Journal of Science and Technology (JST)
Volume 3, Issue 2, March - April 2018, PP 18-20
www.jst.org.in *ISSN: 2456 - 5660*

Speech Emotion Recognition Methods

B.Kishore Babu [1], T Annamani[2]

[1]*(ECE Department, Samskruti college of Engineering and Technology,Hyderabad)*
[2]*(ECE Department, Samskruti college of Engineering and Technology,Hyderabad)*
[1]annadimpu@gmail.com [2]annadimpu@gmail.com

Abstract : *Speech Emotion Recognition is a current topic of research since it has wide range of applications. Speech Emotion Recognition is a vital part of affective human interaction and has become a new challenge to speech processing. The work presented in this paper focus on study of various speech emotions recognition methods.*

Keywords - *SER System; prosodic and spectral features; SVM; HMM; KNN; Ada Boost algorithm*

I. INTRODUCTION

Emotions play an extremely important role in human mental life. It is a medium of expression of one's perspective or his mental state to others. Speech emotion Recognition (SER) can be defined as the extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions- including Neutral, Anger, Surprise, Fear, Happiness, Sadness which any intelligent system with finite computational resources can be trained to identify or synthesize as required. The importance of automatically recognizing emotions in human speech has grown with increasing role of spoken language interfaces in the field of human machine interaction to make the human machine interface more efficient. It can also be used for in-car board system where information of the mental state of the driver maybe provided to initiate his/her safety. In automatic remote call center, it is used to timely detect customers‟ dissatisfaction. In E-learning field, identifying students‟ emotion timely and making appropriate treatment can enhance the quality of teaching [1]. Both spectral and prosodic features can be used for speech emotion recognition because both of these features contain the emotional information. Linear predictive cepstrum coefficients (LPCC) and Mel-frequency cepstrum coefficients (MFCC) are some of the spectral features. Fundamental frequency, loudness, pitch and speech intensity and glottal parameters are the prosodic features used to model the different emotions [2]. Different classifiers available for SER are Kernel Regression and k-nearest neighbors (KNN), Support Vector Machines (SVM), Maximum Likelihood Bayesian Classifier (MLC) [3], Hidden Markov Model (HMM) Artificial Neural Network (ANN) [4]. The potential features are extracted from each utterance for the computational mapping between emotions and speech patterns. The selected features are then used for training and testing by using any classifier method to recognize the emotions.

II. OVERVIEW OF SPEECHRECOGNITION

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified. A speech recognition system consists of five blocks: - Feature extraction, Acoustic modeling, Pronunciation modeling,Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the soundwaves. These sound waves are then captured by a microphone and converted into electrical signals. These electrical signals are then converted into digital form to make them understandable by the speech-system. Speech signal is then converted into discrete sequence of feature vectors, which is assumed to contain only the relevant information about given utterance that is important for its correct recognition. An important property of feature extraction is the suppression of information irrelevant for correct classification such as information about speaker (e.g. fundamental frequency) and information about transmission channel (e.g. characteristic of a microphone). Finally recognition component finds the best match in the knowledge base, for the incoming feature vectors. Sometimes, however the information conveyed by these feature vectors may be correlated and less discriminative which may slow down the further processing. Feature extraction methods like Mel frequency cepstral coefficient (MFCC) provides some way to get uncorrelated vectors by means of discrete cosine transforms (DCT).

III. SPEECH EMOTION ECOGNITION 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 29 recognition methods to identify emotional states. Like typical pattern recognition systems, our speech emotion recognition system contains four main modules: speech input, feature extraction, SVM based classification, and emotion output [5]. The general architecture for SER system has three steps shown in Fig. 1 [6]: i. A speech processing system extracts some appropriate quantities from signal, such as pitch or energy, ii. These quantities are summarized into reduced set of features, iii. A classifier learns in a supervised manner with example data how to associate the features to the emotions. Fig. 1 A basic outline of the Speech Emotion Recognition System [6]

IV. FEATURE EXTRACTION

Feature extraction is based on partitioning speech into small intervals known as frames. To select suitable features which are carrying information about emotions from speech signal is an important step in SER system. There are two types of features: prosodic features including energy, pitch and spectral features including MFCC, MEDC, LPCC.

a. Energy and related features Energy is the basic and most important feature in speech signal. To obtain the statistics of energy feature, we use short-term function to extract the value of energy in each speech frame. Then 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 [7].

b. Pitch and related features The vibration rate of vocal is called the fundamental frequency F0 or pitch frequency. The pitch signal has information about emotion, because it depends on the tension of the vocal folds and the sub glottal air pressure, so the mean value of pitch, variance, variation range and the contour is different in seven basic emotional statuses [5]. Physical quantities (pitch, energy……) F1, F2, F3………… Fn Speech Processing System Feature Extractor Classifier Emotional State Emotional Speech .The following statistics are calculated from the pitch and used in pitch feature vector [8]: • Mean, Median, Variance, Maximum, Minimum (for the pitch feature vector and its derivative) • Average energies of voiced and unvoiced speech • Speaking rate (inverse of the average length of the voiced part of utterance).

c. MFCC and MEDC features Mel-Frequency Cepstrum coefficients is the most important feature of speech with simple calculation, good ability of distinction, anti-noise. MFCC in the low frequency region has a good frequency resolution, and the robustness to noise is also very good. MEDC extraction process is similar with MFCC. The only one difference in extraction process is that the MEDC is taking logarithmic mean of energies after Mel Filter bank and Frequency wrapping, while the MFCC is taking logarithmic after Mel Filter bank and Frequency wrapping. After that, we also compute 1st and 2nd difference about this feature [5].

d. Linear Prediction Cepstrum Coefficients LPCC embodies the characteristics of particular channel of speech, and the same 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) [5].

