UTTERANCE BASED SPEAKER IDENTIFICATION USING ANN

Dipankar Das

Department of Information and Communication Engineering, University of Rajshahi,
Rajshahi-6205, Bangladesh

ABSTRACT

In this paper we present the implementation of speaker identification system using artificial neural network with digital signal processing. The system is designed to work with the text-dependent speaker identification for Bangla Speech. The utterances of speakers are recorded for specific Bangla words using an audio wave recorder. The speech features are acquired by the digital signal processing technique. The identification of speaker using frequency domain data is performed using backpropagation algorithm. Hamming window and Blackman-Harris window are used to investigate better speaker identification performance. Endpoint detection of speech is developed in order to achieve high accuracy of the system.

KEYWORDS

Speaker identification, digital signal processing, speech feature, ANN.

1. INTRODUCTION

Most of us can recognize a known person’s voice without seeing him, this ability of recognition is known as speaker identification. Human’s abilities both to understand the speech and to recognize the speakers from their voices have inspired many scientists to research in this field. However, prior to mid 1960’s, most of speech processing systems were based on analog hardware implementation. Since the advent of inexpensive digital computers and pulse code modulation (PCM), the speech area has undergone many significant advances. Successful speech processing systems require knowledge in many disciplines including acoustic wave spectrum, pattern recognition, and artificial intelligence techniques. In general, speech technology includes the following areas: speech enhancement, speaker separation, speech coding, speech recognition, speech synthesis, and speaker recognition. The area of speaker recognition can be divided into speaker identification and speaker verification. In this paper the main emphasis is on the speaker identification problem.

Speaker identification is important for controlling to secure facilities, personal information, services like banking, credit cheeks, etc. Today, an average person may use many different security items such as PINs (a Personal Identification Numbers) for automatic teller machines, phone cards, credit cards. These can be lost, stolen, or counterfeited.

Speaker verification is one area of general speaker recognition, which also includes speaker identification [1]. For speaker verification, an identity is claimed by the user, and the decision required of the verification system is strictly binary; i.e., to accept or reject the claimed identity. On the other hand, speaker identification is the labelling of an unknown utterance among

DOI : 10.5121/ijcsea.2014.4403

utterances of known speakers. The speaker identification can be done in two ways– text-dependent and text-independent speaker identification. In text-dependent speaker identification, the task is to identify the same utterance both in training and later in testing, where the utterances in training and in testing are not necessarily the same for text-independent speaker identification [16]. In this research, we propose the text-dependent speaker identification system using Artificial Neural Network (ANN).

Speaker differences that both enable and hinder speaker identification include inter-speaker and intra-speaker variations [2]. Inter-speaker variations, (i.e., between speakers) are due to the physical aspect of differences in vocal cords and vocal tract shape, and to the behavioural aspect of differences in speaking styles among speakers. Intra-speaker variations are the differences in the same utterance spoken by the same speaker: speaking rate, his emotional state, his health, etc. Variations in voices translate to variations in acoustic parameters. Good speaker identification system should capture these variations. Therefore, it is desirable to select those acoustic features that have the following characteristics [3]: (i) high inter-speaker and low intra-speaker variability, (ii) easy to measure and reliable over time, (iii) occur naturally and frequently in speech, (iv) stable in different transmission environments, and (v) difficult to imitate. In our research, we extract speech features, namely, the dominant frequency amplitude spectral components to capture the above characteristics of the speaker.

2. RELATED WORKS

Many papers and textbooks have described and proposed many different techniques to extract speaker features and to build automatic speaker identification systems with different assumptions and environments [2]. Most speaker identification systems used either template matching method or probabilistic modeling of the features of the speakers. In template matching method, the reference template of the claimed speaker created during the training phase is compared with the unknown template. On the other hand, probabilistic models employ long-term statistical feature averaging. Neural network technology together with speaker feature has been applied to identify speaker. Time-delay neural networks have been used successfully in both speech recognition and speaker recognition. Linear predictive parameters and their derived parameters related to speaker's vocal tract have been used for speaker identification system [2,4,5]. The linear predictive coding (LPC) derived parameter with hidden Markov models has been used in both speech and speaker identification system [6,7].

Vector quantization representing spectral feature with clustering technique using statistical properties is presented in [8]. The temporal identity mapping neural network has been used for text dependent speaker verification system [9]. Automatic speech recognition and speaker identification using artificial neural network (ANN) is described in [10].

No great work had been done on speaker identification for Bangla speech. However, some of the tasks had been done on feature extraction of Bangla word in [11]. Another feature extraction criterion, such as zero-crossing rate, short-time energy, pitch-extraction, formant frequencies have been studied [12].

3. DATA EXTRACTION AND PRE-PROCESSING

The sound components in a speaker identification system include the sound equipment, and an audio wave recorder. An audio wave recorder is programmed to record the speaker voice via a microphone. The recording is done for the present speaker identification system on a creative wave studio environment. Wave audio data resides in a file, which contains the digital sample values and descriptive information that identifies the particular format of that audio data. For speaker identification system, we extract the audio data from an audio wave file and process them.

