human beings has resonant frequencies. The calculation of frequencies and -3db bandwidth is the elemental evaluation for analyzing the vocal tract of human beings [2].

B. Qualitative: The qualitative features are related to the quality of voice which reflects the emotional content of the audio signal. These features are classified into four categories and they are level of voice; pitch of the voice; phrase, word and boundaries of features; and temporal structures [4].

C. Spectral: Spectral features are estimated from spectrum of the audio signal and are used to describe the scattering of energy over various frequencies [21]. Spectral features include Linear Prediction Coefficients (LPC) and MFCC. Linear predictive analysis is a procedure used to compute the formants. LPCs are estimated in the analysis to track the formants [22]. To extract MFCC features, audio signal is first passed through pre-emphasis filter so as to make the spectrum of signal smooth. The signal after pre-emphasize, is divided into N frames and each frame is having M distance in-between. The frames are windowed to reduce the distortion in spectrum of signal. After windowing, N samples are performed with FFT algorithm to implement DFT. MFCC coefficients are calculated by taking logarithm of output power obtained from filter banks. DCT is used to transform the logarithmic output into time domain [12].

D. Teager Energy Operator (TEO) Based Features: TEO-based features are the ones that occur when additional excitation waveforms are produced from the human vocal folds along with the pitch signals and this condition arises in situations like anger in human speech [22].

V. MACHINE LEARNING MODELS

Machine learning models are used to classify the emotions associated with the audio signal. The different models which are used to recognize the emotions are SVM, GMM, K-NN and neural networks. The purpose of this section is to provide a general description about each learning model used in an emotion recognition system [23].

The machine learning models which are used to recognize emotions from audio signals can be classified in two categories: linear and non-linear classifiers.Linear classifiers are the ones which classifies the emotions on the basis of linear weighted combination of feature values of the audio signals which are given as input in vector form whereas Non-linear classifiers are the ones which are developed on the basis of non-linear weighted combination of feature values of the audio signals. The classifier works as linear or nonlinear by properly selecting the kernel function while implementing an emotion recognition system. Generally, non-linear classifiers are used because in most of the cases the characteristics of input data are unknown [13].

A. SVM: This is a classifier which is based upon transforming an input feature vector space into higher dimension space by performing a non-linear transforming function (kernel function) denoted by Ø. This kernel function reconstructs the feature space like reconstructing 2-D into 3-D feature space and converts a non-linear problem into a linear one [15] as shown in the Fig. 4 given below:

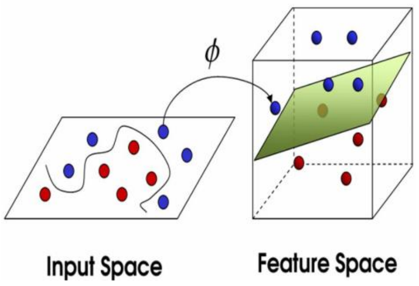

Fig. 4. Basic Concept of SVM [15]

B. GMM: This is a method where emotions from audio signals are described as a mixture of Gaussian densities. The significance of using this model is that the Gaussian parts describe the configuration of spectrum waveforms which provides emotional content and Gaussian mixtures have the capability to demonstrate unpredictable densities. Different emotions produces different probability density functions for irregular features. Therefore, an arrangement of GMM can be used to evaluate the probability of perceiving expression from a specific emotion. A linear weighted GMM is composed of M gaussian densities as expressed in the following equation:

$$\mathrm{P}(\overrightarrow{x|\lambda}) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

Where $\vec{x}$ is an arbitrary vector, $b_i(\vec{x})$ denotes densities of components and p_i denotes the mixture weights and i ranges from 1 to M [14].

C. K-NN: This is a non-parametric machine learning tool and is implemented on various problems related to the analysis of music signals. This model searches k nearest neighbors in training data and makes use of types of k neighbors for determination of the type of given signal [9]. The main features of this machine learning model are that they require high storage space and are sensitive regarding the similarity function used in comparing instances. The limitation of this model is that they consume more time in computations for classifying emotions [24].

VI. Performance Metrics Of Classifiers

The performance of the classifiers can be evaluated by using following parameters:

A. Accuracy R(AR): This metric is defined as the ratio of emotion inputs which are correctly detected to the total number of emotion [14]. Accuracy rate can be evaluated by the expression [15] given below:

$$\mathrm{AR} = \frac{Number\ of\ correctly\ estimated\ test\ data}{Total\ number\ of\ testing} \qquad (2)$$

B. Precision (P): This metric is defined as the ratio of correctly recognized emotions for each class to the correctly recognized emotions for all classes [14]. Precision rate can be calculated by the expression [9] given below:

$$\mathrm{P} = \frac{No.of\ correctly\ recognized\ emotions\ for\ eac□\ of\ t□e\ \ class}{total\ no.\ of\ recognized\ emotions\ for\ all\ emotion\ classes} \ (3)$$

C. F-measure: F-measure is the parameter that is calculated by fusing precision and recognition rate. This factor is evaluated to analyze the comprehensive efficiency of the recognition system [14] and can be calculated by using the following equation:

$$\text{F-measure} = \frac{2*accuracy*precision}{accuracy+precision} \qquad (4)$$

VII. Conclusion

This paper describes the overall emotion recognition system from audio signals by emphasizing on its two main processes which are features extraction and machine learning tools to classify the emotions and discussed the factors from which efficiency of any machine learning model can be evaluated.

