Speech Based Emotion Recognition Using MFCC and ANN

Ms. Swati Shinde[#1], Prof. Mrs. Swati Shilaskar[#2]

#1Vishwakarma Institute of Technology,Pune.
#2Vishwakarma Institute of Technology,Pune.

ABSTRACT

Speech is the most natural mode of communication. This work emphasizes on recognizing different emotions from speech signal. There are two major sections in this project namely feature extraction from speech signal and give this features as input to classifier to recognize emotions. Emotional states of speaker are considered as namely angry, happy, sad and neutral. The testing section classifies the training set of data with the help of back propagation algorithm. For the feature extraction of speech signal Mel Frequency Cepstrum Coefficients (MFCC) is used which gives a set of feature vectors of speech waveform. The Artificial Neural Networks (ANN) is selected as the classifier. The whole simulation is taken place in MATLAB environment. The proposed technique is providing promising results by giving the accuracy rate of 80-85 percent. The human capability to recognize the emotion from speech was also studied and compared with classifiers.

Key words: Emotional speech analysis, emotional speech recognition, Artificial Neural network, MFCC, ANN training, ANN testing, and confusion matrix.

INTRODUCTION

Humans have natural ability to recognize emotions through speech information but the task of emotion recognition for machine is very difficult since machine doesn't have sufficient intelligent to analyze emotions from speech.Speech is the most natural way to communicate for humans.If we want a similar easy and natural communication with machines, we can use speech as an interface.

The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker. Speech emotion recognition system aims at automatically identifying the emotional state of a human being from his or her voice. The goal of project is to fold. The first is to features extractions using Mel Frequency Ceptrum Coefficients (MFCC) from input speech signal and second goal is to use this extracted features to detect emotional speech detection using Artificial Neural Network (ANN). There are several types of classifiers are used for emotion recognition such as Hidden Markov Model (HMM), k-nearest neighbors (KNN), Artificial Neural Network (ANN), GMM super vector based SVM classifier , Gaussian Mixtures Model (GMM) and Support Vector Machine (SVM) [3].

In this project, the basic four emotional states such as angry, happy, sad and neutral state are classified using Artificial Neural Network (ANN) classifier. Accurate detection of emotion from speech has clear benefits for the design of more natural human- machine speech interfaces. In the field of human-computer interaction (HCI), emotion recognition from the computer is still a challenging issue, especially when the recognition is based solely on voice, which is the basic mean of human communication..

EMOTION RECOGNISION IN SPEECH

The block the diagram of speech emotion recognition system is illustrated in Fig 1. Several studies show a high correlation between some statistical measures of speech and the emotional state of the speaker [3], [4].

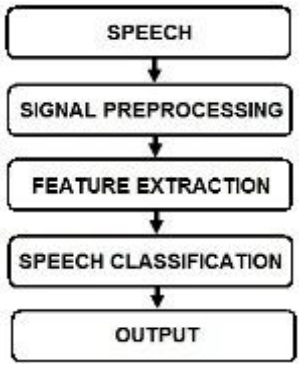


Fig.1 Block Diagram of Speech Emotion Recognition System.

Evaluation of Emotion recognition system through speech is mainly depending on the degree of accuracy of the database used. This proposed system consists of the emotional speech as input, feature extraction, classification of Emotional state using ANN classifier and detection of emotion as the output. The emotional speech input to the system may contain the collection of the acted speech data the real world speech data. After collection of the database containing short signal of emotional speech sample which was considered as the training samples, features such as spectral features (MFCC) were extracted from the speech signal. These feature values are provided to the Artificial Neural Network for training of the classifiers. The emotional speech samples which are recorded by high quality microphone are presented to the classifier as a test input. Then classifier classifies the test sample into one of the emotion from the above mentioned four emotions and gives output as recognized emotion [5].

EXTRACTION OF FEATURES

For the purpose of feature extraction, spectral analysis algorithm such as Mel-frequency Cepstral Coefficients is used. The Mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [2].

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency measured in Hz, a subjective pitch is measured on a scale called the Mel Scale. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above

the perceptual hearing threshold, is defined as 1000 Mels. Therefore, here we are using MFCC for spectral feature extraction. The calculation of the MFCC includes the following steps:

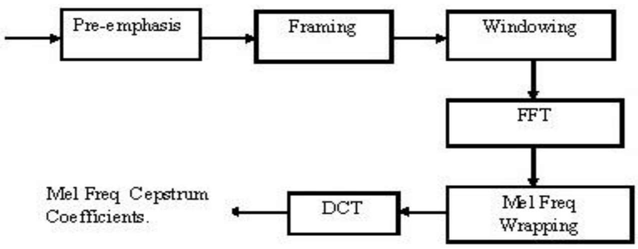


Fig.2 Block Diagram MFCC

1. Pre–emphasis: This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n]= X[n] * a X [n-1] \;\ldots\ldots\ldots\ldots (1)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

