Classification of speech for Clinical Data using Artificial Neural Network

C.R.Bharathi[1],Dr.V. Shanthi[2]

[1.] Research Scholar, Department of ECE, Sathyabama University,
AP, Vel Tech University, Chennai, India.

[2.] Professor, Department of MCA, St. Joseph's College of Engineering,
Chennai, India.

Abstract

A wide range of researches are carried out in speech signal processing for denoising, enhancement and more. Besides the other, stress management is important to improve disabled children speech. In order to provide proper speech practice for the disabled children, their speech is analyzed. Initially, the normal and pathological subjects speech are obtained with the same set of words. In this paper, classification of normal and pathological subjects speech is discussed. Initially Feature Extraction is implemented using well known Mel Frequency Cepstrum Coefficients (MFCC) for both words of normal and pathological subjects' speech. Dimensionality reduction of features extracted is implemented using Principal Component Analysis (PCA). Finally the features are trained using Artificial Neural Network (ANN) for classification.

Keywords*: speech signal, stress management****,*** *Mel Frequency Cepstrum Coefficients (MFCC), Principal Component Analysis (PCA), trained*

1. Introduction

Terminology for Mental Retardation/Intellectual Disability (MR/ID) has been particularly challenging as the term *mentally retarded* carries significant social and emotional stigma. Developmental delay is often used inappropriately as synonymous with MR/ID. Developmental delay is an overlay inclusive term and should generally be used for infants and young children in which the diagnosis is difficult, such as those too young for formal testing.

Speech is the most basic means of communication between humans. Effective determination of quantitative speech disability remains one of the challenges in medical profession. However millions of people are suffering from disabilities associated with speech. Speech synthesis among the disabled children is the first step for making speech corrective tool kit.

Approximately 10% of children have some learning impairment, while as many as 3% manifest some degree of MR/ID. In every verbal communication, the quality and precision of speech are given greater importance [2]. Due to the presence of deleterious properties in the acoustic environment such as multipath distortion (reverberation) and ambient noise, the performance of speech and speaker recognizers are often degraded [3]. Speech communication applications such as voice-controlled devices, hearing aids, and hands-free telephones mostly suffer from poor speech quality because of background noise and room echo [4]. Most of the time, particularly during travel, we meet noisy environment. Signal processing techniques removes the noise from the signal-noise mixture and provides an almost noise-free sound for enhanced communication [2]. Signal processing techniques are exploited in several applications such as speech acquisition, acoustic imaging and communications purposes [5].

In the modern period, there is a great interest for developing techniques for both speech (and character/word sequences) recognition and synthesis [6]. Generally, the speech recognition has two stages: feature extraction and classification [8].

In this paper, we concentrate on the pathological subject's speech and to analyse the speech of them by comparing it with the normal children. This work is very useful for the speech practitioners in the way in which position they have to improve the speech of the abnormal person. Initially, the samples of the normal as well as the pathological subject's speech have been obtained and with the aid of these samples, the further process has to be carried out. Initially, the MFCC of both the speeches are extracted and the PCA is applied to the MFCC to reduce the dimensionality of the speeches. After that the parameters that are extracted from this MFCC of both speeches are then sent as an input to generate the ANN. The abnormal and normal features are used to train the network for classification.

Feature extraction is a key issue for efficient speaker recognition. Additionally, a reduced feature set would allow more robust estimates of the model parameters, and less computational resources would be required. Best features are those that help to discriminate among speakers. A small amount of data is enough to estimate good models.

State-of-the-art systems use the same short-term spectrum features (Mel-Frequency Cepstral Coefficients, MFCC) for speech and speaker recognition, because MFCC convey not only the frequency distribution identifying sounds, but also the glottal source and the vocal tract shape and length, which are speaker specific features. Additionally, it has been shown that dynamic information improves significantly the performance of recognizers, so MFCC are commonly used as features.