3.1. Speech Endpoint Detection Algorithm

In order to extract speaker features we first detect the speech signal. Speech signal detection requires identifying the starting point and endpoint of the signal. In speaker identification system speech endpoint detection algorithm is used to detect the presence of speech, to remove pauses and silences in a background noise. The algorithm to be discussed here is based on the simple time domain measurement– short-term energy. The algorithm of speech endpoint detection is summarized below:

Endpoint Detection Algorithm:

```
Step1: Initialization

    i)      Set frame length L,
    ii)     Compute Speech length N,
    iii)    Set Pointer = 1.
Step2: Compute maximum frame energy

    i)      Read the file noise.wav,
    ii)     Segment data into 10ms (110 points) frame with 50% overlapping,
    iii)    Compute noise energy for each frame,
    iv)     Compute maximum frame energy, $E_m$.

Step3: Repeat step4 to step6 while Pointer < N-L.
Step4: Segment speech data into 10ms frame with 50% overlapping.
Step5: Compute speech frame energy, $E_n$.
Step6: Compare $E_n$ with maximum noise frame energy, $E_m$.

    If $E_n > E_m$
            Append the speech frame to the new file and
            Set Pointer = Pointer + L/2;
    Otherwise,
            Remove the speech frame and
            Set Pointer = Pointer + L/2.
```

The speech endpoint detection algorithm will read the noise data in specific file to determine the threshold of the maximum noise energy. This maximum noise energy level is used to set the threshold in the detection algorithm. First the speech signal is divided into 10ms frames, with 50% overlap. The detection algorithm goes through frame by frame, keeping the valid speech signal frame and throwing away the silence and pause frame according to condition of the threshold value. After processing all frames, all valid speech signal frames are joined together sequentially to create the new all-speech data for speaker feature extraction later.

The performance of the speech endpoint detection algorithm is illustrated in Figure-1(a) and in Figure-1(b). In each example, both original speech signal and the new speech signal with the removal of pauses and silences portion are presented. This endpoint detection algorithm is designed to step over very low signals and weak unvoiced sounds for better speaker identification performance. The frame energy is computed using the equation below:

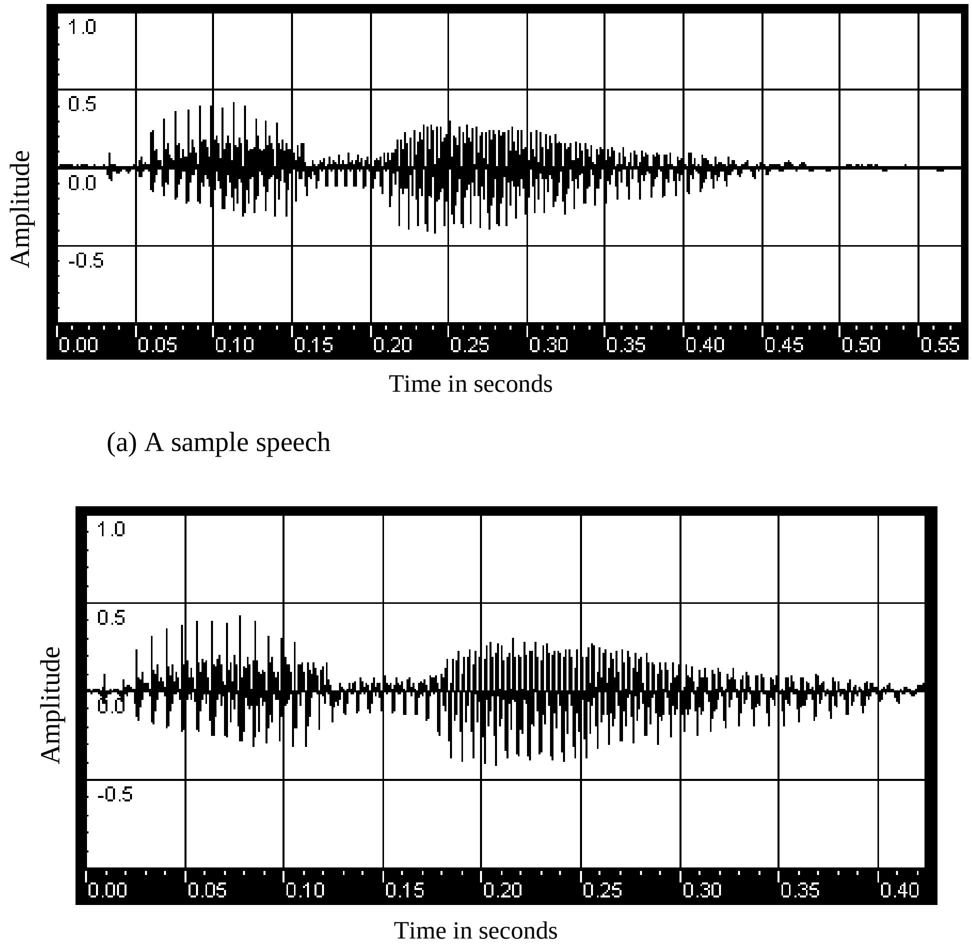
$$E_m = \sum_{n=m-L+1}^{m} s^2(n) \tag{1}$$

4. FEATURE EXTRACTION

Speech feature, namely the peak value of the frequency amplitude spectrum is obtained by averaging the magnitude of K-modified FFT sequence. That is,

$$A(k) = \tfrac{1}{K}\sum_{i=1}^{K} \left|S_i(k)\right| \tag{2}$$

where $S_i(k)$ is the spectra produced by the FFT procedure.



