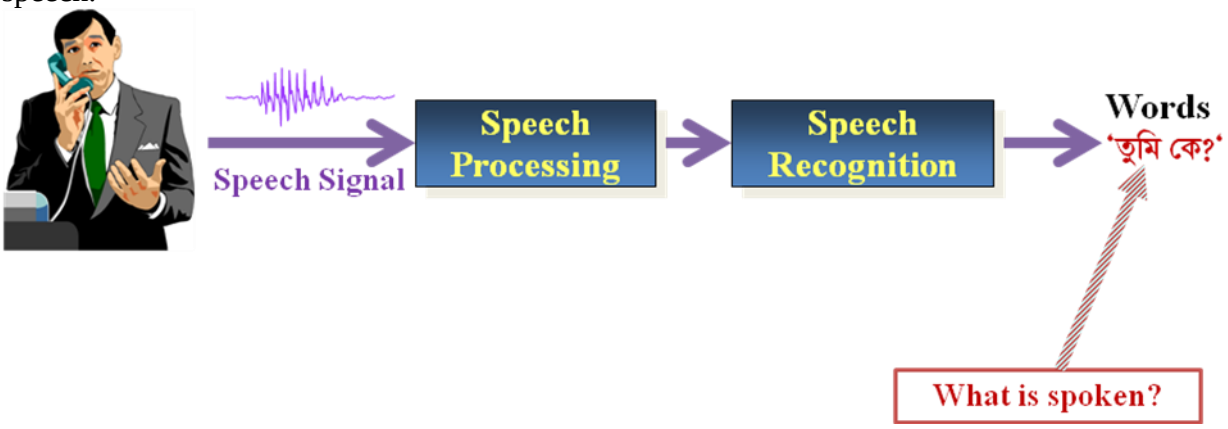
Summary Paper On

Development of Speech Recognition System for Continuous Bangla Speech

GOAL

Our main goal is to automatically extract the string of spoken words from the continuous Bangla speech.



OUTLINES

- ☐ Introduction
- ☐ Research Methodology
- ☐ Speech Segmentation and classification
- ☐ Some approaches of dynamic thresholding on speech segmentation
- ☐ Speech Recognition using feedforward neural network
- ☐ Conclusion

INTRODUCTION

Speech recognition is a computer application that lets people control a computer by speaking to it. It allows computers to understand spoken language. Our main interest is to recognize continuous Bangla speech using feedforard neural network with some improved and faster back-propagation algorithms, like BP with Momentum, Variable learning rate BP, Resilient BP, Conjugate gradient and Levenberg-Marquardt Algorithms. Our research works also include some speech segmentation techniques. Here, we have introduced some approaches of dynamic thresholding on continuous Bangla speech segmentation, like thresholding using k-means algorithm, fuzzy c-means algorithm and Otsu's thresholding algorithm.

In continuous speech, sentences of multiple words tend to be pronounced with little or no gaps between words. Continuous speech is very much difficult to process or recognize without segmentation. An automatic continuous speech recognition system includes-

- ◘ Speech Segmentation
- ◘ Speech Recognition

RESEARCH METHODOLOGY

A collection of techniques are used in this research, as shown in Figure-1.

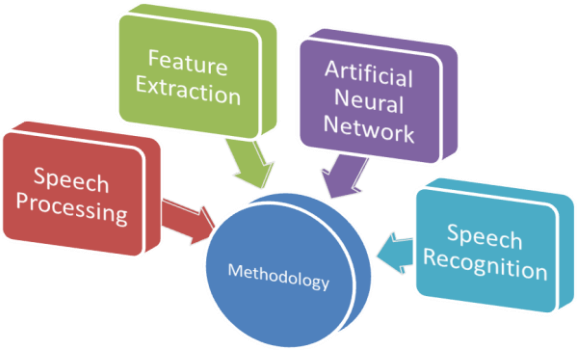


Figure-1. Research Methodology

SPEECH SEGMENTATION

Speech recognition system requires segmentation of speech signal into discrete, non-overlapping acoustic units. It can be the *segment*, *phone*, *syllable*, *word*, *sentence* or *dialog turn* level. Word is the preferred and natural units of speech because word units have well defined acoustic representation. So, we have been chosen word as our basic unit. To segment our continuous speech, we have short-time energy features approaches and some dynamic thresholding approaches.

In general, there are two kinds of segmentation-

- phonemic segmentation
- syllable-like unit segmentation (words or sub-words)

In our research work, Syllable-like unit segmentation by end point detection technique is used.