Principal Component Analysis (PCA), an technique of multivariate statistical analysis [1], consists of computing the eigenvectors of the D*D covariance matrix X, then sorting them according to the corresponding eigenvalues, in descending order, and finally building the projection matrix A (called Karhunen-Loeve Transform, KLT) with the largest K eigenvectors (i.e. the K directions of greatest variance). Each feature vector X is then pre-processed according to the expression Y=A(X-,u), where u represents the mean feature vector. KLT decorrelates the features and provides the smallest possible reconstruction error among all linear transforms, i.e. the smallest possible mean-square error between the data vectors in the original D-feature space and the data vectors in the projection K-feature space.

The study of Neural Networks, was initially inspired by neurobiology, but it has since become a very interdisciplinary field, spanning computer science, electrical engineering, mathematics, physics, psychology, and linguistics as well. Some researchers are still studying the neurophysiology of the human brain, but much attention is now being focused on the general properties of neural computation, using simplified neural models. These properties include:

• **Trainability:** Networks can be taught to form associations between any input and output patterns. This can be used, for example, to teach the network to classify speech patterns into phoneme categories.

• **Generalization:** Networks don't just memorize the training data; rather, they learn the underlying patterns, so they can generalize from the training data to new examples. This is essential in speech recognition, because acoustical patterns are never exactly the same.

• **Nonlinearity:** Networks can compute nonlinear, nonparametric functions of their input, enabling them to perform arbitrarily complex transformations of data. This is useful since speech is a highly nonlinear process.

• **Robustness:** Networks are tolerant of both physical damage and noisy data; in fact noisy data can help the networks to form better generalizations. This is a valuable feature, because speech patterns are notoriously noisy.

• **Uniformity:** Networks offer a uniform computational paradigm which can easily integrate constraints from different types of inputs. This makes it easy to use both basic and differential speech inputs, for example, or to combine acoustic and visual cues in a multimodal system.

• **Parallelism:** Networks are highly parallel in nature, so they are well-suited to implementations on massively parallel computers. This will ultimately permit very fast processing of speech or other data.

An neural network consists of a potentially large number of simple processing elements (called *units*, *nodes*, or *neurons*), which influence each other's behavior via a network of excitatory or inhibitory weights. Each unit simply computes a nonlinear weighted sum of its inputs, and broadcasts the result over its outgoing connections to other units. A training set consists of patterns of values that are assigned to designated input and/or output units. As patterns are presented from the training set, a learning rule modifies the strengths of the weights so that the network gradually learns the training set. This basic paradigm1 can be fleshed out in any different ways, so that different types of networks can learn to compute implicit functions from input to output vectors, or automatically cluster input data, or generate compact representations of data, or provide content-addressable memory and perform pattern completion.

Neural networks are usually used to perform static pattern recognition, that is, to statically map complex inputs to simple outputs, such as an N-ary classification of the input patterns.

Moreover, the most common way to train a neural network for this task is via a procedure called *backpropagation*, whereby the network's weights are modified in proportion to their contribution to the observed error in the output unit activations (relative to desired outputs). To date, there have been many successful applications of neural networks trained by backpropagation.

The rest of the paper is organized as follows. The Feature Extraction and Classification using ANN are described in Section 2. The experimental setup is described in Section 3, including the speech database used to train. Experiments Results are presented and discussed in Section 4: (1) FeatureExtraction using MFCC and PCA and (2) Training and Classification using ANN. Finally, Section 5 summarizes our approach.

2. Methodology

2.1 Feature Extraction

In this work, MR children speech of mild degree is taken as Pathological children speech dataset. For improving the abnormal speech initially Feature extraction, dimensionality reduction and Classification is done. Normal and pathological subject's speech dataset is obtained and then the MFCC is extracted from both the speeches. This speech dataset is developed through in which both the databases are same set of scripts. Subsequently, with the aid of the PCA the dimensionality is reduced for MFCC features extracted. The step by step processes are explained in detail in the following subsections.

