2015 International Conference on Computing Communication Control and Automation

Analysis of Speech Features for Emotion Detection : A review

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.

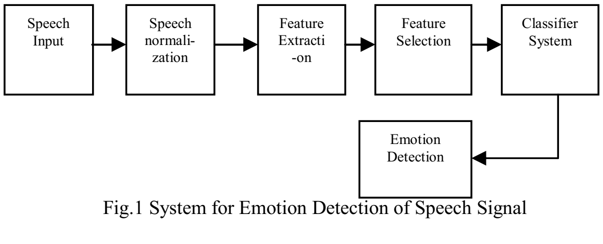

Fig.1 System for Emotion Detection of Speech Signal

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.135

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.

III. REVIEW OF PAPERS

***A. Toward Detecting Emotions in Spoken Dialogs*[1]**

Three sources of information are combined acoustic, lexical, and discourse for emotion detection. To capture emotion information at the language level, an information-theoretic notion of emotional salience is introduced. Optimization of the acoustic correlates of emotion with respect to classification error was accomplished by investigating different feature sets obtained from feature selection, followed by principal component analysis. The results in this paper show that, the best results are obtained when acoustic and language information are combined**.** And also combining all the information improves emotion classification by 40.7% for males and 36.4% for females.

B. *Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection* [2]

This paper has considered pitch features or features of fundamental frequency for the emotion detection of speech. The mean, standard deviation, range, minimum, maximum, median, lower quartile, upper quartile, interquartile range, kurtosis, skewness, slope, curvature and inflexion all these statistics of pitch contour and derivative of pitch contour are taken for emotion detection. Then these statistics are grouped into sentence level and voiced level features, which are further used for emotion detection. After that these characteristics of emotional speech is compared with the characteristics of neutral speech by using KLD. Nested logistic regression models are to quantify the emotionally discriminative power of pitch feature. The results indicate that the pitch contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the pitch shape. Also analyzing the pitch statistics at the sentence level is found to be more accurate and robust than analyzing the pitch statistics for voiced regions.

C. *Robust Recognition of Emotion from Speech* [3]

Prosodic features like pitch, energy, formants and acoustic features are used to extract the intonation patterns and correlates of emotion from speech samples for the emotion detection. To improve the performance features were used on word level emotional utterances. Here the classifiers from WEKA tools are used for emotion detection.

D. *Emotion Recognition in Spontaneous Speech using GMMs* [4]

Author has used MFCC, MFCC-low and variant features for the emotion detection. MFCCs are extracted using pre-emphasized audio, using 25.6ms Hamming window at every 10ms. For each frame 24,FFT based mel warped logarithmic filter bank are placed in 300 to 3400Hz. For MFCC-low filter bank is placed in 20-300Hz. Variant features such as pitch and derivative are used for emotion detection of speech signal. GMM is used as classifier for emotion detection.

E. *Emotion Recognition of Affective Speech based on Multiple Classifiers using Acoustic-Prosodic Information and Semantic Labels* [5]

In this paper, acoustic prosodic information and semantic labels are used for the emotion detection of speech. For acoustic prosodic information detection, acoustic and prosodic features like spectrum, formant and pitch are extracted from input speech. For this three types of base level classifier models GMM, SVM (support vector machine), and MLP (multilayer perceptron) are used and lastly the Meta Decision Tree (MDT) is used for classifier fusion. For SL based detection semantic labels derived from an existing Chinese knowledge base, HowNet are used to extract Emotion Association Rules (EAR) from detected word sequence of speech. Maximum entropy model is then used to explain the relationship between emotional states and EARs for emotion detection.

F. *Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition* [6]

Baseline set and feature based on multiresolution analysis, these two feature sets of heterogeneous domain are used in this paper. The first set includes mel filterbank, pitch, and harmonic to noise ratio and second set includes wavelet packets. After extracting these features, feature integration methods like short – term statistics, spectral moments and autoregressive model are used. Then emotion of the speech is detected by doing the fusion of feature level fusion, fusion of log likelihoods which are produced by temporally integrated feature sets and fusion of temporal integration method.

G. *Emotion Recognition through Speech using Gaussian Mixture Model and Hidden Markov Model* [7]

In this paper, mel frequency cepstrum coefficient (MFCC), linear predictive cepstral coefficients (LPCC) and energy features are used for the emotion detection of speech. Here GMM and HMM is used as classifier for emotion detection of speech. It is observed that both the classifier methods provides relatively similar accuracy. The efficiency of emotion recognition system highly depends on database selection, so it is very necessary to select proper database.

H. *Speech Emotion Recognition using Different Centred GMM* [8]

Here mel frequency Cepstral Coefficients (MFCC) features are used for emotion detection. Here eight different speakers and IITKGP-SEHSC emotional speech corpora are used for emotion detection. And classification is carried out by using GMM. It is observed in this paper that, when the number of centers of centered GMM increases the emotion recognition performance increases.

I. *On the Use of Speech Parameter Contours for Emotion Recognition* [9]

Generally frame based features are used for emotion detection, in this paper temporal contours of parameters like glottal source parameter which is extracted from three component model of speech production is use as a feature for automatic emotion detection of speech. Then automatic classification system for emotion detection is used with front end and back end with HMM in back end.

J. Salient Feature Extraction for Emotion Detection using Modified KullBack Leibler Divergence [10]

Formant frequencies are used for the emotion detection. Here they had taken three formant frequencies f1,f2,f3 and for different vowels the range of f1 lies between 270 to 730Hz, f2 and f3 lies between 840 to 2290HZ and 1690 to 3010Hz respectively. These frequencies are important for analysis of emotion of person. Linear predictive coding technique has been used for estimation of formant frequencies. With the formant frequencies pitch features are also used for detection of emotion. KLD and GMM is used for further process of emotion detection.

IV. CONCLUSION

In this paper, we discussed the basic process for emotional speech detection and the most important part of emotion detection of speech is feature extraction. Different papers have been reviewed considering different features of emotional speech which is to be extracted for emotion detection. Above review states that the prosody features such as pitch, formant, intensity, energy and spectral features such as LPC and MFCC are most commonly used feature. The pitch statistics such as the mean/median, maximum/upper quartile, minimum/lower quartile, and range/interquartile range, are the most emotionally salient pitch features for emotion detection of speech signal while features describing pitch contour shape such as slope, curvature and inflexion are least emotionally prominent. Study shows that by using different features of speech, accuracy upto 85% is achieved in emotion detection. The methods used for emotion detection are KLD or incomplete sparse least square regression while GMM or HMM are used as classifier for emotion detection.

V. 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.