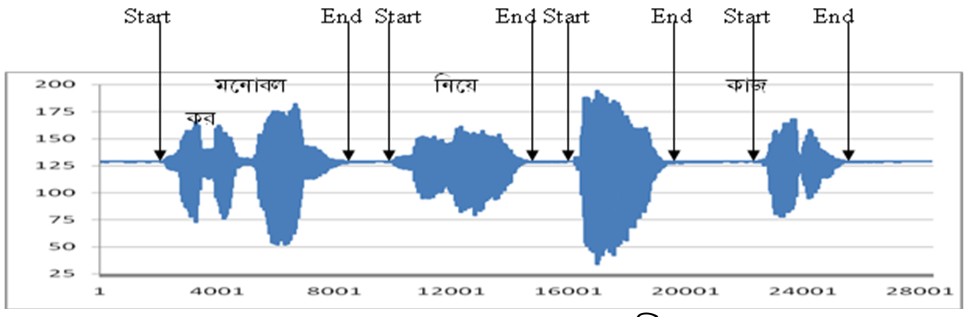


Figure-2. Bangla speech sentence "মনোবল নিয়ে কাজ করো".

Speech Segmentation using Short-time Energy Feature

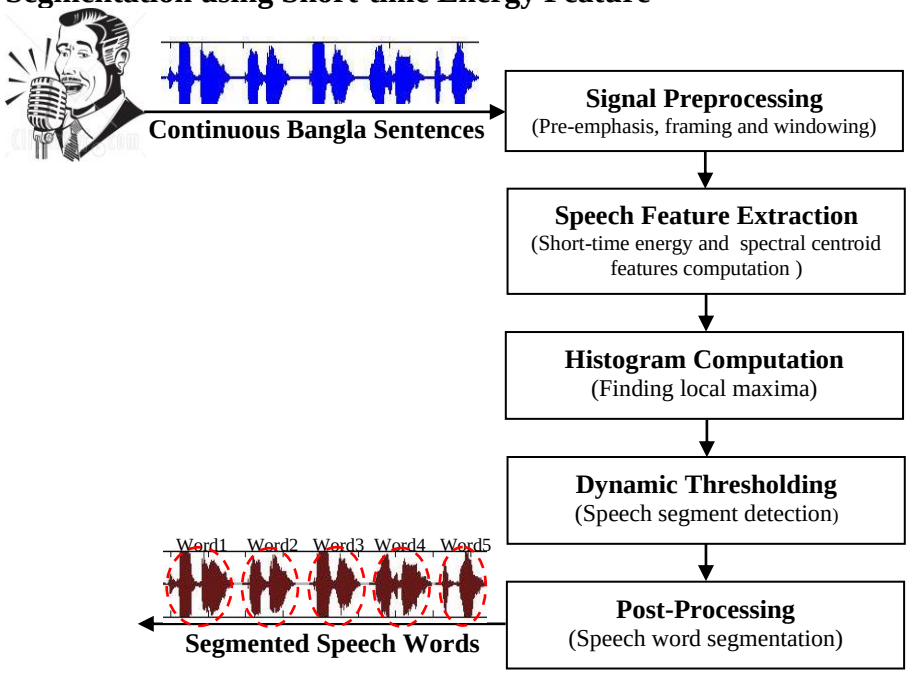


Figure-3. Automatic Speech Segmentation system using short-time energy approaches.

Short-Time Signal Energy

Energy is used to discover voiced sounds. Voiced sounds have higher energy than silence/un-voiced in continuous speech. The short-time energy function of a speech frame with length N is defined as:

$$E_n = \frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2$$

The short-term root mean squared (RMS) energy of this frame is given by:

$$E_{n(RMS)} = \sqrt{\frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2}$$

where: x(m) is the original signal and w(m) is a window function.

