International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

Human Emotion Recognition in Speech using Ant Colony Optimization

Swati Pahune
PG student(Electronics & communication)
Vidarbha Institute of Technology,Nagpur

Nilesh Bodne
Asst.Professor(Electronics& communication)
Vidarbha Institute of Technology,Nagpur

Abstract

Emotional speech recognition is an area of great interest for human-computer interaction. The system must be able to recognize the user's emotion and perform the actions accordingly. It is essential to have a framework that includes various modules performing actions like speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions. The classifications of features involve the training of various emotional models to perform the classification appropriately. Another important aspect to be considered in emotional speech recognition is the database used for training the models [1].

Keywords: Classifier, Emotion recognition, Feature extraction, Feature selection.

I. Introduction

Speech is a complex signal which contains information about the message, speaker, language and emotions. Speech is produced from a time varying vocal tract system excited by a time varying excitation source. Emotion on other side is an individual mental state that arises spontaneously rather than through conscious effort. There are various kinds of emotion which are present in a speech. The basic difficulty is to cover the gap between the information which is captured by a microphone and the corresponding emotion, and to model the specific association. This gap can be bridged by narrowing down various emotions in few, like anger, happiness, sadness, surprise, fear, and neutral. Emotions are produced in the speech from the nervous system consciously, or unconsciously. Emotional speech recognition is a system which basically identifies the emotional as well as physical state of human being from his or her voice [1]. Emotion recognition is gaining attention due to the widespread applications into various domains detecting frustration, disappointment, surprise/amusement etc.There are many approaches towards automatic recognition of emotion in speech by using different feature vectors.

A proper choice of feature vectors is one of the most important tasks. The feature vectors can be distinguished into the following four groups: continuous (e.g., energy and pitch), qualitative (e.g., voice quality) spectral (e.g., MFCC), and features based on the Teager energy operator (e.g., TEO autocorrelation envelope area).For classification of speech, methodologies followed are: HMM, GMM, ANN, k-NN, and several others

as well as their combination which maintain the advantages of each classification technique. After studying the related literature it can be identified that the feature set which is mostly employed is comprised of pitch, MFCCs, and HNR. Additionally, the HMM technique is widely used by the researchers due to its effectiveness. Feature extraction by temporal structure of the low level descriptors or large portion of the audio signal is taken could be helpful for both the modeling and classification processes.

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. Fig.1 indicates the speech emotion system components.

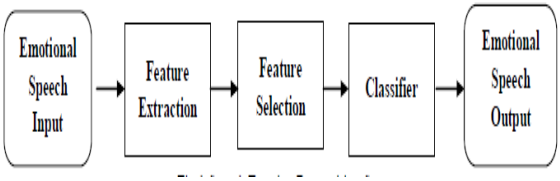

Fig. 1 Speech Emotion Recognition System

Like typical pattern recognition systems, speech emotion recognition system contains four main modules: speech input, feature extraction, feature selection, classification, and emotion output. Since a human cannot classify easily natural emotions, it is difficult to expect that machines can offer a higher correct classification. A typical set of emotions contains 300 emotional states which are decomposed into six primary emotions like anger, happiness, sadness, surprise, fear, neutral. Success of speech emotion recognition depends on naturalness of database. [2].

There are six databases accessible: two freely accessible ones, the Danish Emotional Speech corpus (DES) and Berlin Emotional Database (EMO-DB), and four databases from the Interface venture with Spanish, Slovenian, French and English enthusiastic discourse. These databases contain acted enthusiastic discourse. Regarding legitimacy, there is by all accounts three sorts of databases utilized as a part of the SER .: sort one is acted enthusiastic discourse with human marking. This database is acquired by requesting that an on-screen character talk with a predefined feeling. As of late solid complaints have developed against the utilization of acted feelings. It was demonstrated that acted and unconstrained examples vary in the perspective of elements and correctness's [6], sort 2 is true passionate discourse with human marking. This databases are originating from genuine frameworks (for instance call-focuses) and sort three is evoked enthusiastic discourse with self-report as opposed to marking. Where feelings are incited and self-report is utilized for naming control. [7] Seemingly, diverse sorts of information bases are reasonable for various purposes. Sort 1still can be useful, now and again where mostly hypothetical exploration is pointed, rather that development of a genuine application for the business.