(a) A sample speech

(b) A sample speech with silence and pause removed

Figure-1 Plots of a speech signal Vs. no silence and pause of Bangla speech “Ami”

4.1. Fast Fourier Transform

Each of the Hamming winowed signals (*s(n)w(n)*, i.e., *s(n)* in Figure-4 and *w(n)* in Figure-3) is passed through the FFT procedure to produce the spectra of the windowed signal (Figure-2).

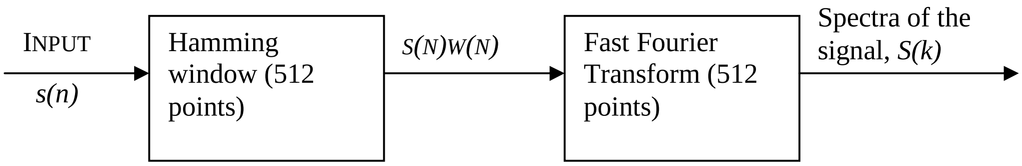

INPUT
s(n)
Hamming window (512 points)
S(N)W(N)
Fast Fourier Transform (512 points)
Spectra of the signal, S(k)

Figure-2 Block diagram of FFT for 46.4ms input signal.

As speech signal is sampled at a rate of 11025 samples per second (F_s=11025Hz). A 46.4ms window is used for short-time spectral analysis, and the window is moved by 23.2ms in consecutive analysis frame. Therefore,

□ Each section of speech is 512 samples in duration.
□ The shift between consecutive speech frames is 256 samples.
□ To avoid time aliasing in using the DFT to evaluate the short-time Fourier transform, we require the DFT size to be at least as large as the frame size of the analysis frame. Since we are using a radix-2 FFT, we require 512 point FFT to compute the DFT without time aliasing.

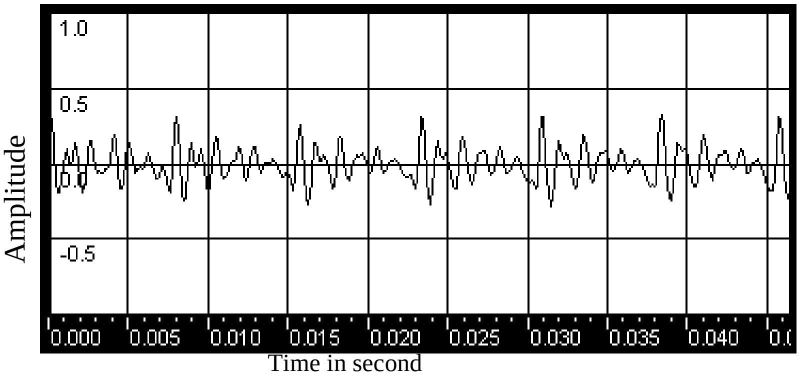


Figure-3 46.4ms segment of the input wave Ami

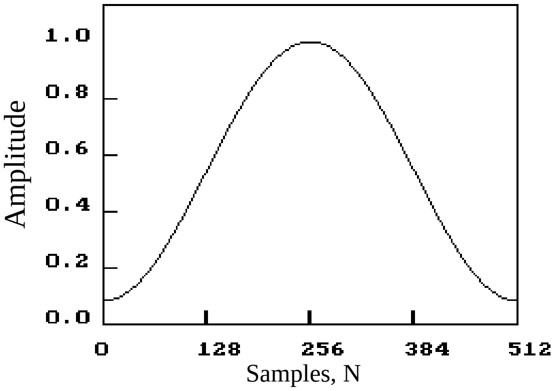


Figure-4 Hamming window (N=512)

4.2. Fast Fourier Transform

The speech features are acquired by signal processing technique. The time dependent frequency analysis (spectrogram) is used. The spectrogram computes the windowed discrete time Fourier transform of a signal using the sliding window [13]. Figure-5 shows a wave form representation of the Bangla utterance “Ami”. The spectrogram splits the wave signal into the segment and applies the windowed parameter to each segment. After this, it compute the discrete time Fourier transform of the each segment with the length equal to FFT length. The frequency amplitude spectrum is obtained by using the Eq.(2) and is shown in Figure-6. The sampling frequency of the wave signal is 11025Hz and the FFT length 512. The Hamming window is used and its length is kept equal to FFT length. The frame overlap used is 256 points i.e., 50%. The speech feature namely, the peak values of the frequency amplitude spectrum are obtained by taking the highest magnitude at the frequency interval of 128Hz up to a maximum frequency of 5160Hz (Figure-7).

