Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.528 – 531

RESEARCH ARTICLE

Feature Extraction and Classification of Emotions in Wave Files Using Crossbreed Algorithm

Aastha Joshi[1], Rajneet Kaur[2]

[1]Department of Computer Science and Engineering Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[2]Department of Computer Science and Engineering Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[1] joshiaastha10@yahoo.com; [2] rosy.rajneet@gmail.com

Abstract— The importance of automatically recognizing emotions in human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. In this paper, emotion classification method based on hybrid of SVM and HMM algorithm is presented. Four primary human emotions, including anger, aggressive, happiness and sadness are investigated. For speech emotion recognition, we extracted 15 features to form the feature vector. Extracted features were sent into the improved crossbreed algorithm (hybrid of HMM & SVM) for classification and recognition. Results show that the selected features are robust and effective for the emotion recognition and give better accuracy compared to individual SVM & HMM classifiers.

Key Terms: - SER System; features extraction; SVM & HMM; GA algorithm

I. INTRODUCTION

Speech emotion recognition is one of the latest challenges in speech processing. Besides human facial expressions speech has proven as one of the most promising modalities for the automatic recognition of human emotions. Automatic Emotion Recognition (AER) can be done in two ways, either by speech or by facial expressions. In the field of Human Computer Interaction (HCI), speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures. The importance of automatically recognizing emotions in human speech has grown with increasing role of spoken language interfaces in the field of human machine interaction to make the human machine interface more efficient. It can also be used for in-car board system where information of the mental state of the driver maybe provided to initiate his/her safety. In automatic remote call center, it is used to timely detect customers‟ dissatisfaction. In E-learning field, identifying students‟ emotion timely and making appropriate treatment can enhance the quality of teaching [1].

Our system has been fully implemented (in matlab) and tested for audio wave files. The objective is to efficiently extract the features from the uploaded wave file and train the system and the different emotions are classified using crossbreed (combination of HMM & SVM) algorithm.

II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker’s voice, and taking appropriate pattern

recognition methods to identify emotional states [2]. Our speech emotion recognition system contains four main modules: speech input, feature extraction, hybrid classification algorithm using HMM and SVM classifier, and emotion output.

The general architecture for SER system has three steps [3]:

i. A speech processing system extracts some appropriate quantities from signal, such as minimum & maximum frequency, noise, loudness, compression ratio, bit depth, sample rate, spectral spread, spectral flatness, spectral roll off, spectral centroid.
ii. These quantities are summarized into reduced set of features,
iii. A classifier is trained in a supervised manner with example data how to associate the features to the emotions.

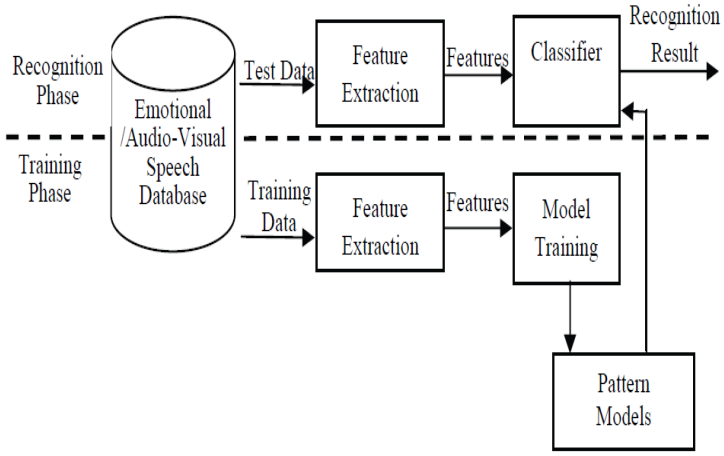

Fig 1. Speech Emotion Recognition System

III. FEATURE EXTRACTION

In pattern recognition, feature extraction is one of the special form of dimensionality reduction. Feature resources required to describe a large set of data accurately. Basically feature extraction is based on partitioning speech into small intervals known as frames. Sixteen features has been evaluated for use in the system.

The features uses in our system are:

- *Frequency:* It is the main feature of an audio file. Maximum, minimum and average frequencies are calculated for each wave file.
- *Spectral Centroid:* Spectral Centroid is the weighted mean frequency. It indicates where the "center of mass" of the spectrum is. Because the spectral centroid is a good predictor of the "brightness" of a sound [4], it is widely used in digital audio and music processing as an automatic measure of musical timbre [5].

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=0}^{N} M_t[n]}$$

where $M_t(n)$ is magnitude of Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to “brighter” textures with more high frequencies [7].

- *Spectral Spread* : We define the spectrum spread as the spread of spectrum around its mean value, i.e. the variance of the above defined distribution [5].
- *.Spectral Flatness :* Spectral flatness is a measure used to characterize an audio spectrum. The spectral flatness can also be measured within a specified sub band, rather than across the whole band. Spectral flatness provides a way to quantify how tone-like a sound is, as opposed to being noise-like [6].

$$Flatness = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} = exp^{\frac{1/N \sum_{n=0}^{N-1} \ln x(n)}{1/N \sum_{n=0}^{N-1} x(n)}}$$

where *x(n)* represents the magnitude of bin number *n* [15].

