www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 4 Issue - 12 December, 2015 Page No. 15364-15367

Speech Classification Using Mfcc, Power Spectrum And Neural Networkand.

Mr. Ashish Gadewar, Mahesh Navale

[1]Dept of E&TC , G.S.M.C.O.E. Balewadi, Pune
&
[2]Astt. Prof. Dept of E&TC , S.K.N.C.O.E, Pune

Abstract:Speech signal carries rich emotional information except semantic information. Speech and emotion recognition improve the quality of human computer interaction and allow more easy to use interfaces for every level of user in software applications. Common emotion namely exclamatory, neutral and question mark were discussed and recognize through a propose frame work which combines of Mel Frequency Cepstrum Coefficients (MFCC) and Power spectrum are used for feature extraction and back propagation neural network are used for recognition of the emotional speech signals. This further will be used for transplanting that emotion in synthetic speech so that output quality of synthesis is improved.

Keywords – Back propagation neural network, Mel frequency Cepstrum Coefficient, Power spectrum.

I. INTRODUCTION

Speech and emotion recognition improve the quality of human computer interaction and allow more easy to use interfaces for every level of user in software applications. This paper is all about the emotion recognition system to classify the voice signals for emotion recognition. Speech is one of the oldest tools humans use for interaction among each other. It is therefore one of the most natural ways to interact with the computers as well. Although speech recognition is now good enough to allow speech to text engines, emotion recognition can increase the overall efficiency of interaction and may provide everyone a more comfortable user interface. It is often trivial for humans to get the emotion of the speaker and adjust their behavior accordingly. Emotion recognition will give the programmer a chance to develop an artificial intelligence that can meet the speaker's feelings that can be used in many scenarios from computer games to virtual sales-programs.

Speech is one of the natural forms of communication. Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. Voice may be a normal speech or produced along with an emotion like happy, sad, fear, angry etc. Voices differ for men and women in several aspects such as speaking pitch, pitch range, the space between the vocal folds, formant frequency,

and the incidence of voice problems. Females speak with a higher fundamental frequency (voice pitch) when compared to males [1]. The higher pitch in women compared to men means the vocal folds vibrate or come together almost twice as many times per second in females than in males. The other differences found in voice quality are caused by the way the vocal folds vibrate between male and female. Usually males speak creakier than female and females speak breathier than males.

Two base emotions, Question mark/Exclamatory and neutral are taken into account. Various speech sets that belong to these emotion groups are taken and used for training and

testing. The ERNN will be capable of distinguishing these test samples. Neural networks are chosen for the solution because a basic formula cannot be devised for the problem. The neural networks are also quick to respond which is a requirement as the emotion should be determined almost instantly. The training takes a long time but is irrelevant as the training will be mostly off-line and on-line both.

II. EMOTION RECOGNITION SYSTEM

Emotion recognition is not a new topic and both research and applications exist using various methods most of which require extracting certain features from the speech [1]. A common problem is determining emotion from noisy speech, which complicates the extraction of the emotion because of the background noise [2]. To extract the emotion signatures inherent to voice signals, the back propagation-learning algorithm [3,4,5,6] is used to design the emotion recognition neural network (ERNN).The proposed system is shown in figure 2.1. The output of it is normalized and is presented to a three layer, fully interconnected neural network for classification. The output layer of the neural network is inputted by the weighted sum of outputs of the hidden and bias nodes in the hidden layer. A set of desired output values is then compared to the estimated outputs of the neural network for every set of input values of the power spectrum of the voice signals. The weights are appropriately updated by back propagating the gradient of the output error through the entire neural network.

The experimental data, used for both training and testing the ERNN, is recorded in WAVE format. Each experimental data segment is composed of number of data points. The input signal to the feature extraction block is created using a sliding window and the step between two successive windows is number of samples. The first set is taken from the beginning of the voice data. The second set is number of samples to the right of the first set, and this is repeated until the window covers the entire number of samples of the voice signal. The sampled value of the voice signal, *T* is the time sampling interval, which span the entire frequency range is taken into consideration as an input to the neural network for signal.