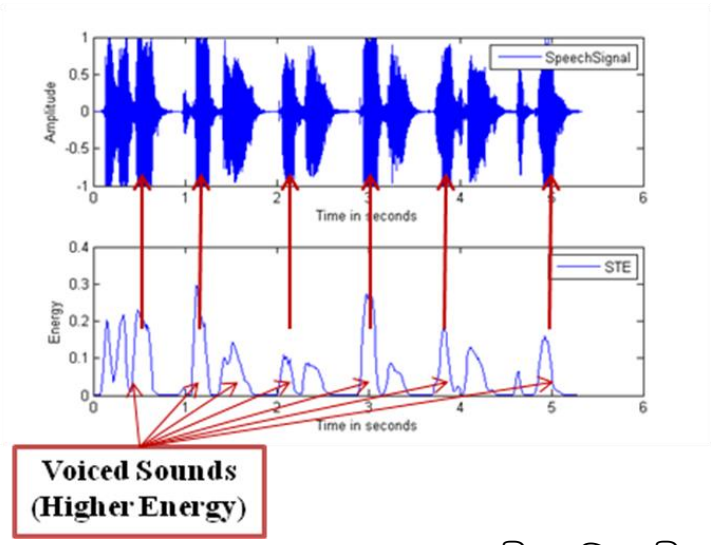


Figure-4. The energy signal of speech sentence: "আমাদের জাতীয় কবি কাজী নজরুল ইসলাম".

Spectral Centroid

It indicates where the "center of gravity" of the spectrum is. High values correspond to "brighter" sounds. The spectral centroid (SC_i) of the i-th frame is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m) X_i(m)}{\sum_{m=0}^{N-1} X_i(m)}$$

Here, f(m) represents the center frequency of i-th frame with length N and $X_i(m)$ is the amplitude corresponding to that frame in DFT spectrum.

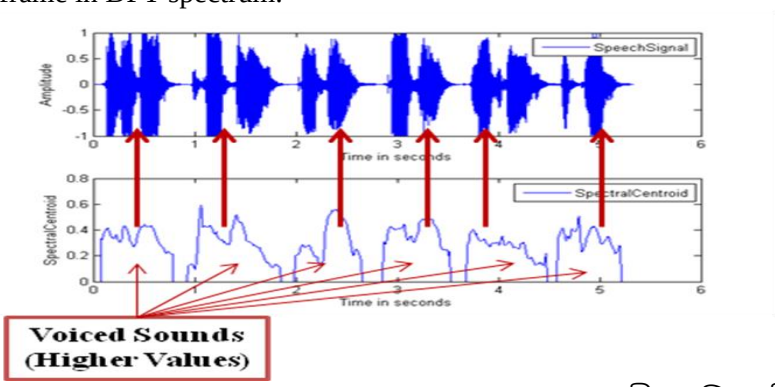


Figure-5. The spectral centroid energy signal of speech sentence: "আমাদের জাতীয় কবি কাজী নজরুল ইসলাম".

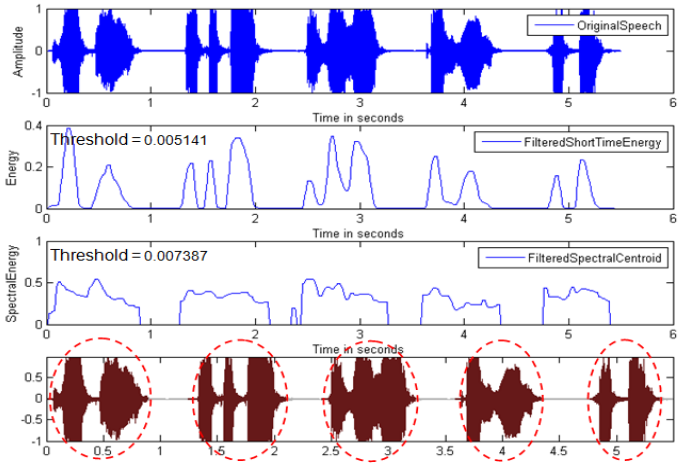


Figure-6. The segmentation results for a speech sentences "জাতীয় পতাকার ডিজাইনার কামরুল হাসান" which contains 5 (five) speech words, marked by dashed circles.

Segmentation Results

To test the segmentation performance, 100 Bangla sentences were recorded from 5 male speakers. The segmentation system achieved the average accuracy of 95.9%.

Table-1. Segmentation Result of short-time energy appraoches

Speaker ID	No. of Sentences	No. of words present	No. of properly segmented words	Seg. accuracy (%)
S1	100	656	627	95.58
S2	100	656	634	96.65
S3	100	656	603	91.92
S4	100	656	638	97.26
S5	100	656	644	**98.17**
Total	**500**	**3280**	**3146**	**95.91**

Dynamic Thresholding on Speech Segmentation

Some approaches of dynamic thresholding on speech segmentation:

- K-Means clustering algorithm
- FCM Clustering Algorithm
- Otsu's Method

The thresholding technique is used to separate voiced regions from silence/un-voiced in continuous speech. The Blocking Black Area technique is used to identify/detect speech word boundary.