III. Feature Extraction and Selection

Speech signal composed of large number of parameters which indicates emotion contents of it. Changes in these parameters indicate changes in the emotions. Therefore proper choice of feature vectors is one of the most important tasks. There are many approaches towards automatic recognition of emotion in speech by using different feature

vectors. By employing feature extraction technique number of features can be extracted from the emotional speech. To achieve accurate identification of emotion classifier should provided with single best feature. Therefore there is need of systematic feature selection to reduce unuseful features from the base features. To select best features Forward Selection method can be used. The remaining features can be used by classifier to increase classification accuracy.

Feature vectors can be classified as long-time and short-time feature vectors. The long-time ones are estimated over the entire length of the utterance, while the short-time ones are determined over window of usually less than 100ms. The long-time approach identifies emotions more efficiently. Short time features uses interrogative phrases which has wider pitch contour and a larger pitch standard deviation. Most common features used by researchers are

Energy and related features:

The Energy is the basic and most important feature in speech signal. We can obtain the statistics of energy in the whole speech sample by calculating the energy, such as mean value, max value, variance, variation range ,contour of energy [1].

Pitch and related features:

The value of pitch frequency can be calculated in each speech frame and the statistics of pitch can be obtained in the whole speech sample. These statistical values reflect the global properties of characteristic parameters. Each Pitch feature vector has the same 19 dimensions as energy.

Qualitative Features:

Emotional contents of a utterance is strongly related with its voice quality. The voice quality can be numerically represented by parameters estimated directly from speech signal. The acoustic parameters related to speech quality are:

(1) Voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level;

(2) voice pitch;

(3) phrase, phoneme, word and feature boundaries;

(4) temporal structures.

Linear Prediction Cepstrum Coefficients (LPCC):

LPCC embodies the characteristics of particular channel of speech. Person with different emotional speech will have different channel characteristics, so we can extract these feature coefficients to identify the emotions contained in speech. The computational method of LPCC is usually a recurrence of computing the linear prediction coefficients (LPC), which is according to the all-pole model.

Wavelet Based features:

Speech signal is a non-stationary signal, with sharp transitions, drifts and trends which is hard to analyze. Wavelets have energy concentrations in time and are useful for the analysis of transient signals. A time frequency representation of such signals can be performed using wavelets. The Discrete Wavelet Transform (DWT) is computed by successive low-pass and high-pass filtering of the discrete time-domain signals of Speaker emotional state.

Mel-Frequency Cepstrum Coefficients (MFCC):

MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used feature of the speech, with a simple calculation, good ability of the distinction, anti-noise and other advantages [7].,

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

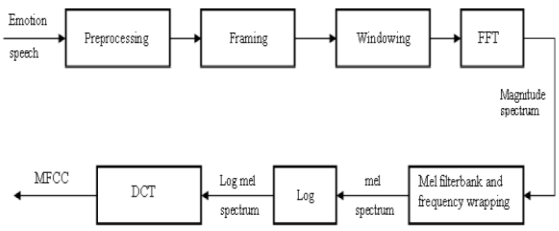

Fig. 2 Block Diagram of the MFCC feature extraction process

Pre-processing: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This pre-emphasis is done by using a filter

Framing: It is a process of segmenting the speech samples obtained from analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 msec. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal.

Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame [8].

FFT: Fast Fourier Transform (FFT) algorithm is widely used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain [8].

Mel Filter bank and Frequency wrapping: The mel filter bank consists of overlapping triangular filters with the cut-off frequencies determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale [8].

Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition [8].

Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalisation step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components [8].

In this work, we have used mfcc for feature extraction where input is a wav file containing emotional speech utterances from Danish Emotion Database. The five emotions used are, anger, happiness, sadness, fear, and neutral state. Then Segmenting all concatenated voiced speech signal into 25.6ms-length frames. Then Estimate the logarithm of the magnitude of the discrete Fourier Transform (DFT) for all signal frames. Filtering out the center frequencies of the sixteen triangle band-pass filters corresponding to the mel frequency scale of individual segments. Estimate inversely the IDFT to get all 10-order MFCC coefficients. After that analyzing all extracted MFCC dataset for Ant Colony Optimization.