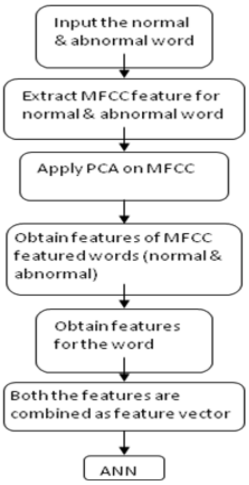

Input the normal & abnormal word
Extract MFCC feature for normal & abnormal word
Apply PCA on MFCC
Obtain features of MFCC featured words (normal & abnormal)
Obtain features for the word
Both the features are combined as feature vector
ANN

Fig.1 Data Flow for Feature Extraction

2.1.1 MFCC

In this segment, the speech samples are extracted from the normal and pathological subjects with the aid of the audio synthesizer. Let D_a and D_b are the abnormal children, normal children speech datasets respectively and from these datasets the MFCC feature is extracted.

$$D_a = \{w_1, w_2, w_3 ... w_{N_w - 1}\} \qquad (1)$$

$$D_b = \{\omega_1, \omega_2, \omega_3 ... \omega_{N_w - 1}\} \qquad (2)$$

MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [13]. The following steps are involved in extracting the MFCC feature

- Fourier transform is taken for the signal
- With the aid of triangular overlapping windows, the power spectrums are compared with mel scale
- At each of the mel frequency the logs of powers are taken
- Discrete Cosine Transform (DCT) is taken for the mel log powers
- Obtaining the resulting spectrums is the amplitude of MFCCs.

With the utilization of above steps, the MFCC features are obtained from the normal as well as abnormal datasets which is referred as M_a and M_b

2.1.2 Principal Component Analysis (PCA)

PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics - because it is a simple, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it.

$$N = w - \mu \qquad (3)$$

$$cv = \frac{N \times N^T}{n-1} \qquad (4)$$

$$Y = N * E^T \qquad (5)$$

$$\frac{Y}{E^T} * \mu = w \qquad (6)$$

N – Mean Deviation
w - window
μ - mean
cv - covariance vector
E – *Eigenvector*

The aforesaid equations are utilized to obtain the PCA of both M_a and M_b, where given equations are the general sets of equations to generate PCA. In PCA the data are processed by window by window. After PCA the inverse PCA also applied to obtain the dimensionality reduced original information again. After this process completed, the following parameters are obtained from the MFCC featured vectors M_a, M_b. The parameters are mean, standard deviation, maximum amplitude value and its id, minimum amplitude value and its id, MFCC length are extracted for the MFCC featured word and as well as for the original word also extracted and hence for each word we have 14 inputs.

2.2 Classification through Neural Network (NN)

The Artificial neural network (ANN) used in the model was a multilayer perceptron (MLP) with two layers of neurons. The number of neurons in the hidden layer is dependent on the size of the input vector [13]. The output layer has one neuron. Both the words of normal and pathological subjects are inputted to the ANN to be trained and to identify the abnormal word. The 14 features extracted for each word is given as input to neural network.

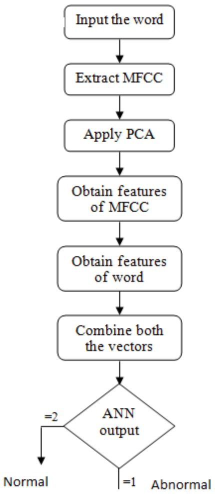


Fig. 2 Data Flow of Classification

NN is utilized for the classification purpose in order to identify the normal speech and the abnormal speech. In order to identify this, initially the MFCC features are obtained from both the normal and abnormal words. After that, both the words are inputted to the NN to identify the abnormal word. A artificial neural network is developed with a systematic step-step procedure which optimizes a criterion commonly known as the learning rule. The input/output training data is fundamental for these networks as it conveys the information which is necessary to discover the optimal operating point.

Once an input is presented to the neural network, and a corresponding desired or target response is set at the output, an error is composed from the difference of the desired response and the real system output.The error information is fed back to the system which makes all adjustments to their parameters in a systematic fashion (commonly known as the learning rule). This process is repeated until the desired output is acceptable.

