Emotion Detection of Speech Signals with Analysis of Salient Aspect Pitch Contour

Rode Snehal Sudhkar
ME Student
Dept. of Electronics and Telecomunication Engineering
JSPM's Jaywantrao Sawant College of Engineering
Hadapsar, Pune – 411028

Manjare Chandraprabha Anil
Research Scholar
Dept. of Electronics and Telecomunication Engineering
JSPM's Rajashri Shahu College of Engineering
Tathawade, Pune – 411033

Abstract— Emotion detection of speech in human machine interaction is very important. Framework for emotion detection is essential, 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 features used for emotion detection of speech are prosody features, spectral features and voice quality features. The classifications of features involve the training of various emotional models to perform the classification appropriately. The features selected to be classified must be salient to detect the emotions correctly. And these features should have to convey the measurable level of emotional modulation.

Keywords—Prosody, Classifier, KLD, GMM, HMM, pitch contour

I. INTRODUCTION

A speech signal is naturally occurring signal and hence is random in nature. The signal expresses different ideas, communication and hence has lot of information. There are number of automatic speech detection system and music synthesizer commercially available. However despite significant progress in this area there still remain many things which are not well understood. Detection of emotions from speech is such an area. The speech signal information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people. Detection of human emotions will improve communication between human and machine. The human instinct detects emotions by observing psycho-visual appearances and voices. Machines may not fully take human place but still are not behind to replicate this human ability if speech emotion detection is employed. Also it could be used to make the computer act according to actual human emotions. This is useful in various real life applications as systems for real life emotion detection using corpus of agent client spoken dialogues from call centre like for medical emergency, security, prosody generation, etc. The alternative emotion detection is through body, face signals, and bio signals such as ECG, EEG. However in certain real life applications these methods are very complex and sometimes impossible, hence emotion detection from speech signals is the more feasible option. Good results are obtained by the signal processing tools like MATLB and various algorithms (HMM, SVM) but their performance has limitations, while combination and ensemble of classifiers could represent a new step towards better emotion detection.

II. BASIC THEORY FOR EMOTION DETECTION

In general, emotion detection system consist of speech normalization, feature extraction, feature selection, classification and then the emotion is detected.

Figure.1 gives the basic flow for the emotion detection from input speech. First noise and d.c components are removed in speech normalization then the feature extraction and selection is carried out. The most important part in further processing of input speech signal to detect emotions is extraction and selection of features from speech. The speech features are usually derived from analysis of speech signal in both time as well as frequency domain. Then the data base is generated for training and testing of the extracted speech features from input speech signal. In the last stage emotions are detected by the classifiers. Various pattern recognition algorithms (HMM, GMM) are used in classifier to detect the emotion.

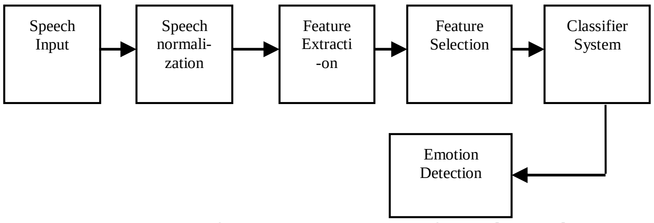


Fig.1 System for Emotion Detection of Speech Signal

A. Speech Normalization

The collected emotional data usually gets degraded due to external noise (background and "hiss" of the recording machine). This will make the feature extraction and classification less accurate. Hence normalization is critical step in emotion detection. In this preprocessing stage speaker and recording variability is eliminated while keeping the emotional discrimination. Generally two types of normalization techniques are performed they are energy normalization and pitch normalization.

B. *Feature Extraction and Selection from Emotional Speech*

After normalization of emotional speech signal, it is divided into segments to form their meaningful units. Generally these units represent emotion in a speech signal. The next step is the extraction of relevant features. These emotional speech features can be classified into different categories. One classification is long term features and short term features. The short term features are the short time period characteristics like formants, pitch and energy. And long term features are the statistical approach to digitised speech signal. Some of the frequently used long term features are mean and standard deviation. The larger the feature used the more improved will be the classification process. After extraction of speech features only those features which have relevant emotion information are selected. These features are then represented into n- dimensional feature vectors [10]. The prosodic features like pitch, intensity, speaking rate and variance are important to identify the different types of emotions from speech. In Table 1 acoustic characteristics of various emotions of speech is given. The observations which are expressed in below table 1 are taken by using Paart software.