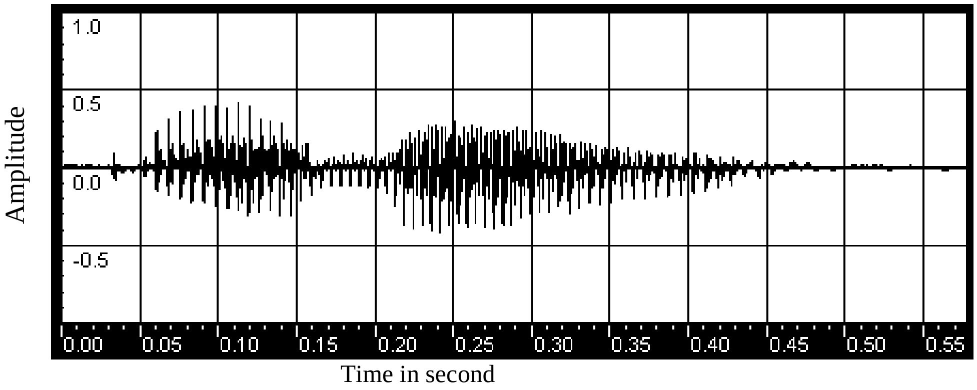


Figure-5 Input signal for the utterance “Ami”

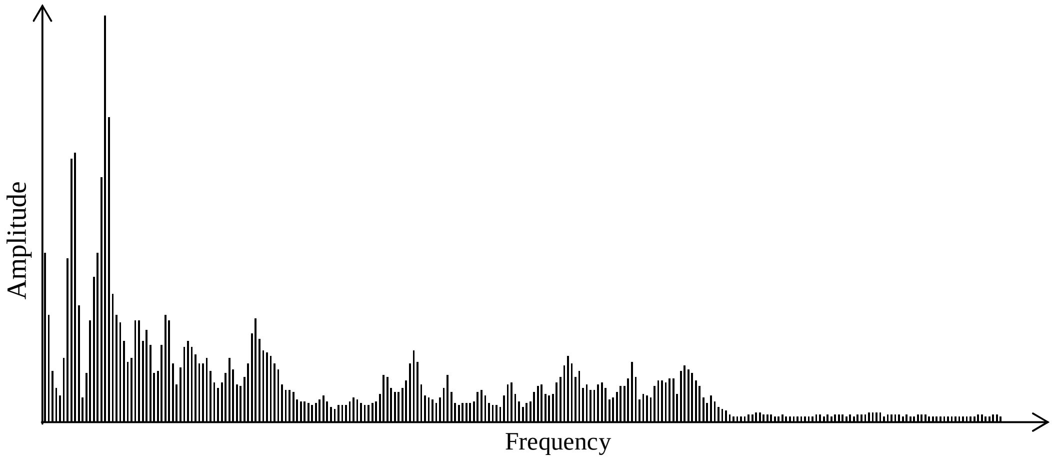


Figure-6 Amplitude of the frequency response

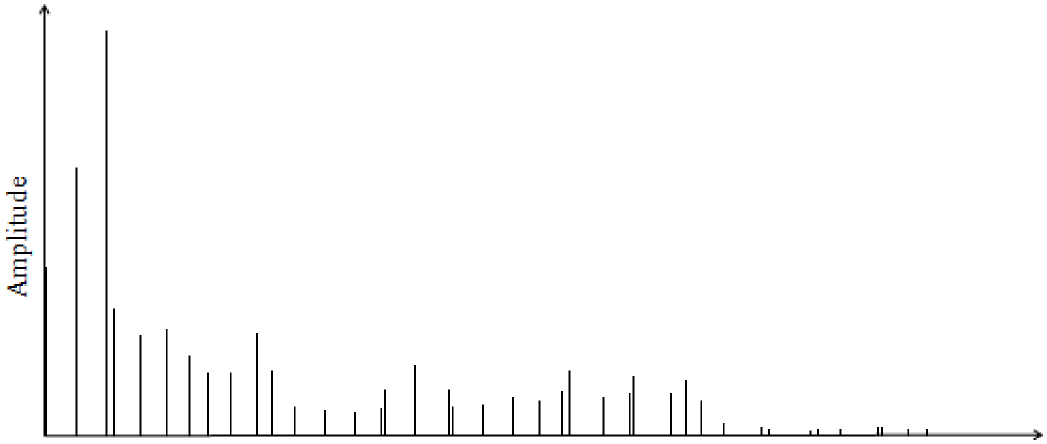


Figure-7 Peak amplitude of the frequency response

4.3. Training Set Generation

The data set used for training and testing the system consists of 80 utterances of selected Bangla words for ten speakers. The training data set (feature vectors) is generated by selecting one utterance for each speaker. A set of 40 features (peak value of frequency amplitude spectrum in different range of frequencies) is extracted for each speaker for both one syllable and two syllable words. These features are used to represent the input of a multilayer perceptron for learning purpose. Figure-8 illustrates the steps for generating the training data set for our ANN classifier.

