IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
ISSN: 2278-2834, ISBN: 2278-8735. Volume 4, Issue 4 (Jan. - Feb. 2013), PP 34-37
www.iosrjournals.org

Speech Emotion Recognition: A Review

Dipti D. Joshi[1,] Prof. M. B. Zalte[2]
[1, 2] *(EXTC Department, K.J. Somaiya College of Engineering, University of Mumbai, India)*

Abstract: *Field of emotional content recognition of speech signals has been gaining increasing interest during recent years. Several emotion recognition systems have been constructed by different researchers for recognition of human emotions in spoken utterances. This paper describes speech emotion recognition based on the previous technologies which uses different methods of feature extraction and different classifiers for the emotion recognition are reviewed. The database for the speech emotion recognition system is the emotional speech samples and the features extracted from these speech samples are the energy, pitch, linear prediction cepstrum coefficient (LPCC), Mel frequency cepstrum coefficient (MFCC). Different wavelet decomposition structures can also used for feature vector extraction. The classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, fear, neutral state, etc. The classification performance is based on extracted features. Conclusions drawn from performance and limitations of speech emotion recognition system based on different methodologies are also discussed.*

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 [2]). 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.

Paper [3] computes speech features that represent entire utterance using the average values of various features which belong to the time or to the frequency domain. For classifying the unknown speech samples vector quantization, ANN and GMM methods are used.

One of researcher constructed an emotion recognition system that uses the combination of both statistic and temporal features. GMM and HMM likelihoods are integrated which is then fed to a Bayesian and an MLP classifier. Their feature set was comprised of the F0, feature contours of log energy, and syllable duration.

A different direction is followed in paper [4], where a variety of descriptors (MFCC, prosodic, speech quality and articulatory) is computed on frame as well as on turn level. During the frame level analysis they used a GMM classifier, while an SVM classifier was used with for turn level. Database used were the BERLIN Emotional Speech database, as well as the Speech Under Simulated and Actual Stress (SUSAS), They found average recognition rates respectively 89.9 and 83.8 percent.

Last but not least, an interesting approach is followed in [5], where sentence-level emotion recognition is investigated. Information from subsentence segments was used for sentence level decisions. Segment level emotion classifier used to generate predictions for segments within a sentence. Second component combines the predictions from these segments to obtain a sentence level decision. Different segment units (words, phrases, time-based segments) and different decision combination methods (majority vote, average of probabilities, and a Gaussian Mixture Model (GMM)) were evaluated. Experimental results show that proposed method significantly outperforms the standard sentence-based classification approach. In addition, they found that time-based segments achieves the best performance, and thus no speech recognition or alignment is needed when using this method, which is important to develop language independent emotion recognition systems.

The paper is organized as follows: section two describes the overall structure of the speech emotion recognition system. The different features extracted in the feature extraction and the details about the feature selection are discussed in section three. Different classification schemes which could be use in the speech emotion recognition 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.

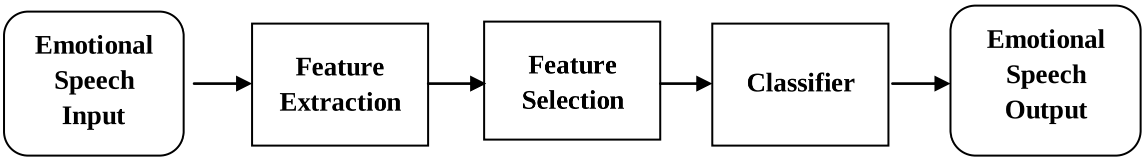

Fig.1 Speech Emotion Recognition System

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 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 1 still 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 100 ms. 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. [2]

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 [8].

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. [9] 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. [2]

IV. Classifier Selection

For choice of classifier there is no fixed criterion. Selection of classifier depends on the geometry of the input feature vector. Some classifiers are more efficient with certain type of class distributions, and some are better at dealing with many irrelevant features or with structured feature sets. Performance comparison of classifiers can be done on the same large and representative database. Most advanced researches on a speaker-independent mode achieve recognition rates from 55% to 95%, whereas humans could hardly reach emotion recognition rates of about 60% from unknown speakers [4].