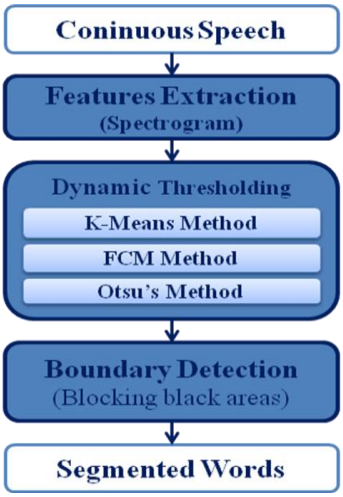


Figure-7. Block diagram of Speech Segmentation using dynamic thresholding approaches.

Spectrogram Feature

Speech spectrogram shows how the spectral density of a signal varies with time. The spectrogram of a speech frame is computed by the following equation:

$$spectrogram(t,w) = |STFT(t,w)|^2 \; ; \; w \, is \, a \, window.$$

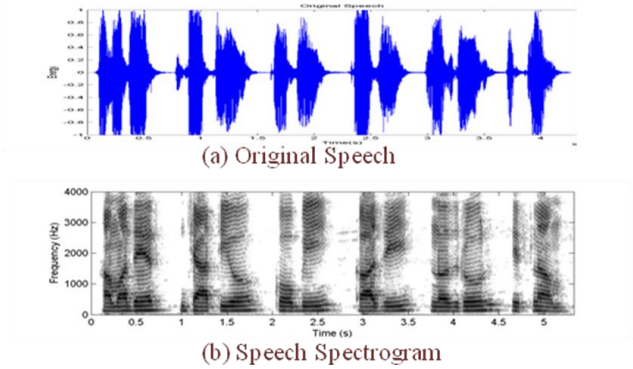


Figure-8. Speech Spectrogram.

Thresholding using K-Means Clustering Algorithm

Here, the modified algorithm is used to compute the threshold from speech spectrogram. The algorithm classifies a given data set into a certain number of clusters (i.e., k clusters) based on distance measures and to define k-centers, one for each cluster and finally computes the threshold. This algorithm is an iterative process until no new moves of data, as shown in Figure-9.

Thresholding using FCM Clustering Algorithm

FCM allows clusters to be fuzzy sets, so that each object belongs to all clusters with varying degrees of membership and the following restriction: The sum of all membership degrees for any given data point is equal to 1. It works by assigning membership to each object on the basis of distance between the cluster center and the object. More the object is near to the cluster center more is its membership. The algorithm is an iteration process, as shown in Figure-10. After each iteration, membership and cluster centers are updated, and finally compute a threshold from the cluster centroids.

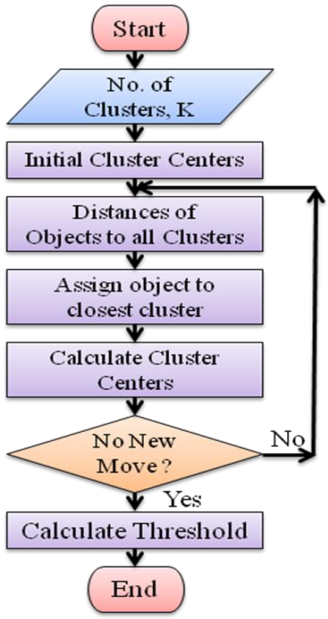


Figure 9. K-Means Clustering Algorithm - Calculating the desired threshold.

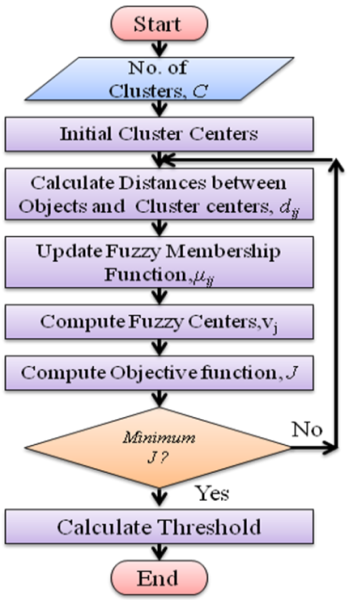


Figure 10. FCM Clustering Algorithm - Calculating the desired threshold.

OTSU'S THRESHOLDING METHOD

Thresholding is the simplest method of image segmentation; can be used to create binary image from gray scale image. The algorithm assumes that the image is composed of two basic classes: Foreground and Background. It then computes an optimal threshold value that minimizes the weighted within class variance; also maximizes the between class variance of these two classes.