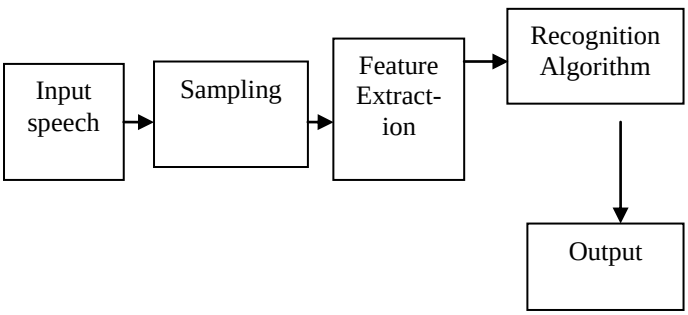

Fig 1 proposed system of speech classification

Hence, the neural network needs little number of inputs and 2 neurons in its output layer to classify the voice signals. The hidden layer has number of neurons. This number was picked through experimentation and experience. If the network has trouble classifying, then additional neurons can be added to the hidden layer.

III. FEATURE EXTRACTION

Any emotion from the speaker's speech is represented by the large number of parameters which is contained in the speech and the changes in these parameters will result in corresponding change in emotions. Therefore an extraction of these speech features which represents emotions is an important factor in speech emotion recognition system [8]. The speech features can be divide into two main categories that is long term and short term features. The region of analysis of the speech signal used for the feature extraction is an important issue which is to be considering in the feature extraction. The speech signal is divided into the small intervals which are referred as a frame [8]. The prosodic features are known as the primary indicator of the speakers emotional states. Research on emotion of speech indicates that pitch, energy, duration, formant,Power,Mel frequency cepstrum coefficient (MFCC), and linear prediction cepstrum coefficient (LPCC) are the important features [5, 6]. With the different emotional state, corresponding changes occurs in the speak rate, pitch, energy, and spectrum.

In feature extraction all of the basic speech feature extracted may not be helpful and essential for speech emotion recognition system. If all the extracted features gives as an input to the classifier this would not guarantee the best system performance which shows that there is a need to remove such a unusefull features from the base features. Therefore there is a need of systematic feature selection to reduce these features. So that for this system we have used only two feature and that are MFCC and Power specrum.

IV. RECOGNITION ALGORITHM

In the speech emotion recognition system after calculation of the features, the best features are provided to the classifier. A classifier recognizes the emotion in the speakers speech utterance. Various types of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbors (KNN),

Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others.

Only when the global features are extracted from the training utterances, Gaussian Mixture Model is more suitable for speech emotion recognition. All the training and testing equations are based on the supposition that all vectors are independent therefore GMM cannot form temporal structure of the training data. For the best features a maximum accuracy of 78.77% could be achieved using GMM. In speaker independent recognition typical performance obtained of 75%, and that of 89.12% for speaker dependent recognition using GMM [10].

In speech recognition system like isolated word recognition and speech emotion recognition, hidden markov model is generally used; the main reason is its physical relation with the speech signals production mechanism. In speech emotion recognition system, HMM has achieved great success for modeling temporal information in the speech spectrum. The HMM is doubly stochastic process consist of first order markov chain whose states are buried from the observer [10]. For speech emotion recognition typically a single HMM is trained for each emotion and an unknown sample is classified according to the model which illustrate the derived feature sequence best [11]. HMM has the important advantage that the temporal dynamics of speech features can be caught second accessibility of the well established procedure for optimizing the recognition framework. The main problem in building the HMM based recognition model is the features selection process. Because it is not enough that features carries information about the emotional states, but it must fit the HMM structure as well. HMM provides better classification accuracies for speech emotion recognition as compared with the other classifiers [9]. HMM classifiers using prosody and formant features have considerably lower recall rates than that of the classifiers using spectral features [9]. The accuracy rate of the speech emotion recognition by using HMM classifier is observed as 46.12% for the speaker dependent in the previous study and that for the speaker independent it was 44.77% [1].