V. CLASSIFIERS

For each extracted features of emotional speech classification algorithm is applied on different set of inputs. Different classifiers are discussed below:

A. Support Vector Machine (SVM) SVM, a binary classifier is a simple and efficient computation of machine learning algorithms, and is widely used for pattern recognition and classification problems, and under the conditions of limited training data, it can have a very good classification performance compared to other classifiers [9]. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. Therefore non-linear problems can be solved by doing this transformation.

B. Hidden Markov Model (HMM) The HMM consist of the first order markov chain whose states are hidden from the observer therefore the internal behavior of the model remains hidden. The hidden states of the model capture the temporal structure of the data. Hidden Markov Models are statistical models that describe the sequences of events. HMM is having the advantage that the temporal dynamics of the speech features can be trapped due to the presence of the state transition matrix. During clustering, a speech signal is taken and the probability for each speech signal provided to the model is calculated. An output of the classifier is based on the maximum probability that the model has been generated this signal [10]. For the emotion recognition using HMM, first the database is sort out according to the mode of classification and then the features from input waveform are extracted. These features are then added to database. The transition matrix and emission matrix has been made according to the modes, which generates the random sequence of states and emissions from the model. Final is estimating the state sequence probability by using Viterbi algorithm [11].

C. K Nearest Neighbor (KNN) A more general version of the nearest neighbor technique bases the classification of an unknown sample on the “votes” of K of its nearest neighbor rather than on only it‟s on single nearest neighbor. Among the various methods of supervised statistical pattern recognition, the Nearest Neighbor is the most traditional one, it does not consider a priori assumptions about the distributions from which the training examples are drawn. It involves a training set of all cases. A new sample is classified by calculating the distance to the nearest training case, the sign of that point then determines the classification of the sample. Larger K values help reduce the effects of noisy points within the training data set, and the choice of K is often performed through cross validation [12]

D. AdaBoost Algorithm AdaBoost algorithm is an adaptive classifier which iteratively builds a strong classifier from a weak classifier. In each iteration, the weak classifier is used to classify the data points of training data set. Initially all the data points are given equal weights, but after each iteration, the weight of incorrectly classified data points increases so that the classifier in next iteration focuses more on them. This results in decrease of the global error of the classifier and hence builds a stronger classifier. AdaBoost algorithm is also used as a feature selector for training SVMs [13].

CONCLUSION

In this study, the overview of different SER methods are discussed for extracting audio features from speech sample, various classifier algorithms are explained briefly. Speech Emotion Recognition has a promising future and its accuracy depends upon the combination of features extracted, type of classification algorithm used and the correct of emotional speech database. This study aims to provide a simple guide to the researcher for those carried out their research study in the speech emotion recognition systems correct of emotional speech database. This study aims to provide a simple guide to the researcher for those carried out their research study in the speech emotion recognition systems.

REFERENCES

[1] Ayadi M. E., Kamel M. S. and Karray F., „Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases', Pattern Recognition, 44(16), 572-587, 2011.

[2] Zhou y., Sun Y., Zhang J, Yan Y., "Speech Emotion Recognition using Both Spectral and Prosodic Features", IEEE, 23(5), 545-549, 2009.

[3] Yashpalsing Chavhan, M. L. Dhore, Pallavi Yesaware, “Speech Emotion Recognition Using Support Vector Machine”, International Journal of Computer Applications, vol.1, pp.6-9,February 2010.

[4] Xia Mao, Lijiang Chen, Liqin Fu, “Multi-level Speech Emotion Recognition Based on HMM and ANN”, 2009 WRI World Congress, Computer Science and Information Engineering, pp.225-229, March 2009.

[5] Yixiong Pan, Peipei Shen and Liping Shen, “Speech Emotion Recognition Using Support Vector Machine”, International Journal of Smart Home, Vol. 6, No. 2, April, 2012.

[6][Online].Available: http://crteknologies.fr/projects/emospeech/

[7] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification”, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

[8] F.Yu, E.Chang, Y.Xu, H.Shum, “Emotion detection from speech to enrich multimedia content”, Lecture Notes In Computer Science,Vol.2195, 550-557, 2001.

[9] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, P.-J. Li, “Mandarin emotional speech recognition based on SVM and NN”, Proceedings of the 18th International Conference on Pattern Recognition (ICPR?06), vol. 1, pp. 1096-1100,September 2006.

[10] B. Schuller, G. Rigoll, M. Lang, “Hidden Markov model-based speech emotion recognition”, Proceedings of the IEEE ICASSP Conference on Acoustics,Speech and Signal Processing, vol.2, pp. 1-4, April 2003.