International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 2 134– 137

Emotion Recognition through Combination of Speech and Image Processing: A Review

Swati Pahune,
PG Student,
Swati_pahune@rediffmail.com,

Nilu Mishra
Assistant Professor,
niki.mishra@gmail.com

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 emotions 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 bridge 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.
system describes in section four.

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.

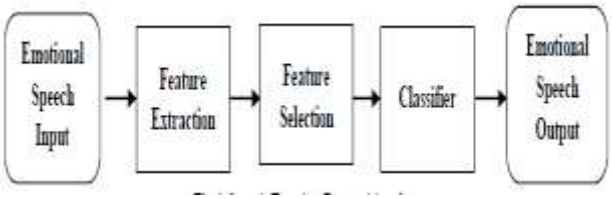

Figure .1 Speech Emotion Recognition System

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 available: two publicly available ones, the Danish Emotional Speech corpus (DES) and Berlin Emotional Database (EMO-DB), and four databases from the Interface project with Spanish, Slovenian, French and English emotional speech. All of these databases contain acted emotional speech. With respect to authenticity, there seems to be three types of databases used in the SER research: type one is acted emotional speech with human labeling. This database is

obtained by asking an actor to speak with a predefined emotion. Recently strong objections have emerged against the use of acted emotions. It was shown that acted and spontaneous samples differ in the view of features and accuracies [6], type 2 is authentic emotional speech with human labeling. This databases are coming from real-life systems (for example call-centers) and type three is elicited emotional speech with self-report instead of labeling. Where emotions are provoked and self-report is used for labeling control. [7] Seemingly, different types of databases are suitable for different purposes. Type 1still can be of use in, some cases where mainly theoretical research is aimed, rather that construction of a real life application for the industry.

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. 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.

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].

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 Speaker emotional state identification applications the Discrete Wavelet Transform offers the best solution. 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.

Basic framework for emotional recognition

The input files are speech signals. Fig.1 gives the basic framework of emotional speech recognition. The feature extraction script extracts the features that represent global statistics. In the Post-processing step, the interface problem between the script for feature extraction and the feature selection technique can be solved. Then feature selection eliminates irrelevant features that hinder the recognition rates. It lowers the input dimensionality and saves the computational time. Distribution models like GMMs are trained using the most discriminative aspects of the feature. The classifiers distinguish the types of emotion. Bio signals such as ECG, EEG,GSR, face and body images are an interesting alternative to detect emotional states, the mechanism of emotion recognition using these bio signals.

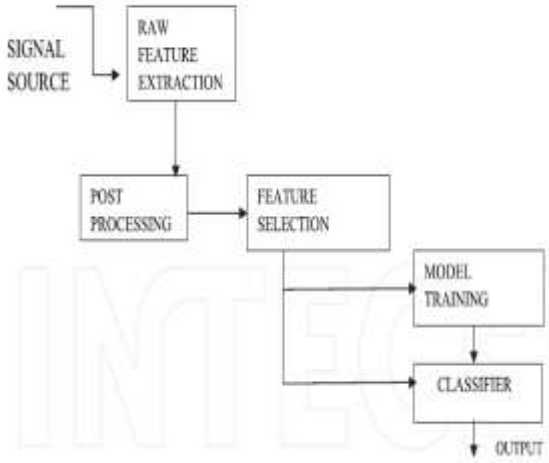

Figure 2 Basic frame of SER[1]

