Linear predictor coefficient, power spectral analysis and two-layer feed forward network for bangla speech recognition

Md. Shafiul Alam Chowdhury
Department of Computer Science and Engineering
Islamic University
Kushtia, Bangladesh
shafiul.a.chowdhury@gmail.com

Md. Farukuzzaman Khan
Department of Computer Science and Engineering
Islamic University
Kushtia, Bangladesh
mfkhanbd2@gmail.com

***Abstract*—Within this research we investigates the tools and techniques how correctly recognize Bangla (Bengali) language. For feature extraction the power spectral analysis and linear predictor coefficient method considered. Speech samples taken from bangla phoneme, word, command and sentence. Each frame of a speech signal's spectrum segmented into number of fragments to take the mean absolute value from each fragment for pattern classification. A two-layer feed forward neural network with maximum likelihood method used. Features extraction by power spectral analysis gives little more accuracy for bangla phoneme and word recognition comparing to linear predictor coefficient analysis. Both methods provide average result for bangla command and sentence recognition. The mel frequency cepstral coefficient could be another approach for future research.**

Keywords—feature extraction; speech recognition; power spectral analysis; linear predictor coefficient; feed forward network.

I. INTRODUCTION

Last hundred years various research conducted in speech recognition in many spoken languages, software with necessary hardware developed for better speech processing and recognition purpose. Bangla speech corpus development for speaker independent continuous speech recognition not successful in noisy and speaker variability environment [1]. Bangla digit automatic speech recognition [2] and word recognition successful in very narrow scale [3]. All the developed systems not able provide full accuracy for speech recognition in any spoken language. So, only few research conducted in bangla (Bengali) spoken language even though it is native language for 300 million people in the globe. This paper is about the description of our developed system how performs for bangla speech recognition.

II. RESEARCH SCOPE

Eight Bangla phonemes, isolated words, commands and six bangla sentences taken for analysis mentioned in experiment and result section. Single male, female and male-female from different age group for different utterances taken as speech sample. Utterances (eight hundred of bangla speech samples) of each phoneme, word, command and sentence recorded several times from the same person in a vacant room at night with necessary recording equipment and computer to avoid noise as much as possible. Power spectral analysis and linear predictive coefficient analysis with Simon haykin's technique [4][12] applied for feature extractions in bangla speech signals. To get the features we have used power spectral analysis and liner predictive coefficient analysis separately for pattern classification and speech recognition in a trained supervised two-layer feed forward neural network with maximum likelihood method. Number of hidden neuron samples with the necessary speech samples trained in this network. The approach to find effective and proper tools and techniques for bangla speech recognition for more accuracy.

III. SHORT TIME ENERGY CALCULATION, SILENCE REMOVAL, FRAMING AND PRE-PROCESSING

We have applied short time energy (STE) technique for silence removal and energy calculation to remove the silence portion from sound signal [5]. The speech signals divided into number of rectangular window frames (16 milliseconds) to remove those frames with less energy present, calculates the short time energy for each frame, then normalized the energy of each frame and reject those frame/s which has energy less than 2% of the maximum energy component, rest of all were remain. We have (separately) split all the speech signals into 20 and 64 milliseconds time durations for the experiment. Frame overlapping not applied. Each frame gone through pre-process technique. Zero padding technique [6] applied in the case of total samples if do not fit an in integer number of frames. We have applied hamming window length of the size of the frame that is defined [7] by the following equation:

$$w(n) = 0.54 - 0.46\cos(2\pi n / N),...(0 < n < N)$$

$$\text{The Window length } L = N + 1$$

The speech is analyzed to extract a set of features representing the spectral envelope.

- Hamming Window (windowing) + Pre-emphasis

@IEEE 978-1-7281-1524-5

Pre-emphasis filter applied to the signal to amplify the high frequencies to balance the frequency spectrum [8]. A typical signal pre-emphasis defined by the equation [9] –

$$y(n) = s(n) - Cxs\ (n-1)$$

where the constant *C* falls generally between 0.9 and 1.0 The pre-emphasis done by using all-zero filter [9].