2. Framing: It is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms.The voice signal is divided into frames of N samples. Framing enables the non-stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40ms.Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

3. Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

The Hamming window equation is given as:

$$W(n) = 0.54 - 0.46\left(\frac{2\pi n}{N-1}\right) \qquad 0 \le n \le N-1 \ldots\ldots\ldots\ldots (2)$$

If the Hamming window is defined as W(n) $0 \le n \le N-1$

Where,

N = number of samples in each frame,

Y[n] = Output signal, X (n) = input signal ,W(n) = Hamming window,

then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n) \ldots\ldots\ldots\ldots (3)$$

4. FFT: FFT converts each frame of N samples from the time domain into the frequency domain. The Fourier Transform is to convert the convolution of the input pulse and the vocal tract impulse response in the time domain. This statement supports the equation below:

$$Y(w) = FFT[h(t) * x(t)] = H(w) * X(w) \ldots\ldots\ldots\ldots (4)$$

If $X(w), H(w)$ and $Y(w)$ are the Fourier Transform of $X(t), H(t)$ and $Y(t)$ respectively.

5. Mel-frequency wrapping: The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter‘s magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components.

$$F_{mel} = 2595 * \log 10\left(1 + \frac{f}{700}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

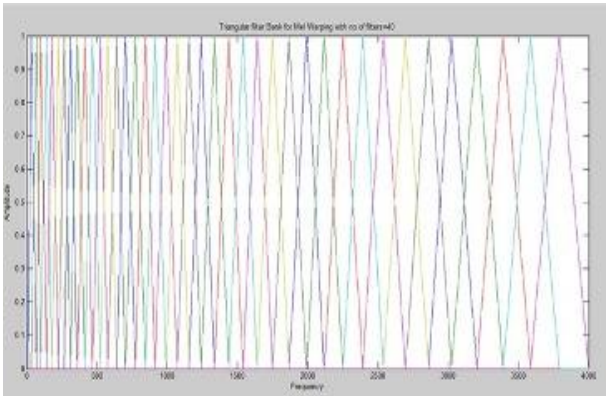


Fig. 3 Response of Mel spaced Filter bank.

6. Cepstrum (Discrete Cosine Transform): This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

METHODOLOGY

Database creation

We recorded emotional speech signal in .wav format. We considered English continuous sentence spoken in four emotions i.e. happy, angry, sad and neutral .The design of the database is specially oriented to speech synthesis purposes, but it can also provide a first approximation to emotional speech analysis and emotion recognition. The recordings were all done in silent rooms and with high quality microphones—a Panasonic 750 by using MATLAB software at sampling rate 8 kHz/16 bit and distance between mouth and microphone was adjusted nearly 20cm [4].

Evaluation of databases

Recorded file was given to 10 nonprofessional listeners to select correctly recorded emotion of one particular category. A total of 120 input signals, 30 per emotion ,one sentence spoken in six different emotions. Each listener had to choose between the four emotional styles considered. If the listener was not satisfied he could also mark a second one. Our results of the evaluation were quite satisfactory. More than an 84% of the first choices were correct and this figure almost reaches 92% if second choice.

ANN as a classifier

The most important aspect of emotion recognition system through speech is classification of emotions. The performance of the system is influenced by the accuracy of classification. On the basis of different features extracted from the utterances of emotional speech samples emotions can be classified by providing significant features to the classifier. We are using Artificial Neural Network (ANN) classifier for emotion recognition due to its superior ability to recognize overlapping patterns [7].Our network consists of an input layer, one hidden layer and an output layer. A general architecture of an artificial neural network is shown in Fig. 4.

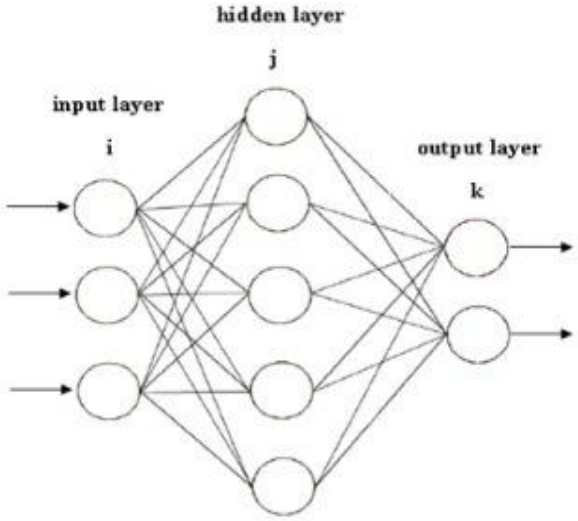


Fig.4: Architecture of an artificial neural network

Initially random values are initialized to weights so in training process we try to adjust these weights so as to minimize error between desired and target classes. Extracted features are given to the input layer consisting input nodes. To successfully classify the patterns into different emotion classes, the coefficients from feature extraction stage are applied to the input layer nodes. The outputs from input layer nodes are connected to the hidden layer, each hidden layer is connected to the output layer. we have used 83 % of total utterances for training purposes and another with 17% utterances for testing. MATLAB Neural Network toolbox is used for creation, training and simulation of the network.