- *Spectral Rolloff :* Spectral rolloff point is defined as the *Nth* percentile of the power spectral distribution, where *N* is usually 85% or 95% [8]. This measure is useful in distinguishing voiced speech from unvoiced: unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, where most of the energy for voiced speech and music is contained in lower bands.

$$\sum_{n=1}^{Rt} M_t(n) = 0.85 * \sum_{n=1}^{N} M_t(n)$$

Where R_t is the frequency below which 85% of the magnitude distribution is concentrated [7].

- *Loudness:* Loudness features aims at simulating the human sensation of loudness. Loudness is "that attribute of auditory sensation in terms of which sounds would be ordered on scale extending from soft to loud" [8].
- *Noise:* Noise is an undesirable component that obscures a wanted signal. **Noise figure** (NF) and **noise factor** (*F*) are measures of degradation of the signal-to-noise ratio (SNR), caused by components in radio frequency (RF) signal chain. The noise factor is defined as the ratio of the output noise power of a device to the portion thereof attributable to thermal noise in the input termination at standard noise temperature T_0 [9].

The **noise factor** *F* of a system is defined as [14]:

$$F = \frac{SNR(in)}{SNR(out)}$$

where SNR(in) and SNR(out) are the input and output signal-to-noise ratios, respectively.

- *Time:* The **real time factor (RTF)** is a common metric of measuring the speed of an automatic speech recognition system. It can also be used in other context where an audio or video signal is processed at nearly constant rate. If it takes time P to process an input of duration I, the real time factor is defined as [12].

$$RTF = \frac{P}{I}$$

- *Compression Ratio :* Audio can often be compressed at 10:1 with imperceptible loss of quality. Data compression ratio is defined as the ratio between the *compressed size* and the *uncompressed size* [11]:

$$\text{Compression Ratio} = \frac{\text{Compressed Size}}{\text{Uncompressed Size}}$$

- *Sample Rate:* This value simply represents the number of samples captured per second in order to represent the waveform; the more samples per second, the higher the resolution, and thus the more precise the measurement is of the waveform. Capturing a sound at a particular ***frequency*** requires a sampling rate of at least twice that frequency (known as the *Nyquist* frequency) [12].
- *Audio Bit Depth:* In digital audio, **bit depth** describes the number of bits of information recorded for each sample. Bit depth corresponds to the **resolution** of each sample in a set of digital audio data [13]. In our study we assume bit depth to be 16.

These features has been extracted for every uploaded wav file and then converted to binary format which is taken as input to plot histogram. The average of features extracted for one kind of emotion has been saved in database as .mat file to form clusters.

IV. RESULTS AND DISCUSSIONS

Firstly wave file is loaded and then converted to binary form to plot histogram which is taken as input to extract relevant features and these features are saved to .mat file.

Using testing toolbox which has been trained using HMM & using SVM category of selected wave file has been checked.

For calculating the accuracy of our crossbreed algorithm, Genetic Algorithm (GA) which is an automatic optimization algorithm is used. It has been discovered that accuracy of our crossbreed algorithm (combination of SVM & HMM) comes out to be more as compared to individual SVM & HMM algorithm.

V. CONCLUSION

The results of speech emotion recognition system are mainly specified in terms of accuracy of matching. The proposed crossbreed technique gives better accuracy of approximately 96% compared to other used techniques.

REFERENCES

[1] Ayadi M. E., Kamel M. S. and Karray F., 'Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases', Pattern Recognition, 44(16), 572-587, 2011.
[2] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Smart Home, Vol. 6, No. 2, April, 2012.
[3] [Online]. Available: http://crteknologies.fr/projects/emospeech/
[4] Grey, J. M., Gordon, J. W., 1978. Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America 63 (5), 1493–1500, doi:10.1121/1.381843
[5] Schubert, Emery; Wolfe, Joe; Tarnopolsky, Alex (2004). "Spectral centroid and timbre in complex, multiple instrumental textures".Proceedings of the 8th International Conference on Music Perception & Cognition, North Western University, Illinois. International Conference on Music Perception & Cognition, Lipscomb, S.D.; Ashley, R.; Gjerdingen, R. O.; Webster, P. (Eds.). Sydney, Australia: School of Music and Music Education; School of Physics, University of New South Wales.
[6] Shlomo Dubnov (2004). "Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes". Signal Processing Letters 11 (8): 698–701. doi:10.1109/LSP.2004.831663. ISSN 1070-9908.
[7] George Tzanetakis and Perry Cook, Musical Genre Classification of Audio Signals IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 10, NO. 5, JULY 2002
[8] Dalibor,Matthias and Christian, "Features of Content Based Audio", Advances in Computers Vol. 78, pp. 71-150,2010.
[9] [Online]. Available: http://en.wikipedia.org/wiki/Noise_figure
[10] [Online]. Available: http://en.wikipedia.org/wiki/Real_time_factor
[11] Data Compression: The Complete Reference. 4th Edition. David Salomon (with contributions by Giovanni Motta and David Bryant). Published by Springer (Dec 2006). ISBN 1-84628-602-6.
[12] [Online]. Available: http://manual.audacityteam.org/man/Digital_Audio
[13] [Online]. Available: http://en.wikipedia.org/wiki/Audio_bit_depth
[14] Agilent 2010, Fundamentals of RF and Microwave Noise Figure Measurements, Application Note, 57-1, p. 5
[15] [Online]. Available: http://en.wikipedia.org/wiki/Spectral_flatness