IV. Feature Extraction

We have done fourier transform of the signal. For normalization First fourier transform (FFT) have done to receive the absolute values. We divides the spectrum of all frames (received after pre-processing). The sampling frequency received 44100 Hz, basically a signal of frames that converts to a frequency domain. For power spectral analysis (FFT) and linear predictor coefficient analysis (LPC) we have divided the spectrum into four and seven segments [4] in the frequency domain for each frame and received four dimensional feature vector in FFT and seven in LPC, i.e.; 0-1 KHz, 1-2 KHz, 2-3 KHz and 3-4 KHz feature vector signal etcetera to that mean value of each spectrum.

A. Power Spectral Analysis (Fast Fourier Transform –FFT)

The power spectrum of a speech signal describes the frequency content of the signal over time. The Fourier Transform for a discreet time signal *f(kT)* is given by:

$$F(n) = \sum_{k=0}^{N-1} f(kT)\, e^{-\frac{j2\pi}{N}.nk} \qquad (1)$$

which can be written as-

$$F(n) = \sum_{k=0}^{N-1} f(k)\, W_N^{-nk} \qquad (2)$$

Where $f(k) = f(kT)$ and $W_N = e^{j2\pi/N}$, W_N is usually referred to as the kernel [10] of the transform.
The power spectrum is defined [11] as-

$$P(f_k) = 1/N^2 \left[|F_k|^2 + |F_{N-k}|^2 \right] \text{-----} k = 1,2,\ldots\ldots,(N/2-1) \qquad (3)$$

Where f_k is defined only for the zero and positive frequencies.

$$f_k \equiv k/N\Delta = 2 f_c\, k/N \text{----} k = 0,1,\ldots., N/2$$

The speech data is segmented into K segments of N = 2M points for the computation of the power spectrum's speech, N is the length of a window here. The power of 2 taken as of the convenient computation for FFT.

B. Linear Predictive Coefficient (LPC) Analysis

The problem addressed by linear prediction is given a linear, time-invariant, discrete-time system whose parameters are not known [12]. The linear prediction model (LPC) shown [13] in Fig. 1 consists of an excitation source *U(z)* supplying input to a spectral shaping filter *H(z)*, to yield output speech $\hat{S}(n)$. A synthetic speech sample, $\hat{S}(n)$, can be modeled by a linear combination of the *p* previous output samples and *q* previous input samples of an LPC synthesizer:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \hat{s}(n-k) + G \sum_{l=1}^{q} b_l u(n-l)$$

Where *G* is a gain factor for the input speech and $b_0 = 1$. Then the spectral shaping filter *H(z)* can be defined by

$$H = \frac{\hat{S}(z)}{U(z)} = G \times \frac{1 + \sum_{i=1}^{q} b_l z^{-1}}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

Most linear predictor coefficient (LPC) work assumes an all-pole model or AutoRegressive (AR) model, where $q = 0$. An all-zero model (p=0) is called a Moving Average (MA) and model with both poles and zeros is known as the AutoRegressive Moving Average (ARMA). Hereafter we focus on the AR model. The fig.1 is LPC Synthesis model.

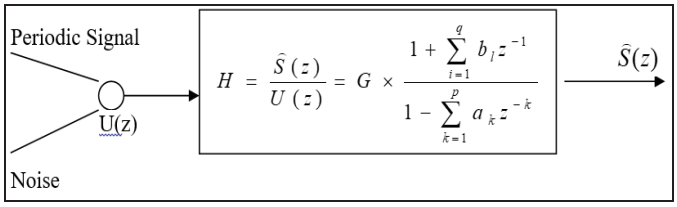

Fig. 1.

$$H(z) = \frac{\hat{S}(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

For the above system, the speech samples *s(n)* are related to the excitation *u(n)* by the simple difference equation-

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (1)$$

Then a predictor with prediction coefficients, α_k is defined as a system whose output is-

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \qquad (2)$$

The system function of a p^{th} order linear predictor is the polynomial-

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (3)$$

The prediction error, *e(n)*, is defined as

@IEEE 978-1-7281-1524-5

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \quad (4)$$

It can be seen from above equation that the prediction error sequence is the output of a system whose transfer function is