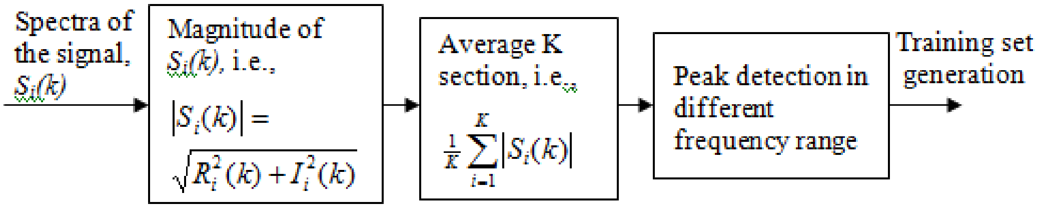

Spectra of the signal, $S_i(k)$
Magnitude of $S_i(k)$, i.e., $|S_i(k)| = \sqrt{R_i^2(k) + I_i^2(k)}$
Average K section, i.e., $\frac{1}{K}\sum_{i=1}^{K}|S_i(k)|$
Peak detection in different frequency range
Training set generation

Figure-8 Block diagram of training set generation

5. Artificial Neural Network Model for Speaker Identification

An attempt has been made to design a neural network as a pattern classifier. A neural network with three-layers, having forty neurons in first layer, eleven neurons in the intermediate hidden layer, and four neurons in the output layer has been used in this model for computer simulation. The model is illustrated in Figure-9.

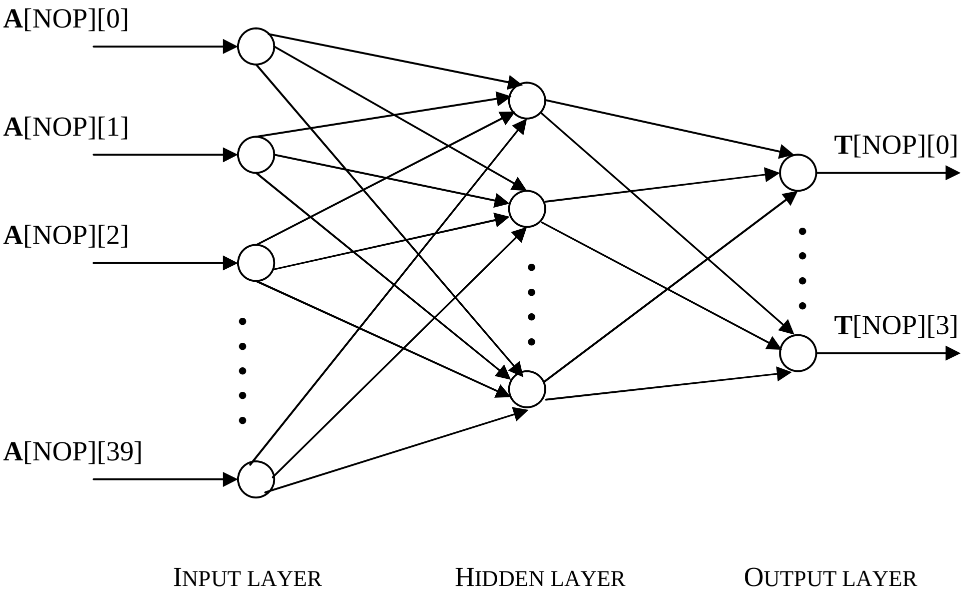


Figure-9 Topology of 40-input, 11-hidden, and 4-output units of a neural network

5.1. Topology

In this work, speech features for ten speakers are considered to be input patterns represented by the vectors **A**[NOP][i], where NOP (Number of person) = 0, 1, 2, ..., 9 and i (peak value of frequency amplitude) = 0, 1, 2, ..., 39, formed 10×40 pattern matrix.

Here NOP = 0 for the first speaker, NOP = 1 for the second speaker, and so on and i = 0 for the first input pattern element, i = 1 for the second input pattern element, and so on. Thus **A**[2][5] represent the fifth input pattern element for speaker 3. The output **T**[NOP][i] represent the target output, where NOP = 0, 1, 2, ..., 9 and i = 0,1,2,3. The weight between input and hidden layer has been denoted by $\mathbf{W}_{ij}$ (from i-th input processing element (PE) to j-th hidden processing element) and hidden to output weight is denoted by $\mathbf{W}_{jk}$ (from j-th hidden PE to k-th output PE). The topology of this network is shown in Figure-9.

5.2. Error Backpropagation Learning Algorithm

Training a network is equivalent to finding proper weights and thresholds values for all the connections such that a desired output is produced for corresponding input . The error backpropagation algorithm [14,15] has been used to train this Multi-Layer Percentron (MLP) network. At first the weight vectors $\mathbf{W}_{ij}$ and $\mathbf{W}_{jk}$ and the threshold values for each processing element's in the network were to be initialized with small random numbers. Then the algorithm is learned to find a proper weight and threshold values.