Transforming the original feature set to a high dimensional feature space by using the kernel function is the main thought behind the support vector machine (SVM) classifier, which leads to get optimum classification in this new feature space. The kernel functions like linear, polynomial, radial basis function (RBF) can be used in SVM model for large extent. In the main applications like pattern recognition and classification problems, SVM classifier are generally used, and because of that it is used in the speech emotion recognition system. SVM is having much better classification performance compared to other classifiers [10, 4]. The emotional states can be separated to huge margin by using SVM classifier. This margin is nothing but the width of the largest tube without any utterances, which can obtain around decision boundary. The support vectors can be known as the measurement vectors which define the boundaries of the margin. An original SVM classifier was designed only for two class problems, but it can be use for more classes. Because of the structural risk minimization oriented training SVM is having high generalization capability. The accuracy of the SVM for the speaker independent and dependent classification are 65% and above 80% respectively [10, 7].

Other classifier that is used for the emotion classification is an artificial neural network (ANN), which is used due to its ability to find nonlinear boundaries separating the emotional states. Out of the many types, Back propagation neural network is used most frequently in speech emotion recognition [5]. Multilayer perceptron layer neural networks are relatively common in speech emotion recognition as it is easy for implementation and it has well defined training algorithm [10].The ANN based classifiers may achieve a correct classification rate of 71.19% in speaker dependent recognition, and that of 72.87% for speaker independent recognition. So for this system as a classifier we choose ANN in which we use back propagation algorithm.

V. CONCLUSION

In this proposed work we will develop a neural network that is designed to classify the voice signals with the help of MFCC and Power Spectrum. In speech signal classifications many biological and emotional factors from speaker side are involved so it is difficult to mathematical model all those parameter which characterizes perfect emotion. So neural network approach is depends on better which

is probabilistic training. In future this neural network will work with near real time applications, like computer games. Role playing games can utilize the player's acting capabilities as another input that might improve gaming experience. It also can be extended to classifying genders for other emotional speeches such as fear, happy, angry, surprise and disgust. By including more number of subjects from various ethnic groups, a user independent and general system for gender classification from any type of speech – normal or emotional can be developed.In conventional TTS toning of speech is monotonic which looks very robotic, so if this system is interface with these TTS then it is helpful to enhance its working and effectiveness

REFERENCES

[1]Emotion Recognition Using Neural Networks MEHMET S. UNLUTURK, KAYA OGUZ, COSKUN ATAY Proceedings of the 10th WSEAS International Conference on NEURAL NETWORKS 2010

[2] DWT and MFCC Based Human Emotional Speech Classification Using LDA M Murugappan, Nurul Qasturi Idayu Baharuddin, Jerritta S 2012 International Conference on Biomedical Engineering (ICoBE),27-28 February 2012,Penang

[3] J. A. Freeman and D. M. Skapura, Neural Networks: Algorithms, Applications and Programming Techniques, Addison Wesley Publishing Company, 1991.

[4] T. Masters, Practical Neural Network Recipes in C++, Academic Press Publishing Company, 1993.

[5] A. Cichocki and R. Unbehauen, Neural Networks for Optimazing and Signal Processing, John Wiley & Sons Publishing Company, 1993.

[6] Werbos, P. The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting, John Wiley & Sons, New York, 1994.

[7] Gelfer M. P., Mikos, Victoria A, "The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based On Isolated Vowels," Journal of Voice 1 December 2005

[8] Lindasalwa Muda M. B., Elamvazuth I. , "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques " Computing, vol. 2, March 2010.

[9] Speech Emotion Recognition, Ashish B. Ingale, D. S. Chaudhari, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[10] 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.

[11] T. Vogt, E. Andre and J. Wagner, "Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization", LNCS 4868, PP.75-91, 2008.

AUTHORS

First name-Mr. Saurabh Padmawar,
ME,S.K.N.C.O.E Pune.
Second name-Prof. Mrs. Pallavi S. Deshpande , M-Tech, S.K.N.C.O.E Pune.