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \quad (5)$$

Comparing equation (1) and (4), if $\alpha_k = a_k$, then *e(u) = Gu(n)* and the prediction error filter, *A(z)*, will be an inverse filter for the system, *H(z)*, i.e., *H(z) = G/A(z)*. Our basic approach is to find a set of predictor coefficients that will minimize the mean-squared prediction error over a short segment of the speech waveform. The short time average prediction error could be defined as-

$$E_n = \sum_m e_n^2(m)$$

$$= \sum_m (s_n(m) - \hat{s}_n(m))^2$$

$$= \sum_m \left(s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right)^2 \quad (6)$$

where $s_n(m)$ is a segment of speech that has been selected in the vicinity of sample n, i.e.,

$$s_n(m) = s_n(m+n)$$

We can find the values of α_k that minimize E_n in equation (6) by setting –

$\partial E_n / \partial \alpha_i = 0$, *i=1,2, 3,……p,* thereby obtaining the equation-

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{\alpha}_k \sum_m s_n(m-i)s_n(m-k) \quad (7)$$

where $\hat{\alpha}_k$ are values of α_k that minimize E_n. For simplicity we can drop the caret on $\hat{\alpha}_k$ and if we define

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \quad (8)$$

then equation (6) and (7) can be written more compactly as

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \quad i = 1,2,...p \quad (9)$$

Using equation (6) and (7), the minimum mean-squared prediction error can be shown to be

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} \alpha_k \sum_m s_n(m)s_n(m-k) \quad (10)$$

and using equation (9)-

$$E_n = \phi_n(0,0) - \sum_{k=1}^{p} \alpha_k \phi_n(0,k) \quad (11)$$

C. The Autocorrelation Method [14]

For a finite length window *w(n), $s_n(m)$* is identically zero outside the window interval $0 \le m \le N-1$, it can be shown that

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k) \; .. 1 \le i \le p \text{ and } 0 \le k \le p$$

can be expressed as-

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k) \; .. 1 \le i \le p \text{ and } 0 \le k \le p$$

Furthermore it can be seen that in this case $\phi_n(i,k)$ is identical to the short-time auto correlation function for *(i-k)*. That is $\phi_n(i,k) = R_n(i-k)$ where

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k)$$

Since $R_n(k)$ is an even function, it follows that

$$\phi_n(i,k) = R_n(|i-k|) \quad i=1, 2,...p \text{ and } k = 0, 1,...p$$

Therefore equation (9) can be written as-

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \quad 1 \le i \le p \text{ and } 0 \le k \le p \quad (12)$$

similarly, the minimum mean-squared prediction error of equation (11) takes the form- $E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k)$

The set of equations (12) can be expressed in matrix form as

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ & & \dots & \dots & \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ R_n(p) \end{bmatrix}$$

The p×p matrix of auto-correlation values is a Toeplitz matrix; i.e., it is symmetric and all the elements along a given diagonal are equal. $s_i(n) = \sum_{j=-(N_{FB_i}-1)/2}^{(N_{FB_i}-1)/2} a_{FB_i}(j)s(n+j)$

@IEEE 978-1-7281-1524-5

D. The Durbin's Recursive Solution [14]

By exploiting the Toeplitz nature of the matrix of coefficients, several efficient recursive procedures have been devised for solving system of equations. The most popular and well known methods are the Levinson and Rabinson algorithms [15] most efficient method known for solving particular system of equations is Durbin's recursive procedure can be stated as follows: $E(0) = R(0)$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right] \Big/ E^{(i-1)} \quad 1 \le i \le p$$

$$\alpha_i^{(i)} = k_i$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

Above equations are solve recursively for $i=1, 2,....,p$ and the final solution is given as $\alpha_j = \alpha_j^{(p)} \quad 1 \le j \le p$