Various Classifiers used by researchers are K-nearest Neighbors (KNN)[1], hidden Markov model (HMM), Gaussian mixtures Model (GMM), support vector machine (SVM) and artificial neural net (ANN). HMM has been studied long time by researchers for speech emotion recognition, as has advantage on dynamic time warping capability. Moreover, it has been proved useful in dealing with the statistical and sequential aspects of the speech signal for emotion recognition [10]. However, the classify property of HMM is not satisfactory.

Gaussian mixture model allows training the desired data set from the databases. GMM are known to capture distribution of data point from the input feature space, therefore GMM are suitable for developing emotion recognition model when large number of feature vector is available. Given a set of inputs, GMM refines the weights of each distribution through expectation-maximization algorithm. GMMs are suitable for developing emotion recognition models using spectral features, as the decision regarding the emotion category of the feature vector is taken based on its probability of coming from the feature vectors of the specific model. Gaussian Mixture Models (GMMs) are among the most statistically matured methods for clustering and for density estimation. They model the probability density function of observed data points using a multivariate Gaussian mixture density. [11]

Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). They are known to be more effective in modeling nonlinear mappings. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. Almost all ANNs can be categorized into three main basic types: MLP, recurrent neural networks (RNN), and radial basis functions (RBF) network. The classification accuracy of ANN is fairly low compared to other classifiers. The ANN based classifiers may achieve a correct classification rate of 51.19% in speaker dependent recognition, and that of 52.87% for speaker independent recognition. [2]

One of the important classifiers is the support vector machine. SVM classifiers are mainly based on the use of kernel functions to nonlinearly map the original features to a high dimensional space where data can be well classified using a linear classifier. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers [8]. SVM has shown to have better generalization performance than traditional techniques in solving classification problems. The accuracy of the SVM for the speaker independent and dependent classification are 75% and above 80% respectively [1, 8].

There are many other classifiers used for speech emotion recognition such as k-NN classifiers fuzzy classifiers, and decision trees. The GMM and the HMM, are the most used ones for speech emotion recognition. The performance of many of them is not significantly different from the above mentioned classification techniques. [2]

V. Conclusion

In this paper, most recent work done in the field of Speech Emotion Recognition is discussed. Most used methods of feature extraction and several classifier performances are reviewed. Success of emotion recognition is dependent on appropriate feature extraction as well as proper classifier selection from the sample emotional speech. It can be seen that Integration of various features can give the better recognition rate. Classifier performance is need to be increased for recognition of speaker independent systems. The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can even provide emotion recognition of real time speech.

References

[1] Jia Rong, Gang Li , Yi-Ping Phoebe Chen "*Acoustic feature selection for automatic emotion recognition from speech*", Elsevier , Information Processing and Management volume 45 (2009)

[2] M. E. Ayadi, M. S. Kamel, F. Karray, "*Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases*", Pattern Recognition 44, PP.572-587, 2011.

[3] Y. Li and Y. Zhao, "*Recognizing Emotions in Speech Using Short- Term and Long-Term Features*," Proc. Int'l Conf. Spoken Language Processing, pp. 2255-2258, 1998.

[4] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll,"*Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech*," Proc. Int'l Conf. Spoken Language Processing, pp. 2225-2228, 2007.

[5] je hun jeon, rui xia, yang liu," *sentence level emotion recognition based on decisions From subsentence segments*", ICASSP 2011, 978-1-4577-0539-©2011 IEEE

[6] Vogt, T. Andre, E. "*comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition*", proc. ICME 2005, Amsterdam, Netherlands 2005

[7] J. Sidorova, " *Speech Emotion Recognition*" DEA report Universitat Pompeu Fabra, July 4, 2007

[8] P.Shen, Z. Changjun, X. Chen, "*Automatic Speech Emotion Recognition Using Support Vector Machine*", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011

[9] S. Mallat, "*A wavelet tour of signal processing"*, NewYork, Academic Press, 1999

[10] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "*Speech Emotion Recognition Using Hidden Markov Model*", Eurospeech, 2001

[11] Neiberg, D., elenius K., Laskowski K. "*Emotion recognition in spontaneous speech using GMM*". Proc. INTERSPEECH'2006, Pittsburgh, 2006