The following components of the model represent the actual activity of the neuron cell. All inputs are summed altogether and modified by the weights. This activity is referred as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be -1 and 1.

Mathematically, this process is described in the figure 3.

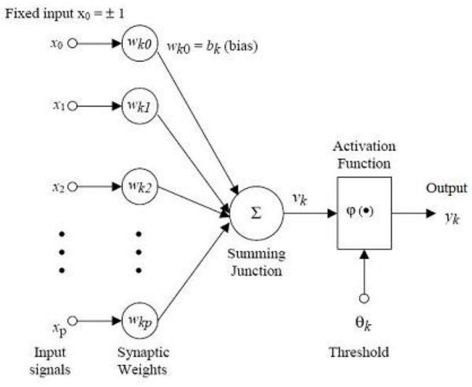


Fig. 3 ANN Model

From this model the interval activity of the neuron can be shown to be:

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

The output of the neuron, y_k, would therefore be the outcome of some activation function on the value of v_k.

3. Experimental Setup

3.1 The speech database

In this with the aid of the **Free Audio Editor** we generate the dataset with the normal and abnormal female children within the age limit 6-10. For normal data we utilized 2 female children and for abnormal data we utilized a female child for our system and their normal frequency range is from 20 - 4 kHz. The proposed technique is tested with the database of 100 words with two normal children and an abnormal child each.

3.2 Classification using NN

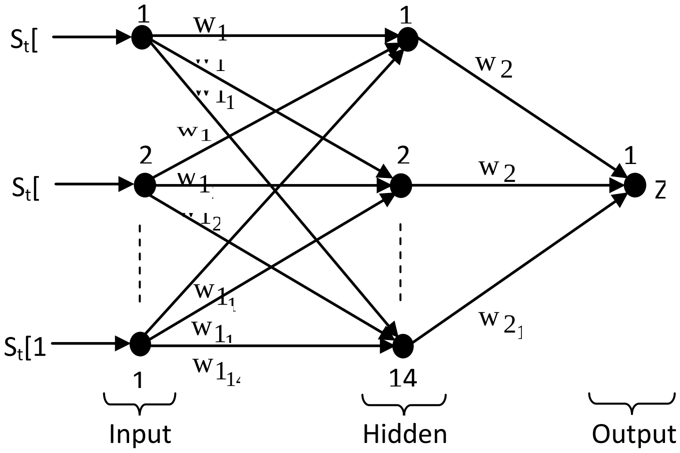
NN is used for the classification purpose in order to separate the normal speech and for identifying the abnormal speech. The following steps details the training process of ANN

Step 1: As the first step, set up the input weights to every neuron, apart from the neurons of the input layer.

Step 2: This neural network has 14 input layers which are the parameters of a word as said in the section 2.2. N_h hidden layers and one output layer to identify the word inputted is either normal or abnormal. The value at the output layer is either true or false depending on whether the input word is abnormal or normal. In this neural network, 14 input neurons and a bias neuron, N_h hidden neurons and a bias neuron and one output neuron are present.

Step 3: The weights are added to the designed network N, also it is biased. The developed N is shown in Fig.4.

Fig. 4.



Step 4: The basis function and the activation function which are chosen for the designed N is given below.

$$Z_i = \alpha + \sum_{j=1}^{N_h} \left(w_{ij} S_t[1] + w_{ij} S_t[1] + w_{ij} S_t[2] + w_{ij} S_t[3] + w_{ij} S_t[4] + ... + w_{ij} S_t[14] \right) \tag{7}$$

$$h(Z_i) = \frac{1}{1+e^{-Z_i}} \tag{8}$$

$$h(Z_i) = Z_i \tag{9}$$

Eq. (7) is the basis function for the input layer, where $S_t[1] - S_t[14]$ are the parameters for the word which are mean, standard deviation, maximum amplitude value and its id, minimum amplitude value and its id, MFCC length for MFCC featured word and for the original word we extract the same word for the MFCC length here we extract the word length. Here w_{ij} is the weight of the neuron and α is the bias. The sigmoid function for the hidden layer is given in Eq.(7) and the activation function for the output layer is given in Eq.(8). The basis function given in Eq. (7) is commonly used in all the remaining layers (hidden and output layer, but with the number of hidden and output neurons, respectively). The output of the ANN is obtained by giving the region vector as its input.