V. Training And Target Data

We have created a data vector for training and the cross-ponding variable 1 uses for defining the target vector. Each column represents Frame or frame number where thousands of frames were created. Each row represents 1st, 2nd, 3rd and 4th feature of each frame (0-1 KHz, 1-2 KHz, 2-3 KHz, 3-4 KHz) for FFT feature. For LPC it is seven features. Table I shows a few *training data* where eight bangla word uttered 40 times (each word uttered five times) by a single male and 689 frames created. Each frame extracted into seven features. So, 689 X 7 = 4823 features created. Columns represent frame or number of frame and rows represents the feature. This is training data where many features for 1st utterance or 1st variable 1, 2, 3 then 4, 5, 6 for 2nd utterance or 2nd data and so on. Fig. 2 shows seven feature of a single frame (1st frame of a word) for LPC. One utterance of phoneme/ word/ command or sentence contains many number of frames. We plot seven values (-13462, 1.2335, -1.6723, 0.8737, -0.5197, 0.2645, 0.2016) received from feature extraction of 1st frame (Table I) for LPC mentioned in Fig. 2.

Fig. 3 shows seven features of 689 frames in the same window for isolated Bangla word. Forty utterances by a same individual male person (eight different word uttered and each word five times). We have plot the seven values from feature extraction of 689 frames. A target vector has been created (Table II) to identify the correct frame for the correct utterance of phoneme, word, command and sentence. Colum represents utterances and row represent different samples (phoneme/ word/ command/ sentence). There is eight row for eight phoneme/ word/ command, six rows for six sentence and many columns for utterances. The neural network can understand which particular frame belongs to which phoneme, word, command or sentence.

TABLE I.

Feature of each frame in KHz	Training data for isolated bangla word (Frame numbers)							
	1st Frame	2nd Frame	3rd Frame	4th Frame	5th Frame	6th Frame	7th Frame	Nth (689th frame)
1st feature (0- 1)	-1.3462	-1.5636	-1.6534	-1.7251	-1.8603	-1.9294	-1.6634	-1.3208
2nd feature (1- 2)	1.2335	1.5989	1.6796	1.7644	1.8988	1.9069	0.9802	1.7477
3rd feature (2- 3)	-1.6723	-1.9016	-1.9300	-2.0543	-2.3130	-2.3661	-1.4773	-1.8560
4th feature (3- 4)	0.8737	1.0980	1.2453	1.4461	1.8421	2.0387	1.4149	1.4729
5th feature (4- 5)	-0.5197	-0.6398	-0.8497	-0.9820	-1.0951	-1.1966	-0.3963	-1.2017
6th feature (5- 6)	0.2645	0.1875	0.3651	0.5135	0.6993	0.8634	0.6067	0.4333
7th feature (6- 7)	0.2016	0.2562	0.1718	0.0667	-0.1541	-0.3066	-0.4610	-0.2718

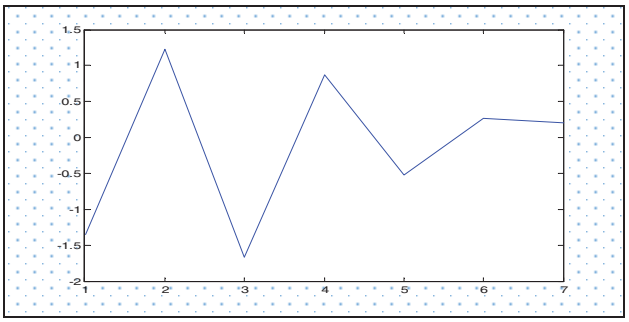

Fig. 2.

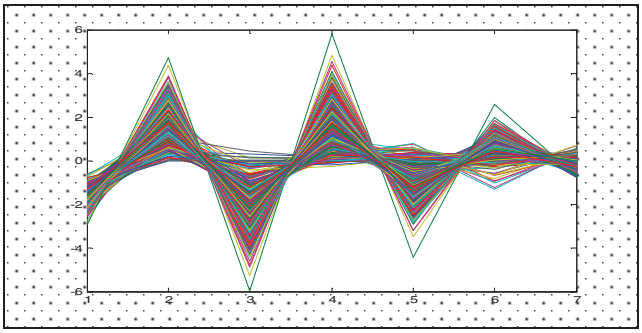

Fig. 3.

TABLE II.