VIII. REFERENCES

[1] H .Palo and M .Mohanty, "Wavelet based feature combination for recognition of emotions", Ain Shams Engineering Journal, 2017.

[2] I .Bisio, A .Delfino, F .Lavagetto, M .Marchese and A .Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications", IEEE Transactions on Emerging Topics in Computing, vol .1, no .2, pp .244-257, 2013.

[3] R. Khosrowabadi, H. Quek, A. Wahab and K. Ang, "EEG-based Emotion Recognition Using Self-Organizing Map for Boundary Detection", 2010 20th International Conference on Pattern Recognition, 2010.

[4] M .Sezgin, B .Gunsel and G .Kurt, "Perceptual audio features for emotion detection", EURASIP Journal on Audio, Speech, and Music Processing, vol .2012, no .1, 2012.

[5] Y .Yang, Y .Lin, Y .Su and H .Chen, "A Regression Approach to Music Emotion Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol .16, no .2, pp .448-457, 2008.

[6] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang and Lian Li, "Speech Emotion Recognition Using Fourier Parameters", IEEE Transactions on Affective Computing, vol .6, no .1, pp .69-75, 2015.

[7] C .Chang, C .Wu, C .Lo, C .Wang and P .Chung, "Music emotion recognition with consideration of personal preference", The 2011 International Workshop on Multidimensional)nD (Systems, 2011.

[8] K. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", IEEE Signal Processing Letters, vol. 13, no. 1, pp. 52-55, 2006.

[9] Tao Li and M. Ogihara, "Toward intelligent music information retrieval", IEEE Transactions on Multimedia, vol. 8, no. 3, pp. 564-574, 2006.

[10] S. Jothilakshmi, V. Ramalingam and S. Palanivel, "Unsupervised speaker segmentation with residual phase and MFCC features", Expert Systems with Applications, vol. 36, no. 6, pp. 9799-9804, 2009.

[11] M .Muller, D .Ellis, A .Klapuri and G .Richard, "Signal Processing for Music Analysis", IEEE Journal of Selected Topics in Signal Processing, vol . 5, no .6, pp .1088-1110, 2011.

[12] B. Yu, H. Li and C. Fang, "Speech Emotion Recognition based on Optimized Support Vector Machine", Journal of Software, vol. 7, no. 12, 2012.

[13] S .Koolagudi and K .Rao, "Emotion recognition from speech :a review", International Journal of Speech Technology, vol .15, no .2, pp .99-117, 2012.

[14] R .Lanjewar, S .Mathurkar and N .Patel, "Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model)GMM (and K -Nearest Neighbor)K-NN (Techniques", Procedia Computer Science, vol .49, pp .50-57, 2015.

[15] N .Nalini and S .Palanivel, "Music Emotion Recognition :The combined evidence of MFCC and Residual Phase", Egyptian Informatics Journal, pp . 1-10, 2015.

[16] H .Palo and M .Mohanty, "Wavelet based feature combination for recognition of emotions", Ain Shams Engineering Journal, 2017.

[17] L. Chen, X. Mao, Y. Xue and L. Cheng, "Speech emotion recognition: Features and classification models", Digital Signal Processing, vol. 22, no. 6, pp. 1154-1160, 2012.

[18] Z. Zhang, E. Coutinho, J. Deng and B. Schuller, "Cooperative Learning and its Application to Emotion Recognition from Speech", IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1-1, 2014.

[19] L. Chen, X. Mao, Y. Xue and L. Cheng, "Speech emotion recognition: Features and classification models", Digital Signal Processing, vol. 22, no. 6, pp. 1154-1160, 2012.

[20] Y. Yang and H. Chen, Music emotion recognition. Boca Raton, Fla: CRC, 2011.

[21] T. Li, M. Ogihara and G. Tzanetakis, Music data mining. Boca Raton, Fla.: CRC Press, 2012.

[22] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods", Speech Communication, vol. 48, no. 9, pp. 1162-1181, 2006.

[23] M. El Ayadi, M. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.

[24] G. Eliot, The mill on the Floss. New York: Open Road Integrated Media, 2016.