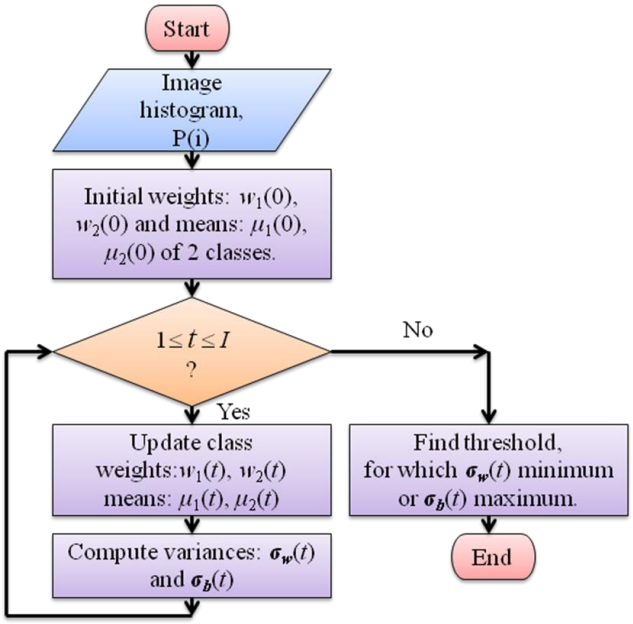


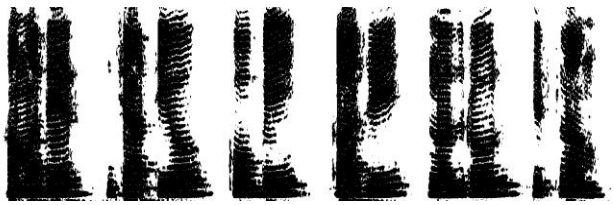
Figure 11. Otsu's Thresholding Algorithm – Converting binary image.

Thresholded Images

We used MATLAB's 3-class '`kmeans`' function and 3-class '`fcm`' function. Both functions return the *centroids* of three clusters. The average of 1st and 2nd largest *centroids* is calculated as the desired threshold value. We also used MATLAB's '`graythresh`' function, that uses Otsu's thresholding method, returns a level or threshold value for which the intra-class variance of the black and white pixels is minimum. The output image replaces all pixels in the input image with luminance greater than or equal to the threshold with the value of 1 (fully white) and less than threshold with 0 (fully black). Figure-12 shows the thresholded spectrogram image of the speech sentence ('আমাদের জাতীয় কবি কাজী নজরুল ইসলাম').

(a) K-Means thresholded Image

(b) FCM thresholded Image



(c) Otsu's thresholded Image

Figure 12. Thresholded Spectrogram Images of Speech signal 'আমাদের জাতীয় কবি কাজী নজরুল ইসলাম'

Blocking Black Area Method

Here, we used a new approach '*Blocking Black Area*' method in the thresholded spectrogram image that produces rectangular black boxes in the voiced regions of the speech sentence, as shown in Figure-13. The method works as follows:

(a) Thresholded Spectrogram Image from Speech Signal

(b) After applying Blocking Black Area Method

Figure 13. Effect of applying Blocking Black Area Method – Producing rectangle black boxes in voiced regions of speech signal.

Boundary Detection

We used MATLAB '*regionprops*' function to identify each rectangular object. The function's '*Extrema*' parameter is used to detect the start (*bottom-left*) and end (*bottom-right*) point of each object.

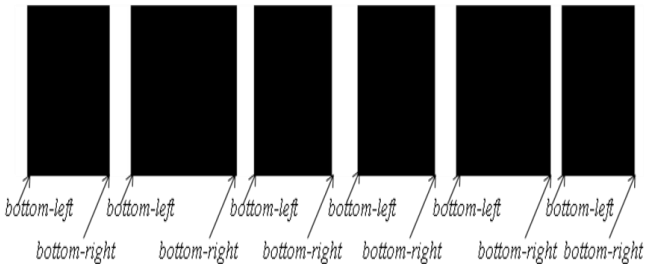


Figure 14. Star and End point Detection of rectangular object.

Final Speech Segments

Each rectangular black box represents a speech word. After detecting the start and end points of each black box, the word boundaries of the original speech sentence are marked automatically by these two points and finally we have cut the word segments from the speech sentence. Figure-15 shows that 6 (six) black boxes represent 6 (six) word segments in the speech sentence **'আমাদের জাতীয় কবি কাজী নজরুল ইসলাম'**.