Target data table (Utterance) Phoneme/ Word/ Command/ Sentence	Utterance-1	Utterance-2	Utterance-3	Utterance-4	Utterance-5	Utterance-6	Utterance-7	Utterance-N
1st Word	1	0	0	0	0	0	0	0
2nd Word	0	1	0	0	0	0	0	0
3rd Word	0	0	1	0	0	0	0	0
4th Word	0	0	0	1	0	0	0	0
5th Word	0	0	0	0	1	0	0	0
6th Word	0	0	0	0	0	1	0	0
7th Word	0	0	0	0	0	0	1	0
8th Word	0	0	0	0	0	0	0	1

VI. Two Layer Feed Forward Network

All the training data and target data (received from feature extraction) separately stored in matlab data file later used in the neural network. We have extracted the feature vector, created the data vector and target vector then trained the network. Once the network trained we need to identify the accuracy of the neural network for the same data. We have tested again the same features to pass through it. Before pass through it the same data has for gone from framing, pre-processing, windowing, feature extraction etcetera and placed in the test data. The test data to check with a function (in neural network) and the result goes to a *'result'* file where maximum likelihood condition [16][17] applied. The *'result'* file where each vector passes to each frame in neural network gives an output that is an amount of belongingness. Based on maximum likelihood condition the particular frame that has the maximum nearest value considered to belong to the particular phoneme/ word/ command/ sentence. Percentage calculation provides how much accuracy of recognition for particular speech sample.

VII. Experiment And Results

We have tried to find the best features and recognition methods among various methods and input category of system by conducting separate experiments for bangla phoneme, word, command and sentence. Each experiment 50 hidden neuron samples with recorded speech samples used. Bangla phoneme, word, command and sentence taken for feature extraction using power spectral analysis and linear predator coefficient analysis, pattern classification and recognition in two layer feed forward neural network for a single male, female, six/five male-female for LPC and FFT. Each speech sample uttered 5 times, i.e., for phoneme it could be 5 X 8 = 40 samples for single person. Maximum 240 samples used for each experiment extracted separately in FFT and LPC with two different window sizes (hamming window) of 20 and 64 milliseconds. For LPC each frame divided into seven parts, (0-1 MHz, 1-2 MHz, 2-3 MHz and till 6-7 MHz) and for FFT four parts.

A. Bangla Phoneme Recognition

Table III about bangla phoneme's feature extraction using FFT & LPC and recognition in neural network for single male, single female and six male-female. Eight Bangla phonemes (vowel and consonant) - "অ (/O/)", "আ (/A/)", " ই (/I/)", " উ (/OO/)", " এ (/EA/)", "ও (/O/)", " ঐ (/OI/)" and "ক (/KO/)" taken. Each phoneme's time duration lies between 1.018 to 1.201 seconds. The accuracy rate for single male and single female up to 95% in FFT and LPC (little varies in 20Ms & 64Ms window sizes). FFT and LPC methods performs almost same. Six male-female participation with 200 speech samples the recognition rate reduces to 54%.

B. Bangla Word Recognition

Table IV about bangla word's feature extraction pattern classification and recognition. Eight isolated bangla words - "অংক (Math)", "আমি (I)", "ইলিশ (Ilish)", "উট (Camel)", " কলা (Banana)", " খরগোশ (Rabbit)", " গরু (Cow)" and "ঘড়ি (Clock)" taken. Each word's time duration 1.201 seconds. Minimum 40 to maximum 200 samples used. Recognition rate for a single male in LPC lies between 55% and 87.5%, in FFT it lies between 65% and 70%. The accuracy rate for single female in LPC stands at 87.5% (20 and 64 Ms window), in FFT it lies between 72.5% and 75%. For five male-female (3 male & 2 female) recognition rate in FFT lies between 50.5% and 60% (both window) higher than LPC (43.5% and 47%). Recognition for single male or female LPC (up to 87.5%) is better than FFT (up to 75%) but for five male-female both (FFT & LPC) are poor (43.5% to 60%).