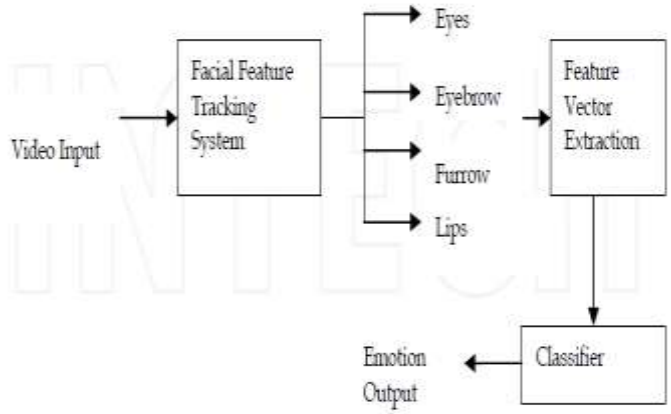

Figure 3 Processing of Feature extraction

Galvanic Skin Response is the measure of skin conductivity. There is a correlation between GSR and the arousal state of body. In the GSR emotional recognition system, the GSR signal is physiologically sensed and the feature is extracted using Immune Hybrid Particle Swarm Optimization (IH-PSO). The extracted features are classified using neural network classifier to identify the type of emotion .In the facial emotion recognition the facial expression of a person is captured as a video and it is fed into the facial feature tracking system. Fig 3 gives a basic framework of facial emotional recognition. In facial feature tracking system, facial feature tracking algorithms such as Wavelets, Dual-view point-based model etc. are applied to track eyes, eyebrows, furrows and lips to collect all its possible movements. Then the extracted features are fed into classifier like Naive Bayes, TAN or HMM to classify the type of emotion.

3. Emotional speech database

There should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. According to some studies the following are the most relevant factors to be considered:

1) Real-world emotions or acted on
2) Who utters the emotions
3) How to simulate the utterances
4) Balanced utterances or unbalanced utterances

Utterances are uniformly distributed over emotions Most of the developed emotional speech databases are not available for public use. Thus, there are very few benchmark databases that can be shared among researchers. Most of the databases share the following emotions: anger, joy, sadness, surprise, boredom, disgust, and neutral.[1]

Types of DB

At the beginning of the research on automatic speech emotion recognition, acted speech was used and now it shifts towards more realistic data. The databases that are used in SER are classified into 3 types. Fig 4 briefs the types of databases. Table 1 gives a detailed list of speech databases.

Type 1 is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion, e.g. DES, EMO-DB.

Type 2 is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come from real-life applications for example call-centers.[1]

Type 3 is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated.

4. Acoustic characteristics of emotions in speech

The prosodic features like pitch, intensity, speaking rate and voice quality are important to identify the different types of emotions. In particular pitch and intensity seem to be correlated to the amount of energy required to express a certain emotion. When one is in a state of anger, fear or joy; the resulting speech is correspondingly loud, fast and enunciated
with strong high-frequency energy, a higher average pitch, and wider pitch range, whereas with sadness, producing speech that is slow, low-pitched, and with little high-frequency energy. In Table 2, a short overview of acoustic characteristics of various emotional states is provided.[1]

EMOTIONS / CHARACTERISTICS	JOY	ANGER	SADNESS	FEAR	DISGUST
Pitch mean	High	very high	very low	very high	very low
Pitch range	High	high	Low	High	high-male low-female
Pitch variance	High	very high	Low	very high	Low
Pitch contour	incline	decline	Decline	Incline	Decline
Intensity mean	High	very high-male high-female	Low	medium/ high	Low
Intensity range	High	high	Low	High	Low
Speaking Rate	High	low-male high-female	high-male low-female	High	very low-male low-female
Transmission Durability	Low	low	High	Low	High
Voice Quality	modal/ tense	Sometimes breathy; Moderately blaring timbre	Resonant timbre	Falsetto	Resonant timbre

Table 2. Acoustic Characteristics of Emotions

REFERENCES

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

[2] Textbook of Speech Enhancement, Modeling and Recognition- Algorithms and ApplicationsEdited by Dr. S Ramakrishnan

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

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

[5] Speech Emotion Recognition: A Review Dipti D. Joshi, Prof. M. B. Zalt, IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) SSN: 2278-2834, ISBN: 2278-8735. Volume 4, Issue 4 (Jan. - Feb. 2013), 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 forEnvironmental Robustness in Speech Recognition', Speech Communication, Special Issue on Speech Under Stress,vol. 20(1-2), pp. 151-170, November 1996.