Characteristics	Happy	Anger	Enquiry	Fear	Surprise
Emotion					
Pitch Mean	High	Very high	High	Very high	Very high
Pitch Range	High	High	High	High	High
Pitch Variance	High	Very high	High	Very high	Very high
Pitch Contour	Incline	Decline	Moderate	Incline	Incline
Speaking Rate	High	High	Medium	High	High

Table. 1 Acoustic Characteristics of Emotions

C. *Database for Training, Testing*

The database is used for training, testing and development of feature vector. A good database is important for desired result. Various databases are available created by speech processing community. The databases can be divided into training data set and testing data set. The famous databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database, TIMIT. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

The databases that are used in SER are classified into 3 types.

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.

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.

D. *Classifiers to Detect Emotions*

Various classifiers like GMM, HMM are used according to their specific usage based on selected features. Emotions are predicated using classifiers and selected feature vectors to predict emotion from training data set and the development data set. For the training data sets the emotion information are known whereas for testing data set the emotion information are unknown. When performing analysis of complex data one of the major problems comes from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still the data with sufficient accuracy.

Typically, in speech recognition, we divide speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage.

E. *Pitch Features*

Pitch features are used for emotional discrimination of speech signal. The features of speech contour such as mean, median, standard deviation, variance, minimum, maximum, range, kurtosis and skewness are used to find emotional discrimination of speech signal. These pitch statistics are grouped into voiced level and sentence level features.

III. METHODOLOGY

A. *Analysis of Pitch Features*

The fundamental frequency or F0 contour which is prosodic feature provides the tonal and rhythmic properties of speech. It predominantly describes the speech source rather than the vocal tract properties. Although it is also used to emphasize linguistic goals conveyed in speech, it is largely independent of the specific lexical content of what is spoken in most languages. The fundamental frequency is also a supra segmental frequency, where the information is conveyed over longer time scales than other segmental speech correlates such spectral envelope features. Therefore, rather than using the pitch contour over an entire utterance or sentence such as the mean, maximum and standard deviation. However, it is not clear that estimating global statistics from the pitch contour will provide local information of emotional modulation. Therefore, in addition to sentence-level analysis, we investigate alternative time units for the F0 contour analysis that is voiced level analysis.