This all processes are implemented in Maltab program

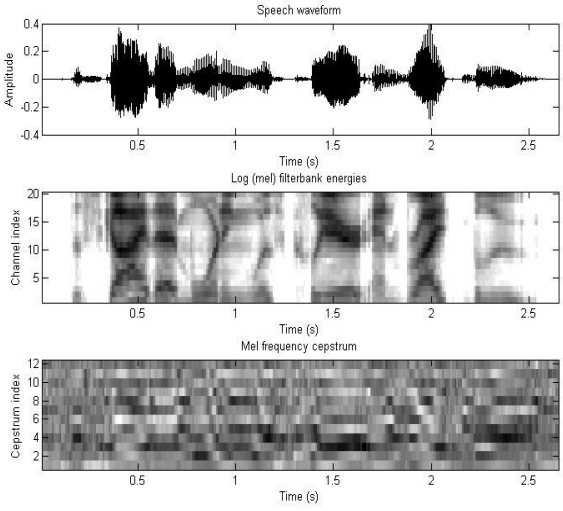

Fig. 3 Speech waveform, filterbank energies & mel frequency ceptrum

IV. Ant Colony Optimization

Ant Colony Optimization (ACO) is a population based met heuristic approach to find

approximate solutions to difficult optimization problems. The inspiring source of ACO is the pheromone trail laying behavior of real ants, which use pheromone as a communication medium.These pheromone trail values are modified at runtime based on a problem-dependent heuristic function and the amount of pheromone deposited by the ants while they traverse between their colonies and a food source. In ACO, pheromone trail values serve as distributed, numerical information, which the ants use to construct solutions probabilistically. There is one solution per ant. The higher the pheromone value (initial edge), the higher the probability of an ant choosing that particular trail will be.As mentioned above, it is repeatedly applied until a termination condition is satisfied. In practice, a termination condition may be the maximum number of solutions generated, the maximum CPU time elapsed, or the maximum number of iterations without improvement in solution of favoring the exploration of new areas in the search space.

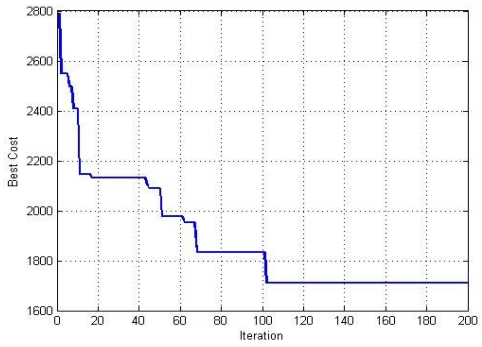

Fig.4 iteration using ACO

V Conclusion

Conclusion in this paper, latest work done in the field of Speech Emotion Recognition is talked about and most utilized strategies for highlight the extraction are also looked into. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. In this research work we extracted the feature from mfcc which are now going to give as input to the ACO which is going to use as a classifier, so it may give good result.Also it opens the new method for correspondence amongst human and machine.

References

[1] S. Ramakrishnan, " Recognition of Emotion from Speech: A Review", Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi,India. ISBN 978-953-51-0291-5 Published Intec China and Europe, pg no. 121-138, 14 March, 2012.

[2] Dr.S. Ramakrishnan, " Speech Enhancement, Modeling and Recognition- Algorithms and Applications".

[3] Daniel Erro, Eva Navas, Inma Hernáez, and Ibon Saratxaga, "Emotion Conversion Based on Prosodic Unit Selection" , IEEE Transactions On Audio, Speech And Language Processing, Vol. 18, No. 5, pp.974-983, July 2010.

[4] Panagiotis C. Petrantonakis , and Leontios J. Hadjileontiadis, " Emotion Recognition From EEG Using Higher Order Crossings", IEEE Trans. on Information Technology In Biomedicine, Vol. 14, No. 2,pp.186-197, March 2010.

[5] Dipti D. Joshi, Prof. M. B. Zalt (Jan. - Feb.2013) on Speech Emotion Recognition: A Review,IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) SSN: 2278-2834, ISBN: 2278-8735. Volume 4, Issue 4, PP 34-37, www.iosrjournals.org

[6] Ellen Douglas-Cowie , Nick Campbell , Roddy Cowie , Peter Roach, "Emotional Speech: Towards a New Generation Of Databases", Speech Communication Vol. 40, pp.33–60 ,2003.

[7] John H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise

for Environmental Robustness in Speech Recognition", Speech Communication, Special Issue on Speech under Stress, vol. 20(1-2), pp. 151-170, November 1996

[8] Vaishali M. Chavan, V.V. Gohokar , " Speech Emotion Recognition by using SVM-Classifier ", International Journal of Engineering and Advanced Technology (IJEAT)