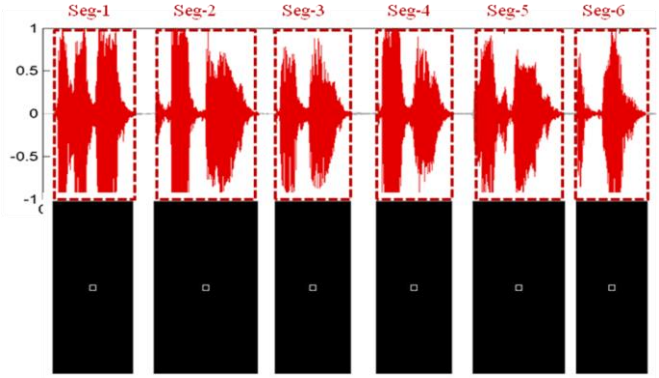


Figure 15. 6 (six) detected word segments in the speech sentence 'আমাদের জাতীয় কবি কাজী নজরুল ইসলাম'.

EXPERIMENTAL RESULTS

To test the performance of the segmentation approach, 100 sentences from 5 speakers are used for segmentation. The developed system achieved the average segmentation accuracy 95.55% with K-Means; 96.2% with FCM and 90.58 with Otsu's method.

Table 2. The details segmentation results

Speaker ID	No. of Sentences	No. of words present	No. of properly segmented words (#) and accuracy (%)						Avg. Accuracy
			K-Means		FCM		OTSU		
			#	%	#	%	#	%	%
S1	100	656	621	**94.67**	629	**95.88**	517	**78.81**	**89.79**
S2	100	656	621	**94.67**	622	**94.82**	601	**91.62**	**93.70**
S3	100	656	622	**94.82**	624	**95.12**	612	**93.29**	**94.41**
S4	100	656	631	**96.18**	643	**98.02**	619	**94.36**	**96.19**
S5	100	656	639	**97.41**	637	**97.10**	622	**94.82**	**96.44**
Total	500	3280	3134	**95.55**	3155	**96.19**	2971	**90.58**	**94.11**

SPEECH CLASSIFICATION

Syllable-based classification: Classify the segmented speech words/sub-words into different classes based on the number of syllables. The classification algorithm works as follows-

- Compute spectrogram image from segmented words/sub-words.
- Identify syllables (i.e., brighter voiced regions) in the segmented speech using-
  - Otsu's thresholding method and
  - Blocking black area method
- Finally, count the number of syllables of speech words and assign to the corresponding class:
  - Mono (1)-syllabic words belong to **Class-1**
  - Di (2)-syllabic words belong to **Class-2**
  - Tri (3) or poly (3+)-syllabic words belong to **Class-3**

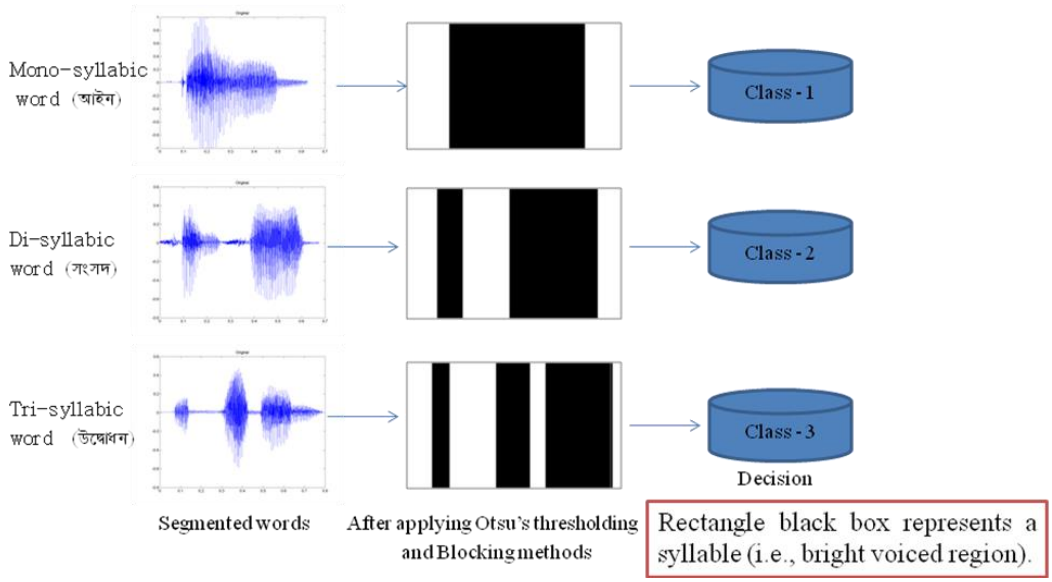


Figure-16. Example of classified words based on syllables.