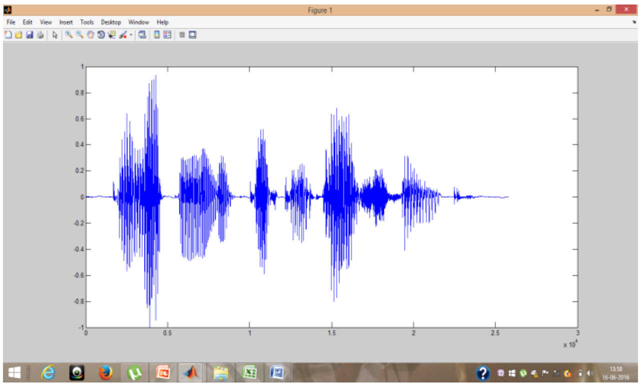


Fig 2: Input Speech Signals

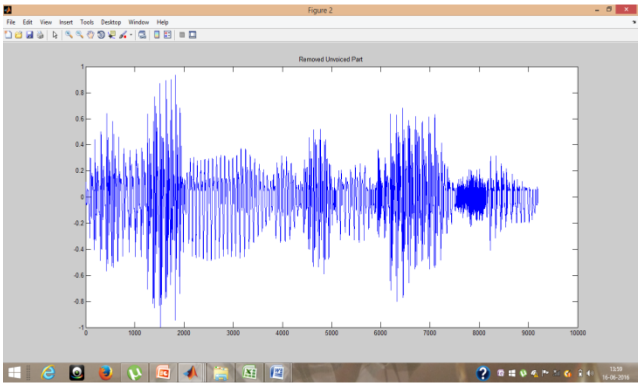


Fig 3: Input signal without unvoiced part

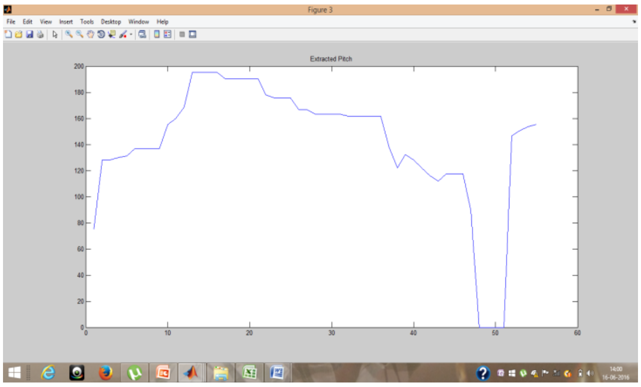


Fig 4: Extracted pitch part of input speech

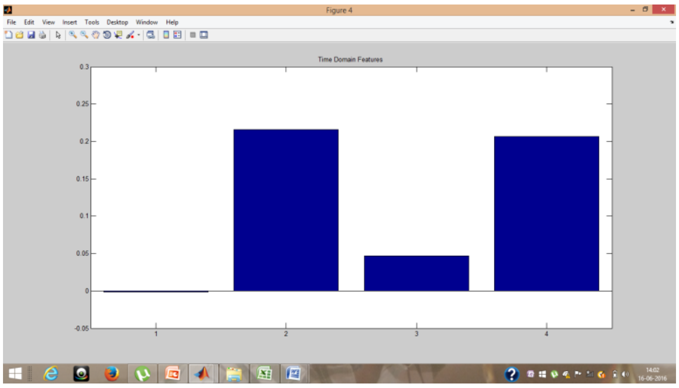


Fig 5: Time domain features of input speech

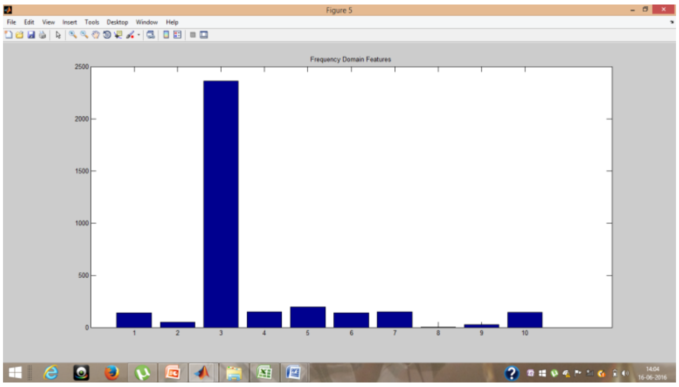


Fig 6: Frequency domain features of Input speech

B. *Emotional Discrimination of Speech Signal*

In our project, we use Kullback–Leibler divergence method for comparing the neutral and emotional data. KLD is the approach to identify and quantify the pitch features with higher level of emotional modulation. The distribution of pitch features extracted from emotional database is compared with distributions of pitch features extracted from the neutral reference database using KLD. KLD provides measure of distance between two distributions. It is approach to robustly estimate the differences between the distributions of two random variables.

IV. RESULTS

Below Table 2 gives the time domain features of input signal and Table 3 gives the features of pitch contour. Values of features of pitch contour as well as values of derivate of pitch curve changes with emotional modulation.

Feature	Neutral	Happy	Angry
Mean	-0.00065729	-0.00079131	-0.00072557
Standard deviation	0.081136	0.10628	0.07917
Variance	0.006583	0.011296	0.0062679
Zero crossing	0.080526	0.10459	0.077847

Table 2: Time domain features of Input Neutral, happy and angry speech signal

	Neutral	Happy	Angry
Maximum F0	333.3333	363.6364	347.8261
Minimum F0	212.9398	224.2913	228.1448
Maximum DF0	290.9091	363.6364	320
Mean DF0	0.53989	-0.27076	-0.32468
Standard deviation DF0	67.3122	61.0276	63.7329
Range	290.3692	363.9071	320.3247
Kurtosis	5.2283	3.7977	5.5485
Skewness	-1.8881	-1.5092	-1.9185
Smoothness	-22.1618	-35.3783	-27.0541
Moment	22.0934	35.2583	26.9578
Entropy	-0.69272	-0.69311	-0.70315

Table 3: Frequency domain features of Input neutral, happy and angry speech signal