Neural Network Training

After creation of the network, it can be trained for recognizing emotions by presenting training inputs and their corresponding targets. Type of training used is supervised training in which user has provided desired output for each input pattern. The mode used for training a network is batch mode. Batch mode means that the weights and biases of the network are updated only after the entire training set has been applied to the network [8].

Neural Network Testing

Once we have determined the best network, we need to evaluate its performance on a test set. In order to have an honest evaluation, the data in the test set must not occur in either the training set or the development set. i. e. has tested on unknown data, this is known as the testing phase or the validation phase. If the network does not meet the required validation standards, then further training of the network is required. The result analysis can be done from confusion matrix. The overall accuracy calculated from Table 1 is 85%.

CONCLUSION

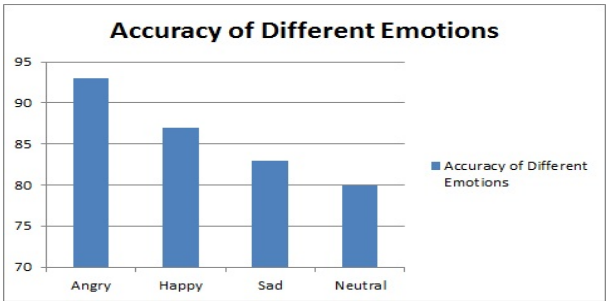
It presents framework for speech emotion recognition system based on Mel-frequency Cepstral coefficient (MFCC) and Neural Network and this implement an effective method for recognition of emotional state of speaker. Here Mel Frequency Cepstral Coefficients (MFCC) is used for extraction of features from speech signal and neural network is used to classify the given testing samples into different emotional states. By analyzing proposed approach, we can say that neural network can give better performance for recognizing emotions through speech signal. Combination of neural network and MFCC gives overall accuracy of about 85 percent. Further accuracy can be improved by increasing training data but correspondingly processing time increases.

RESULTS

Table 1: Confusion matrix of the emotional database

Confusion Matrix for Accuracy calculation		Predicted Emotions			
		Angry	Happy	Sad	Neutral
Actual Emotions	Angry	26	3	1	0
	Happy	3	24	2	1
	Sad	0	2	25	3
	Neutral	1	1	2	26

Graph 1: Percentage accuracy of different emotions



REFERENCES

[1] L Rabinar, B H Juang, B Yegnanrayana, Fundamental of Speech Recognition, Pearson 2012.

[2] Rahul b. Lanjewar, D. S. Chaudhari,Comparative analysis of speech emotion recognition system using Different classifiers on Berlin emotional speech database, ,*International Journal of Electrical and Electronics Engineering Research*, Vol. 3, Issue 5, page no:145-156,Dec 2013.

[3] JiaRong, Gang Li, Yi-Ping Phoebe Chen,Acoustic feature selection for automatic emotion recognition from speech,*Journal of Information Processing and Management*, Vol.4,page no:315–328,2009.

[4] Dipti D. Joshi, M. B. Zalte, Recognition of emotions from Marathi speech using mfcc and dwt algorithms,(ENTC Department K .J. Somaiya College of Engineering, University Mumbai, India) ISSN (Print): 2278-5140, Volume-2, Issue – 2,2013.

[5] JagvirKaur, AbhilashSharma,Speech emotion-speaker recognition using Mfcc and neural network, *Global Journal of Advanced Engineering Technologies*, Vol3, Issue 3- ISSN: 2277-6370,2014.

[6] JagvirKaurAbhilash Sharma Kaur et al.,Emotion Detection Independent of User Using MFCC Feature Extraction,*International Journal of Advanced Research in Computer Science and Software Engineering* 4(6), pp. 230-234,June – 2014.

[7] Jin Yu, Artificial Neural Network Assignment (s0105853) MAI-ECS,2004.

[8] Bhagyashree Kale, AnandKakade, Speech Emotion Recognition through ANN, *International Journal of Research in Management, Science & Technology* (E-ISSN: 2321-3264) Vol. 2, No. 1,April 2014.