Table-3. Details Result of Syllable-based Classification

Speaker ID	No. of Segmented words/sub-words	No. of properly classified segments	No. of segments belong to **Class-1**	No. of segments belong to **Class-2**	No. of segments belong to **Class-3**	Accuracy (%)
S1	687	641	211	314	162	93.30
S2	621	577	171	280	170	92.91
S3	615	566	167	260	188	92.03
S4	632	558	192	279	161	88.29
S5	649	587	214	288	147	90.45
Total	**3204**	**2929**	**955**	**1421**	**828**	**91.42**

SPEECH RECOGNITION PROCESS

The block diagram of speech recognition system is given below:

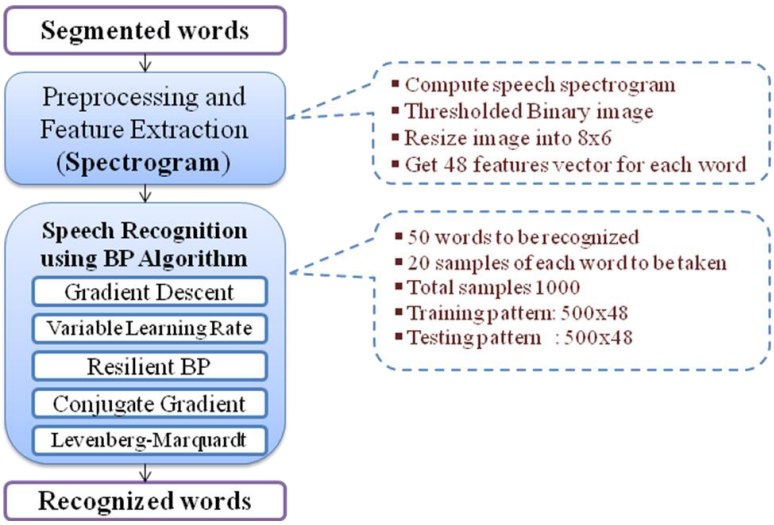


Figure-17. Block diagram of Speech Recognition System

BP Algorithm

The Backpropagation algorithm is the supervised learning algorithm, where the target of the output is given. It does gradient descent as the opposite to the gradient of the error between target and actual output of the net.

- Error, $E = 1/2 \sum_k (t_k - o_k)^2$
- Weight change, $\Delta w = -\eta \nabla E(w) = -\eta \frac{\partial E}{\partial w}$; η *is the learning rate.*

The training of a net by BP algorithm involves three stages:

- Feedforward Pass: The feedforward of the input training pattern.
- Backward Pass: The calculation and backpropagation of associated errors.
- Weights Adjustment : After all of the δ factors have been determined, the weights of all layers are updated simultaneously.

The Flowchart of BP Algorithm is given below:

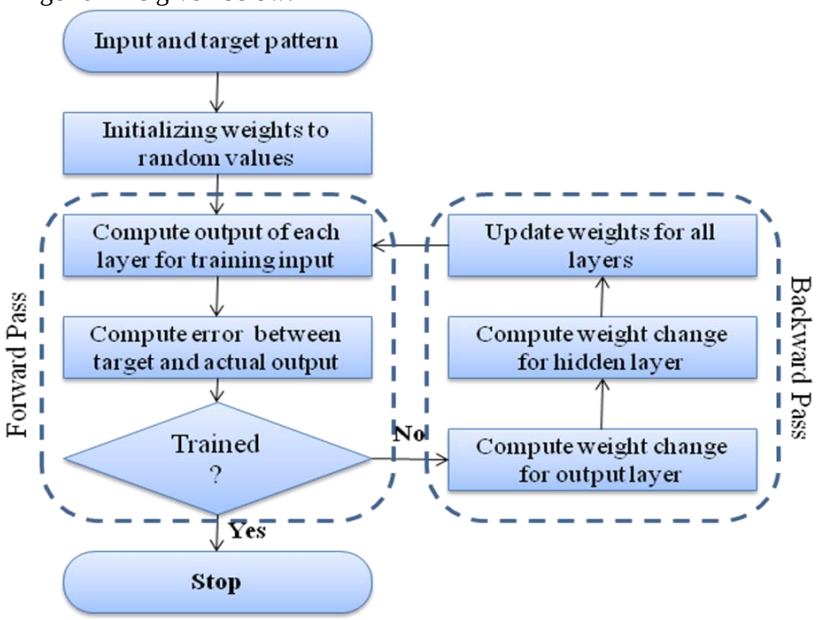


Figure-18. Flowchart of BP Algorithm.