6. TRAINING PROCEDURE FOR SPEAKER IDENTIFICATION

The speaker identification system read each train utterance of each speaker from the train speaker data set of 10 speakers. Then 11025Hz, 8-bits, monoral utterance signal *s(n)* is passed through the speech endpoint detection algorithm to remove pauses, silences, and weak unvoiced sound signals. The resulting signal is then passed separately to a 512 points Hamming window, 512 points Blackman-Harris window (3-term).

Next, the spectral analysis is performed to obtained feature vectors (40-peak frequency amplitude in different frequency range). These feature vectors are then fed to a MLP to learn the network. Finally, proper weight and threshold values are saved in files for identification purpose. The learning algorithm is consisted of the following steps:

1. One utterance is selected for each speaker from the speaker data set.
2. The starting point and endpoint of each speech signal is determined for each speaker using the speech endpoint detection algorithm.
3. Speaker feature is extracted from each utterance that is used to create the training vector for each speaker.
4. The training vector is generated for all of the ten speakers.
5. These training vectors are then fed into a MLP to train the network.
6. Finally, the common weight and threshold values for all speakers are stored.

7. TESTING PROCEDURE FOR SPEAKER IDENTIFICATION

The testing procedure for speaker identification read an unknown utterance from the test speaker data set. The speech signal *s(n)* is passed through the speech endpoint detection algorithm to valid speech signal. The resulting signal is then passed separately to a 512 points Hamming, 512 points Blackman-Harris window. The speaker features are then extracted from the speech signal and fed to the MLP network. The network uses the predefined knowledge (weight and threshold values) to calculate error for each speaker. The identification system then selects the smallest error value. This error value is compared with a threshold and a decision of whether to accept or reject the speaker is made.

8. EXPERIMENTAL RESULT

Experiment has been done to observe two things the behaviour of the neural network and the speaker identification accuracy rate. The behaviour of the network has been observed with respect to different parameters used in the proposed neural network model. The effect of the hidden layer units has been studied also.

The speaker identification accuracy has been tested depending on the network behaviour. Hamming window and Blackman-Harris window were used to investigate the better identification accuracy.

8.1. Network Behaviour Study

It was wise to test some simple cases to verify the system. For this purpose, a simple train set was produced which represent digit 0 to 9 for a seven segment with the output patterns to classify these digits and is given in Table-1. The network was learned with this pattern. Next the test pattern set was same as the train set. It was seen that the network learns all the patterns in a few cycles and successfully classifies them.

Table-1 A simple test pattern and their respective output

Input	Digit	Output
0111111	0	0000
0000110	1	0001
1011011	2	0010
1001111	3	0011
1100110	4	0100
1101101	5	0101
1111101	6	0110
0000111	7	0111
1111111	8	1000
1101111	9	1001

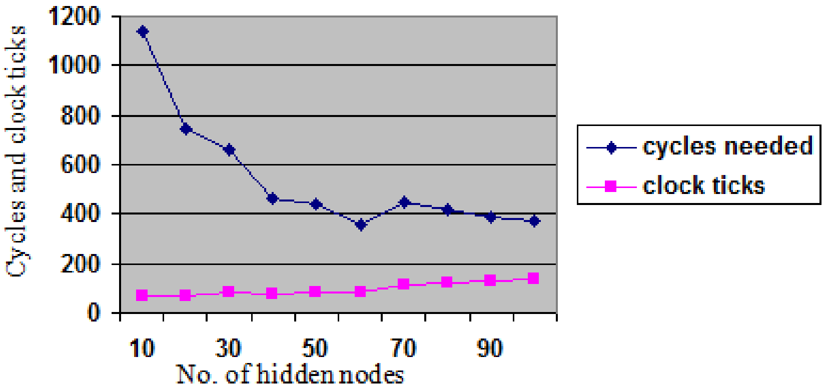


Figure-10 Hidden nodes vs. learning cycles and learning time.

To see the effect of number of hidden layer nodes the same train and the test pattern set were used to train and test the network. Only one hidden layer is used in this case. Time needed (clock ticks) to learn for a given error tolerance and the cycles of phases are noted. The result is summarized in Figure-10.

The effect of network parameters such as learning rate and spread factors was observed. Both the learning rate and spread factors are real numbers and in the range of 0 to 1. The effect of learning rate on learning time was very much dependent on input pattern. If the inter-pattern distance is large, a high learning rate (η>0.7) swiftly converges the network. For small inter-pattern distance a small value of learning rate is needed unless the weight bellow up. The learning time as a function of learning rate is shown in Figure-11. Here the number of hidden layer is 40 and the spread factor is fixed at 0.7.