C. Bangla Command Recognition

Table V is about eight bangla command's feature extraction pattern classification and recognition. Minimum 40 to maximum 200 speech samples used. Eight Bangla command - " এই কাজ কর (Do this job)", "দরজা খোল (Open the door)", "টেবিল পরিষ্কার কর (Clean the table)", " বাম দিক যাও (Go to left)", "পশ্চিম দিক সরো (Move towards west)", "অফিস যাও (Go to office)", "এই চেয়ার আনো (Bring this chair)" and "জানালা বন্ধ কর (Close the window)" taken. Each command's time duration lies between 1.802 to 2.716 seconds. Bangla command's recognition rate for single male lies within 70% to 65% in LPC, it is 40% and 70% in FFT. The recognition rate for the single female lies within 32.5% to 25% in FFT, it is 35% and 50% in LPC. For five male-female recognition rate lies within 16.5% to 32.5% in FFT, it is 21% and 32% in LPC.

D. Bangla Sentence Recognition

Table VI about six bangla sentence's feature extraction pattern classification and recognition. Minimum 30 to maximum 150 speech samples used. Six bangla sentences - "আমরা কলা খাই (We eat bananas)", " কলা ভালো ফল (Banana is good fruits)", "ফল স্বাস্থের জন্য ভাল (Fruits are good for health)", "তারা তিন বন্ধু (They are three friends)", "তিন বন্ধু খেলা করে (Three friends play)" and "তিন বন্ধু খায় (Three friends eat)" taken. Each sentence's time duration lies between 2.011 to 3.213 seconds. The recognition rate for single male lies within 66.66% to 70% in FFT, it is 53.33% and 50% in LPC is poor rate. For single female recognition rate lies within 50% to 63.33% in FFT, it is 43.33% and 46.66% in LPC also poor rate. The recognition rate for five male-female lies within 42.66% and 44.66% in FFT, it is 34.66% and 31.33% in LPC which is poor rate too.

TABLE III.

Bangla phoneme recognition (Window length in 20 & 64 milliseconds, 50 Hidden neuron samples)												
Number of bangla Phoneme : 08	Single male				Single female				Six male-female			
	FFT		LPC		FFT		LPC		FFT		LPC	
	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms
No. of Utterances recognize	38 (Out of 40)	34 (Out of 40)	33 (Out of 40)	36 (Out of 40)	29 (Out of 40)	37 (Out of 40)	38 (Out of 40)	29 (Out of 40)	141 (Out of 240)	138 (Out of 240)	130 (Out of 240)	134 (Out of 240)

@IEEE 978-1-7281-1524-5

Percentage	95%	85%	82.5%	90%	72.5%	92.5%	95%	72.5%	60%	54.5 %	54.16 %	55.83 %

TABLE IV.

Bangla word recognition (Window length in 20 & 64 milliseconds, 50 Hidden Neuron samples)												
Number of bangla Word: 08	Single male				Single female				Five male-female			
	FFT		LPC		FFT		LPC		FFT		LPC	
	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms
No. of Utterances recognize	26 (Out of 40)	28 (Out of 40)	35 (Out of 40)	22 (Out of 40)	30 (Out of 40)	29 (Out of 40)	35 (Out of 40)	35(Out of 40)	101 (Out of 200)	120 (Out of 200)	87 (Out of 200)	94 (Out of 200)
Percentage	65%	70%	87.5%	55%	75%	72.5%	87.5%	87.5%	50.5%	60%	43.5%	47%

TABLE V.

Bangla Command recognition (Window length in 20 & 64 milliseconds, 50 Hidden Neuron samples)												
Number of bangla Command : 08	Single male				Single female				Five male-female			
	FFT		LPC		FFT		LPC		FFT		LPC	
	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms
No. of Utterances recognize	16 (Out of 40)	28 (Out of 40)	28 (Out of 40)	26 (Out of 40)	13 (Out of 40)	10 (Out of 40)	14 (Out of 40)	20 (Out of 40)	33 (Out of 200)	65 (Out of 200)	42 (Out of 200)	64 (Out of 200)
Percentage	40%	70%	70%	65%	32.5%	25%	35%	50%	16.5%	32.5%	21%	32%

TABLE VI.