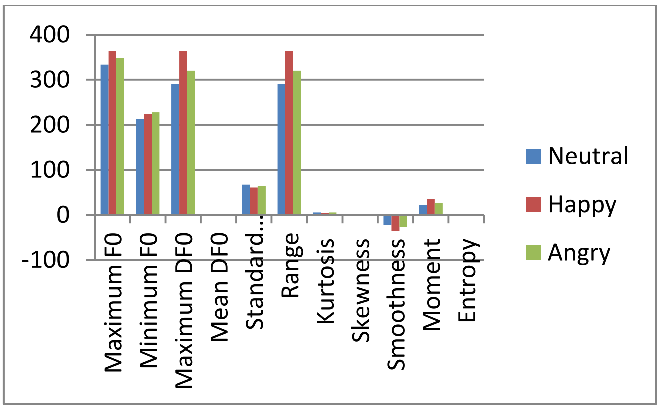
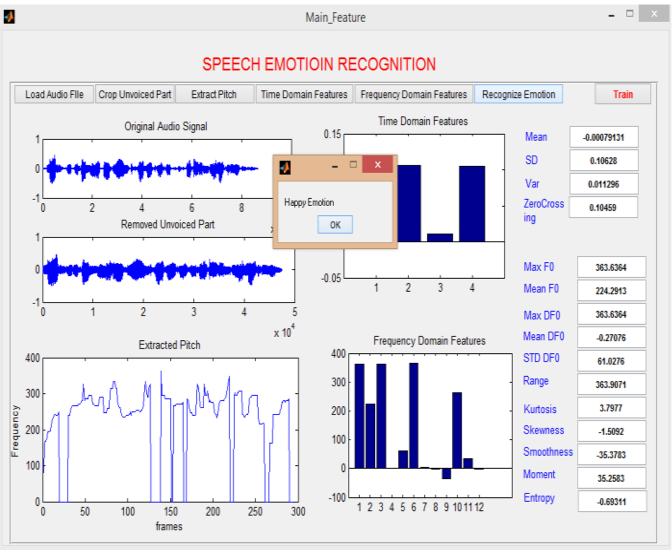


Fig 8: Graphical representation of Table 3



Happy Emotion

Mean	-0.00079131
SD	0.10628
Var	0.011296
ZeroCrossing	0.10459
Max F0	363.6364
Mean F0	224.2913
Max DF0	363.6364
Mean DF0	-0.27076
STD DF0	61.0276
Range	363.9071
Kurtosis	3.7977
Skewness	-1.5092
Smoothness	-35.3783
Moment	35.2583
Entropy	-0.69311

V. CONCLUSION AND FUTURE SCOPE

From the results in section IV, we can say that the there is variation in values of F0 for different emotions. Values of all the feature of F0 in case of neutral signal are found to be less than the other two signals. We will use this result to build the reference neutral model to detect the emotion in speech signal using KLD.

In this paper, the database which we have used is taken from the professional actors. So the future scope of paper will be the analysis of natural emotional databases recorded from real life scenarios.

VI. REFERENCES

[1] Chul Min Lee, *Student Member, IEEE,* and Shrikanth S. Narayanan*, Senior Member, IEEE, "*Toward Detecting Emotions in Spoken Dialogs*" in* IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 2, MARCH 2005

[2] Carlos Busso*, Member, IEEE*, Sungbok Lee*, Member, IEEE*, and Shrikanth Narayanan*, Fellow, IEEE, "* Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection*" in* IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 4, MAY 2009

[3] Mohammed E. Hoque, Mohammed Yeasin, Max M. Louwerse, "Robust Recognition of Emotion from Speech"

[4] Daniel Neiberg, Kjell Elenius and Kornel Laskowski," Emotion Recognition in Spontaneous Speech Using GMMs"

[5] Chung-Hsien Wu, Senior Member, IEEE, and Wei-Bin Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels", in IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 2, NO. 1, JANUARY-MARCH 2011

[6] Stavros Ntalampiras and Nikos Fakotakis, "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition", in IEEE TRANSACTIONS ON AFFECTIVECOMPUTING, VOL.3, NO. 1, JANUARY-MARCH 2012

[7] Akshay S. Utane Dr. S.L.Nalbalwar, " Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model" in International Journal of Advanced Research in Computer Science and Software Engineering, April 2013

[8] Biswajit Nayak, Mitali Madhusmita, Debendra Ku Sahu, " Speech Emotion Recognition using Different Centred GMM" in International Journal of Advanced Research in Computer Science and Software Engineering, sep 2013

[9] Vidhyasaharan Sethu*, Eliathamby Ambikairajah and Julien Epps, "On the use of speech parameter contours for emotion recognition" in Sethu et al. EURASIP Journal on Audio, Speech, and Music Processing 2013

[10] Md. Touseef Sumer, "Salient Feature Extraction For Emotion Detection Using Modified Kullback Leibler Divergence" in Internationl Journal of Research in Engineering and Applied Science(IJREAS),Jan 2014

[11] Dr. Shaila D. Apte, "Speech and Audio Processing", book.