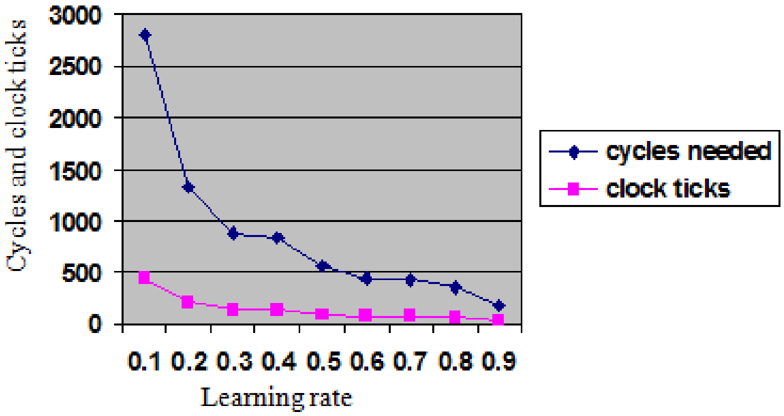


Figure-11 Cycles and clock ticks as a function of learning rate

The parameter spread factor *k*, controls the “spread” of the sigmoid function. It also acts as an automatic gain control, since for small input signals the slope is quite steep and so the function is changing quite rapidly, and producing a large gain. For large inputs, the slope and thus the gain is much less. The effect of spread factor on the network behavior is shown in Figure-12. Here the learning rate is fixed to 0.9 and the number of hidden layer is 40.

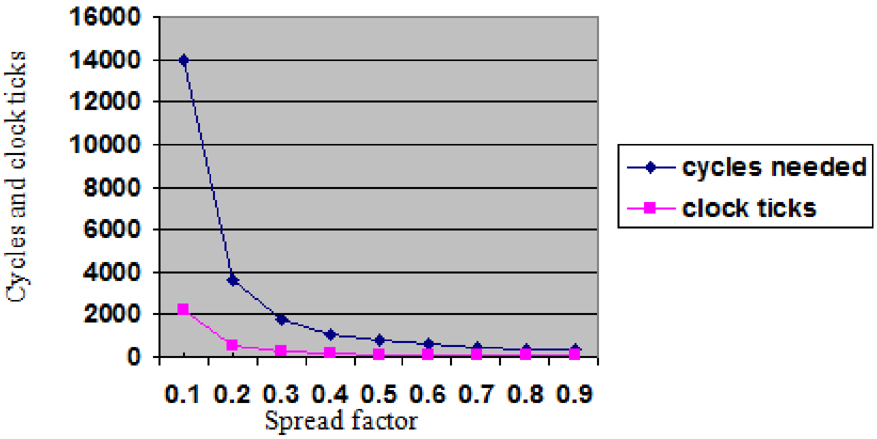


Figure-12 Cycles and clock ticks as a function of spread factors

8.2. Speaker Identification Accuracy Rate

For the present study the utterances of 10 speakers were recorded using the audio wave recorder. Each of them uttered the prominent selected Bangla words, “Ami” for one syllable word and “Bangla-desh” for two syllable word. Therefore, two sets of test data set are produced. Each set contains forty samples of ten speakers. In the following section, speaker identification accuracy rate using different method is considered.

8.2.1. Use of Hamming Window

The input (peak values of frequency amplitude) of the ANN is obtained by the frequency analysis for the given input Bangla words using Hamming window. The detail of the ANN was specified by representing the input in the form of matrix. The error goal is less than 0.01 for this network.

The number of iterations in which the network reached the specified error goal is equal to 1,34000 for one syllable word “Ami”. The learning rate of the network is set to $\eta_1=\eta_2=0.9$ and the spread factors is $k_1=k_2=0.5$. For both cases the error tolerance level is fixed at 0.05. The speaker identification accuracy based on the speaker features using the Hamming window is presented in Table-2.

Table-2 Speaker identification accuracy using Hamming window

	One Syllable Word	Two Syllable Word
Sample Utterances	40	40
Correct Identification	33 (82.5%)	26 (65%)
False Inclusion	1 (2.5%)	4 (10%)
False Rejection	6 (15%)	10 (25%)

8.2.2. Use of Blackman-Harris Window (3-term)

The speaker features that are extracted by applying the Blackman-Harris window are used to train the network. The number of iterations (cycles) in which the network reached the specified error goal is equal to 71,000. The network parameter are selected to $\eta_1=\eta_2=0.9$, and $k_1=k_2=0.5$. The identification accuracy based on the sample size is given in Table-3. The error tolerance level is selected to 0.05 or 5% for this identification score.

Table-3 Speaker identification accuracy using Blackman-Harris window

	One Syllable Word	Two Syllable Word
Sample Utterances	40	40
Correct Identification	26 (65%)	24 (60%)
False Inclusion	3 (7.5%)	5 (12.5%)
False Rejection	11 (27.5%)	11 (27.5%)

8.3 Discussion

The network we have proposed depends on the numbers of hidden layer nodes. If the number of hidden layer nodes increases, the number of iterations in which the network reaches the specified error goal decreases. Since the computational load of the network increases with the increasing of the hidden layer nodes, so the network takes more time (clock ticks) to reach the error goal. It is seen that 10 nodes in hidden layer take only 68 clock ticks, whereas 100 nodes in hidden layer take 133 clock ticks (Figure-11). The effect of learning rate is very much dependent on input patterns. The learning rate of *0.9* ($\eta=0.9$) swiftly converges the network and provides the faster learning. The spread factor of 0.5 provides a good nonlinear smoothed function in this speaker identification system.