Bangla sentence recognition (Window length in 20 & 64 milliseconds, 50 Hidden Neuron samples)												
Number of bangla Sentence: 06	Single male				Single female				Five male-female			
	FFT		LPC		FFT		LPC		FFT		LPC	
	20 Ms	64 Ms	20 Ms	64 Ms	20Ms	64 Ms	20Ms	64 Ms	20 Ms	64 Ms	20 Ms	64 Ms
No. of Utterances recognize	20 (Out of 30)	21 (Out of 30)	16 (Out of 30)	15 (Out of 30)	15 (Out of 30)	19 (Out of 30)	13 (Out of 30)	14 (Out of 30)	64 (Out of 150)	67 (Out of 150)	52 (Out of 150)	47 (Out of 150)
Percentage	66.66%	70%	53.33%	50%	50%	63.33%	43.33%	46.66%	42.66%	44.66%	34.66%	31.33%

VIII. Discussion And Conclusion

For phoneme recognition the accuracy rate for single male and female up to 95% both in FFT and LPC. Single male or female's word recognition in LPC (up to 87.5%) better than FFT (up to 75%). For male word recognition in LPC (up to 70%) slightly better than FFT. For female recognition LPC (50%) performs well than FFT (32.5%). FFT performs better than LPC for sentence recognition. For all the cases increasing number of participants and speech samples reduces the recognition rate. Changing of speaker (male/female) and window length slightly influences recognition rate. FFT & LPC approaches for feature extraction and recognition with two layer feed forward network is a good approach for bangla phoneme and word recognition, but not worked very well for command and sentence recognition. In a noisy environment with many people new tools and techniques (very large scale) could be introduced and implemented for experiment. Mel frequency cepstral coefficient analysis, hidden markov model could be applicable in the future research.

Acknowledgment

The research is a part of PhD research work by Md. Shafiul Alam Chowdhury under supervision of Professor Dr. Md. Farukuzzaman Khan, Department of Computer Science & Engineering, Islamic University, Kushtia, Bangladesh.

References

[1] Biswajit D, Sandipan M and Pabitra M, Bengali Speech Corpus for continuous automatic speech recognition system, 978-1-4577-0931-9 IEEE, 2011.

[2] G Muhammad, Yousef A., M N Huda, "12th International Conference on Computers and Information Technology", December 2009.

[3] Rahman, M., & Khatun, F. Development of Isolated speech recognition system for bangla words, DIUJST, vol. 6, pp. 30-35, 2011.

[4] Simon haykin, Communication systems, 4th edition, John Wiley & Sons Inc., 2001.

[5] Muhammad A and Shibli N, "A Silence removal and endpoint detection approach for speech processing, "3rd International Multidisciplinary research Conference on global prosperity through research & innovation", Sarhad University of Science and Information Technology, September 2016.

[6] Md Sah Hj S, Dzulkifli M and Shussain S, "Temporal Speech Normalization Methods Comparison in Speech Recognition Using Neural Network", International Conference of Soft Computing and pattern recognition, December 2009.

[7] Oppenheim, Alan V., Ronald W. Schafer, and John R. Buck. Discrete-Time Signal Processing. Upper Saddle River, NJ: Prentice Hall, 1999.

[8] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC", International Conference on Computer Graphics, Simulation and Modeling, Thailand, July 2012.

[9] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition", Canadian Conference on Electrical and Computer Engineering, Vol. 2, October 1995.

[10] M.A. Sid. Ahmed, Image Processing: Theory, algorithms and architectures, McGraw Hill, New York, 1995.

[11] William H. Press, Saul A. Teukolsky, William T. Vellerling and Brain P. Flannery, Numerical Recipes in C, Cambridge University Press, 1992.

[12] Richard A. Haddad and Thomas W. Parsons, Digital Signal Processing: Theory, Applications and Hardware, Computer Science Press, New York, USA. 1991.

[13] Jean-Claude Junqua & Jean-Paul Haton, Robustness in Automatic Speech Recognition: Fundamentals and Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[14] Lawrence R. Rabiner and Ronald W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, pp. 396-417, 1978.

[15] J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, New York, 1976.

[16] David Faraggi and Richard Simon, The Maximum Likelihood Neural Network As A Statistical Classification Model, Elsevier Journal of Statistical Planning and Inference, pp. 93-104, Vol. 46, July 1995.

[17] S.M. Kay and V. Nagesha, Maximum likelihood estimation of signals in autoregressive noise, IEEE Transactions on Signal Processing, Vol. 42, January 1994.