Step 5: The learning error is determined for the network N as follows

$$Er = \frac{1}{N_h}\sum_{o=0}^{N_h-1} D_o - Z_o z \tag{10}$$

Here, Er is the error in the FF-ANN, D_o is the desired output and Z_o is the actual output.

3.2.1 Error minimization by BP algorithm

The steps involved in the training of BP algorithm based NN is given below.

(1) Randomly generate weights in the interval $[0,1]$ and assign it to the neurons of the hidden layer and the output layer. But all neurons of the input layer have a constant weight of unity.

(2) Determine the BP error using Eq. (9), and give the training gene data sequence as input to the N Eq. (7), Eq. (38) and Eq. (9) show the basis function and transfer function.

(3) Adjust the weights of all the neurons when the BP error is determined as follows,

$$w_{ij} = w_{ij} + \delta w_{ij} \tag{11}$$

The change in weight δw_{ij} given in Eq. (7) can be determined as $\delta_{w_{ij}} = \gamma \, . Z_{ij} \quad . Er$, where Er is the BP error and γ is the learning rate and it normally ranges from 0.2 to 0.5.

(4) After adjusting the weights, repeat steps until the BP error gets minimized. Normally, it is repeated till the criterion, $E < 0.1$ is satisfied.

Once the error gets minimized to a minimum value it is concluded that the designed FF-ANN is well trained for its further testing phase and the BP algorithm is terminated. Thus the neural network is trained using the parameters for each word.

4. Experimental Results

4.1 Feature Extraction

Initially, the words are extracted from the both normal and abnormal children and then the MFCC feature has been extracted from it. Subsequently, the PCA is applied to reduce the dimensionality of the words. Few speech samples which are sent as input to MFCC Feature Extraction.

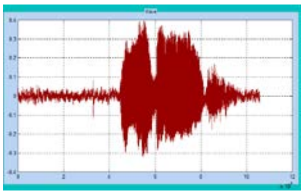


Fig. 5 Normal speech 1 “wild animals”

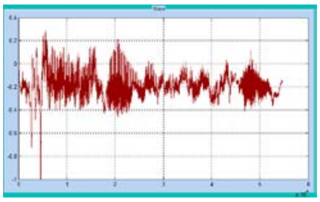


Fig. 6 Normal speech 2-“wild animals”

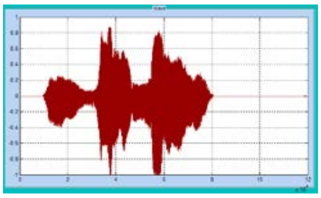
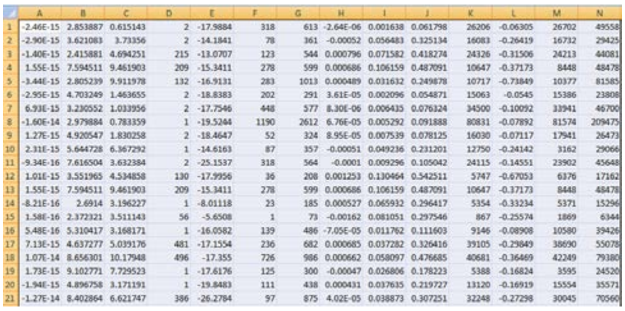


Fig. 7 Abnormal speech – “wild animals”