The identification performance using Hamming window and Blackman-Harris (3-term) window for both one syllable and two syllable words are presented in Table-2 and Table-3. The best identification score is 82.5%, which is obtained for one syllable word “Ami” using Hamming window. There is no false inclusion for 9 speakers. The highest false inclusion error is 12.5%, which is obtained for two syllable word using the Blackman-Harris window and have the high security risk.

9. CONCLUSION

A model of simple speaker identification for Bangla speech using artificial neural network and digital signal processing technique is described in this thesis. In this research, we simulate the artificial neural network (ANN) model for speaker identification system in Bangla. The spectral information used in this research is affected by the sound pressure level of the speaker, i.e., the distance between a speaker and microphone is important. Thus, some kind of normalization is required to eliminate the influence of any variable transmission characteristics on the spectral data. The current system can be termed as a language independent speaker identification system. It will act as speaker identification in English if the training set and the testing set are in the form the English. So the output of the system depends on the input training set, and the identification process is same for all languages.

The speech parameter used in this model, i.e., peak value of the frequency amplitude in the different range of frequencies consists of the sufficient information about the speakers and further varies among the speakers. Using the proper normalization on speech signal this system can be used to identify speaker more accurately. By adopting some filtering method (i.e., preemphasis, noise elimination etc.) prior to signal processing and by using better method in feature extraction (i.e., LPC instead of FFT), the performance of the system can be improved. In near future, it is very important to extend this technique so that more accurate and real-time speaker identification becomes possible.

REFERENCES

[1] Lawrence R. Rabiner and Ronald W. Schafer (1978), “Digital Processing of Speech Signal,” Prentice-Hall Inc., Englewood Cliffs, New Jersey.

[2] Michael Tran, “An Approach to A Robust Speaker Recognition System,” A Ph.D. Thesis Paper, Dept. of Electrical Engineering, Virginia Polytechnic Institute and State University.

[3] B.S. Atal (1976), “Automatic Recognition of Speaker from Their Voices,” Proc. IEEE, Vol. 64, No. 4.

[4] M.R. Sambur (1976), “Speaker Recognition using Orthogonal Linear Prediction,” IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP□24.

[5] M. Shridhar and M. Baraniecki (1979), “Accuracy of Speaker Verification Via Orthogonal Parameters for Noise Speech,” Proc. Int. Conf. Acoust., Speech and Signal Processing.

[6] M. Savic and S.K. Gupta (1990), “Variable Parameter Speaker Verification Based on Hidden Markov Modeling,” Proc. Int. Conf. Acoust., Speech and Signal Processing.

[7] Y.C. Zheng and B.Z. Yuan (1988), “Text-Dependent Speaker Identification using Circular Hidden Markov Models,” Proc. Int. Conf. Acoust., Speech and Signal Processing.

[8] S. Vela and Hema A. Murthy (1998), “Speaker Identification A New Model Based on Statistical Similarity,” Proc. Int. Conf. on Computational, Linguistics, Speech and Document Processing (ICCLSDP), Calcutta, February 18-20.

[9] R. Srikanth, Dr. Y.G. Srinivasa (1998), “Text-Dependent Speaker Verification using Temporal Identity Mapping Neural Network,” Proc. ICCLSDP, Calcutta, February 18-20.

[10] A.H. Waibel and J.B. Hampshire H, “Neural Network Application to Speech,” School of Computer Science, Carnegie Mellon University.

[11] M.M. Rashid, M. Meftauddin and M.N. Minhaz (1998), “Speech Password Security System using Bangla Neumerals as Fixed Text,” Int. Conf. on Comp. and Info. Tech. Dhaka, December 18-20.

[12] M.N. Minhaz, M.S. Rahamn and S.M. Rahamn (1998), “Feature Extraction for Speaker Identification,” Int. Conf. on Comp. and Info. Tech., Dhaka, December 18-20.

[13] Rafael C. Gonzalez, Richard E. Woods (1998), “Digital Image Processing,” Addison-Wesley.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams (1986), “Learning Internal Representation by Error Backpropagation,” Vol. 1, pp□318-362, MIT Press, Cambridge, MA.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams (1986), “Learning Representation by Backpropagating Errors, Nature.

[16] M. A. Bashar, Md. Tofael Ahmed, Md. Syduzzaman, Pritam Jyoti Ray and A. Z. M. Touhidul Islam, “Text-Independent Speaker Identification System Using Average Pitch And Formant Analysis”, International Journal on Information Theory (IJIT), Vol. 3, No. 3, pp 23-30.

AUTHORS

Dipankar Das received his B.Sc. and M.Sc. degree in Computer Science and Technology from the University of Rajshahi, Rajshahi, Bangladesh in 1996 and 1997, respectively. He also received his PhD degree in Computer Vision from Saitama University Japan in 2010. He was a Postdoctoral fellow in Robot Vision from October 2011 to March 2014 at the same university. He is currently working as an associate professor of the Department of Information and Communication Engineering, University of Rajshahi. His research interests include Object Recognition and Human Computer Interaction.