Practically, the convergence obtained from standard BP learning algorithm is very slow with the increasing of the number of nodes. Comparison study on some improved and faster BP algorithms:

- BP with Momentum and Variable learning rate
- Resilient BP
- Conjugate gradient and
- Levenberg-Marquardt Algorithm

We used Matlab Neural Network Toolbox 7.12.0 to create, train and simulate the BP feedforward neural network. The net has 48 input (each word 48-features data), 40 hidden and 50 output neurons. It is used to recognize 50 different

words at a time. Both hidden and output layer used 'logsig' (log sigmoid) transfer function. A set of 500 samples of 50 words (10 samples of each word) is used as training data and a set of another 500 samples of 50 words is used as testing data.

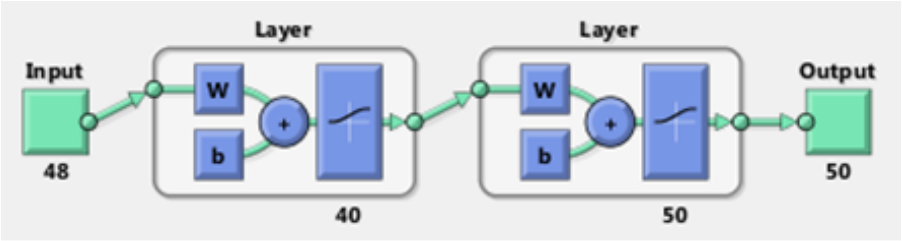


Figure-19. Practical Neural Network Configuration.

Table-4. Recognition Result Analysis.

Algorithms	Recognizing 10 words Pt : 100x48 Ptest : 100x48			Recognizing 20 words Pt : 200x48 Ptest : 200x48			Recognizing 30 words Pt : 300x48 Ptest : 300x48			Recognizing 40 words Pt : 400x48 Ptest : 400x48			Recognizing 50 words Pt : 500x48 Ptest : 500x48		
	Epoch	Time (s)	Recog. Accuracy (%)	Epoch	Time (s)	Recog. Accuracy (%)	Epoch	Time	Recog. Accuracy (%)	Epoch	Time	Recog. Accuracy (%)	Epoch	Time	Recog. Accuracy (%)
Gradient Descent BP ('traingd')	5000	34	98	5000	31	98	5000	36	94.33			N/A			N/A
BP With Momentum ('traingdm')	5000	38	98	5000	31	98	5000	36	93.33			N/A			N/A
BP with Variable LR ('traingdx')	231	04	94	1070	18	97	1866	22	94.67			N/A			N/A
Resilient BP (trainrp)	38	1	96	82	01	96.5	216	2 s	90.30	483	4s	84.50	521	6s	83.2
Conjugate Gradient BP ('trainscg')	32	1	98	68	01	97	84	1 s	94.30	161	2s	93.25	119	7s	90.2
Levenberg-Marquardt BP ('trainlm')	5	7	94	9	67	93.5	16	2m 43s	92.64	11	6m 8s	90	20	24m 13s	90

LIMITATIONS

Some words are not properly segmented. No or very little gap between two successive words causes two or more words in a single segment. Also the gap within a word causes sub-word segmentation. Segmentation using fuzzy logic takes much time than other approaches. In feature extraction, selecting a window function is not a simple task. In classification, some same words belong to different classes due to the variations of speaker's articulation. Increasing the number of training pattern caused slow down the system and decreased the recognition performance. Pre-processing of more input speech pattern takes much time than training the network with these pattern. No proper way of choosing the number of units in hidden layer.

CONCLUSION AND FUTURE WORK

Some approaches of dynamic thresholding on speech segmentation using K-means algorithm, fuzzy c-means algorithm and Otsu's algorithm have been proposed. This produced satisfactory result (94%). The performance of thresholding using K-Means and fuzzy c-means on speech segmentation is better than using Otsu's method. Introduced syllable-based speech classification and this produced satisfied result (91%). We have done comparison study and result analysis of some improved BP algorithms on speech recognition. From experimental results, it is concluded that the Conjugate Gradient and Resilient BP algorithms are good choices for training large networks and speech recognition problems. It has much faster convergence.

Our future works include- reducing segmentation time using fuzzy logic; to classify speech words using fuzzy logic; to speed up the recognition system when using Levenberg-Marquardt algorithm. We have a continuous effort to develop such a system that a robot can understand Bangla speech and react appropriately.