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	-2.46E-15	2.853887	0.615143	2	-17.9884	318	613	-2.64E-06	0.001638	0.061798	26206	-0.06305	26702	49558
2	-2.90E-15	3.621083	3.73356	2	-14.1841	78	361	-0.00052	0.056483	0.325134	16083	-0.26419	16732	29425
3	-1.40E-15	2.415881	4.694251	215	-13.0707	123	544	0.000796	0.071582	0.418274	24326	-0.31506	24213	44081
4	1.55E-15	7.594511	9.461903	209	-15.3411	278	599	0.000686	0.106159	0.487091	10647	-0.37173	8448	48478
5	-3.44E-15	2.805239	9.911978	132	-16.9131	283	1013	0.000489	0.031632	0.249878	10717	-0.73849	10377	81585
6	-2.95E-15	4.703249	1.463655	2	-18.8383	202	291	3.61E-05	0.002096	0.054871	15063	-0.0545	15386	23808
7	6.93E-15	3.230552	1.033956	2	-17.7546	448	577	8.30E-06	0.006435	0.076324	34500	-0.10092	33941	46700
8	-1.60E-14	2.979884	0.783359	1	-19.5244	1190	2612	6.76E-05	0.005292	0.091888	80831	-0.07892	81574	209475
9	1.27E-15	4.920547	1.830258	2	-18.4647	52	324	8.95E-05	0.007539	0.078125	16030	-0.07117	17941	26473
10	2.31E-15	5.644728	6.367292	1	-14.6163	87	357	-0.00051	0.049236	0.231201	12750	-0.24142	3162	29066
11	-9.34E-16	7.616504	3.632384	2	-25.1537	318	564	-0.0001	0.009296	0.105042	24115	-0.14551	23902	45648
12	1.01E-15	3.551965	4.534858	130	-17.9956	36	208	0.001253	0.130464	0.542511	5747	-0.67053	6376	17162
13	1.55E-15	7.594511	9.461903	209	-15.3411	278	599	0.000686	0.106159	0.487091	10647	-0.37173	8448	48478
14	-8.21E-16	2.6914	3.196227	1	-8.01118	23	185	0.000527	0.065932	0.296417	5354	-0.33234	5371	15296
15	1.58E-16	2.372321	3.511143	56	-5.6508	1	73	-0.00162	0.081051	0.297546	867	-0.25574	1869	6344
16	5.48E-16	5.310417	3.168171	1	-16.0582	139	486	-7.05E-05	0.011762	0.111603	9146	-0.08908	10580	39426
17	7.13E-15	4.637277	5.039176	481	-17.1554	236	682	0.000685	0.037282	0.326416	39105	-0.29849	38690	55078
18	1.07E-14	8.656301	10.17948	496	-17.355	726	986	0.000662	0.058097	0.476685	40681	-0.36469	42249	79380
19	1.73E-15	9.102771	7.729523	1	-17.6176	125	300	-0.00047	0.026806	0.178223	5388	-0.16824	3595	24520
20	-1.94E-15	4.896758	3.171191	1	-19.8483	111	438	0.000431	0.037635	0.219727	13120	-0.16919	15554	35571
21	-1.27E-14	8.402864	6.621747	386	-26.2784	97	875	4.02E-05	0.038873	0.307251	32248	-0.27298	30045	70560

Fig. 8 Feature Extraction output of Normal Speech of Child 1

4.2 Classification using ANN

Subsequently, for MFCC features the PCA is applied to reduce the dimensionality of the words and then they are inputted to the neural network to identify the abnormal and the normal word. The target is fixed as 1 and 2 for abnormal and normal speech respectively. Fig.9 is the generated neural network structure.

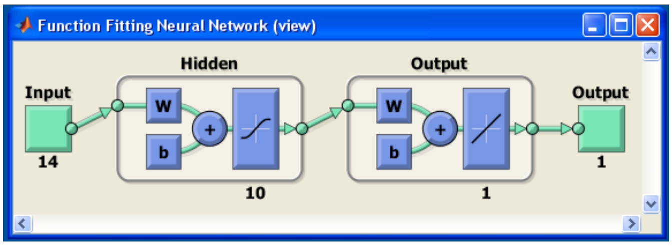


Fig. 9 Generated ANN for classification

5. CONCLUSIONS AND FUTURE WORK

The proposed system was implemented in the working platform of **MATLAB (version 7.11).** In this with the aid of the **Free Audio Editor** we generate the dataset with the normal and abnormal female children within the age limit 6-10. For 100 normal data (words) we utilized 2 female children each and for 100 abnormal data we utilized a female child for our system and their normal frequency range is from 20 – 4khz. The speech input is recorded at a sampling rate of 44.1kHz. To develop an effective system to identify the abnormal word and the spot, where the speech has to be improved, initially the MFCC is obtained from both the normal and abnormal words and then PCA dimensionality reduction is done. Then classification of normal and abnormal words are done using ANN. The result of this work is discussed with outputs. After that, the work will be extended by (i) Testing Phase (ii) acute Spotting aberration in speech of pathological subject respectively.

References

[1] Shirbahadurkar and Bormane, "Speech Synthesizer Using Concatenative Synthesis Strategy for Marathi language (Spoken in Maharashtra, India)", International Journal of Recent Trends in Engineering, Vol. 2, No. 4, pp. 80-82, November 2009

[2] Singaram, Guru Raghavendran, Shivaramakrishnan and Srinivasan, "Real Time Speech Enhancement using Blackfin Processor BF533", J. Instrument Society of india, Vol. 37, No. 2, pp. 67-79, 2009

[3] Qiguang Lin Ea-Ee Jan and James Flanagan, "Microphone Arrays and Speaker Identification", IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 622-629, October 1994

[4] Thomas Lotter, Christian Benien and Peter Vary, "Multichannel Direction-Independent Speech Enhancement Using Spectral Amplitude Estimation", EURASIP Journal on Applied Signal Processing, Vol. 2003, No. 11, pp. 1147-1156, 2003

[5] Sven Nordholm, Thushara Abhayapala, Simon Doclo, Sharon Gannot, Patrick Naylor and Ivan Tashev, "Microphone Array Speech Processing", EURASIP Journal on Advances in Signal Processing, Vol. 2010, pp. 1-3, 2010

[6] Marius Crisan, "Chaos and Natural Language Processing", Acta Polytechnica Hungarica, Vol. 4, No. 3, pp. 61-74, 2007

[7] Aida–Zade, Ardil and Rustamov, "Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems", World Academy of Science, Engineering and Technology, Vol. 3, No. 2, pp. 74-80, Spring 2007

[8] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go and Chungyong Lee, "Optimizing Feature Extraction for Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 1, pp. 80-87, January 2003

[9] Rashad, Hazem M. El-Bakry and Islam R. Ismail, "Diphone Speech Synthesis System for Arabic Using MARY TTS ", International journal of computer science & information Technology (IJCSIT), Vol. 2, No. 4, pp. 18-26, August 2010

[10] Stelzle, Ugrinovic, Knipfer, Bocklet, Noth, Schuster, Eitner, Seiss and Nkenke, "Automatic, computer-based speech assessment on edentulous patients with and without complete dentures - preliminary results", Journal of Oral Rehabilitation, Vol. 37, No. 3, pp. 209-216, March 2010

[11] Sumathi and SanthaKumaran, "Pre-Diagnosis of Hypertension Using Artificial Neural Network", Global Journal of Computer Science and Technology, Vol.11, No.2, pp.43-47, February 2011

[12] El-Shafie, Mukhlisin, Najah and Taha, "Performance of artificial neural network and regression techniques for rainfall-runoff prediction", International Journal of the Physical Sciences, Vol.6, No.8, pp.1997-2003, April 2011

[13] Rashidul Hasan, Mustafa Jamil, Golam Rabbani and Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", In proceedings of 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, December 2004

C.R.Bharathi received AMIE (ECE) and M.E.(Applied Electronics) in 2001 and 2005. Since June 2002, she has been a Lecturer in an Engineering College and at present working in Vel Tech University, Avadi, Chennai.She is an Research scholar at Sathyabama University.

Dr. V. Shanthi working as Professor in St.Joseph's College of Engineering,Chennai. Her research interest includes Artificial Intelligence, cloud computing. Dr. V. Shanthi is co-author.