Isolated English Word Recognition System: Appropriate for Bengali-accented English

Tanjin Taher Toma[1], Abu Hasnat Md. Rubaiyat[2], A.H.M Asadul Huq[3]

[1,3]Dept. of Applied Physics, Electronics and Communication Engineering, University of Dhaka, Dhaka, Bangladesh

[2]Dept.of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Abstract: ***Accents of English have been investigated for many years both from the perspective of native and non-native speakers of the language. Various research results imply that non-native speakers of English language produce certain speech characteristics which are uncommon in native speakers' speech. This is because non-native speakers do not produce the same tongue movement as native speakers. This paper presents an isolated English word recognition system devised with the speech of local Bangladeshi people, who are also non-native speakers of English language. Here, we have also noticed a different speech characteristic which is not available within the speech of native English speakers. Two acoustic features, 'pitch' and 'formants' have been utilized to develop the system. The system is speaker-independent and stands on Template based approach. The recognition method applied here is very simple and the recognition accuracy is also very satisfactory.***

Keywords: Isolated word recognition, Speaker-independent, Pitch, Formants, Template based approach

1. Introduction

Automatic speech recognition has been the area of active research for many years. It is still considered as one of the complex problems in the field of signal processing. Automatic speech recognition systems have been implemented in many languages around the world. Researchers have also investigated the effect of non-native speakers on automatic speech recognition performance. They have attempted to explore the difference in characteristics between native and non-native speakers of a particular language. Studies indicate that non-native speakers of some languages provide several speech characteristics that are unusual in native speakers' speech of those languages. For example, the authors of (Sidaras et al., 2009) demonstrated that the duration and the first and second formant frequencies of English vowels spoken by Spanish speakers had different characteristics from those of native English speakers [1]. Similarly, it was noticed that the tongue location of the English vowels by nonnative speakers had different characteristics from that of native speakers (Wade et al., 2007). This paper conducts a study on isolated English words recognition using speech characteristics of local Bangladeshi people, who are also non-native speakers of English language.

Two features of human speech, 'Pitch' and 'Formant' have been analyzed on the isolated speech samples taken from local speakers. The experiment indicates that pitch can separate the utterances of male and female speakers despite the variation of pitch value with emotional states. On the other hand, 'Formant' feature is found capable of making distinction between different utterances of dissimilar vowel sounds. It is very well known that the frequencies of the first two or three formants are sufficient for the perceptual identification of vowels (Pols et al, 1969; Flagan, 1972; Minifie et al, 1973). But, in our study it is noticed that with Bengali accented English, only the second formant is sufficient enough to distinguish vowel sounds.

Thus, analyzing features from speech samples, a speech recognition system is designed and simulated for identifying dissimilar vowel sound utterances of English language. The system follows Template-based approach [2]. Besides, it is a speaker-independent one, i.e. it responses to random speakers including both male and female. In addition, our vocabulary set consists of the isolated utterances which are very useful commands to control a home-appliance or a vehicle by a handicapped person.

2. System Model

The speech recognition system in this work is based on the following model shown in Figure 1.The system has two phases: Training and Recognition. In both phases, the spectrum of input speech is analyzed using a frame-by-frame method to separate the speech from the silence part. The length of each frame is 20 ms and the overlapping between frames is 10ms. In the training phase, word-templates have been generated from the extracted feature of speech signal and the system has been trained with those templates. In the recognition phase, the input speech feature is compared with each of the reference templates to obtain similarity values. The word of the most similar word-template is then selected as the recognition result.

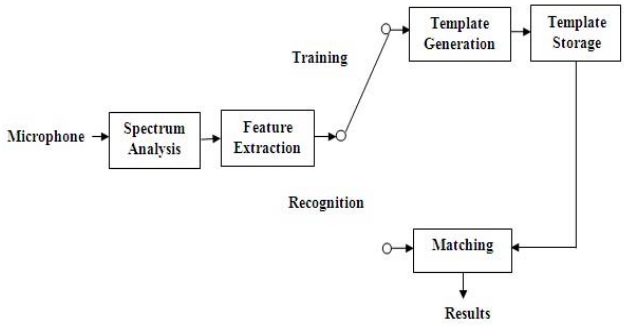

Figure 1: Isolated word recognition model based on template matching

3. Speech Features

3.1 Pitch

Pitch refers to the perceived fundamental frequency of a complex speech signal [3].It is produced due to the vibration of the vocal folds [4]. It depends on the tension of the vocal folds and the sub glottal air pressure when speech is generated. Pitch in human voice is dependent on the length and thickness of the vocal cord as well as the tightening and relaxation of the muscles surrounding them. Since women possess shorter vocal cord than that of the men, they generally have higher pitch value than men do. Hence, pitch contour of an utterance is very useful for gender identification.

3.2 Formants

The term Formant refers to peaks in the harmonic spectrum of a complex sound [5]. In speech science and phonetics, formants refer to acoustic resonance of a human vocal tract [4]. The spectrum of a speech signal might consist of several formants, but the first three have great significance in speech recognition.

Formants are estimated mainly when pronouncing a vowel, because recognition of vowels based on them is easier and gives better result [6]. These formants are quite similar for the same vowel at different fundamental frequencies, enabling it to be recognized regardless of the pitch. In contrast, they are completely different for dissimilar vowel sounds. Hence, words containing dissimilar vowel sounds (e.g., go, head, turn, etc) can be easily detected based on the difference of formant values. Another notable characteristic of formants is that vowel formants of female speakers tend to be located at higher frequencies than those of the males. It is due to some anatomical issue of vocal tract length, that is, the distance from the vocal folds to the lips. Researches imply that male vocal tract is longer than that of the female and the longer the vocal tract, the lower resonant frequencies [9]. Hence, female resonant frequencies or formants do posses higher values compared to man. Thus, it is not plausible to identify a particular utterance by the same formants' values for both male and female speakers.

4. Proposed Approach

A dissimilar-vowel isolated English word recognition system has been developed in this work through the following consecutive steps. Self-formed database is used here. The recordings used in this work were collected via a headset microphone in a closed room. Our database contains 60 speech samples per utterance of the vocabulary set while 30 samples come from male and another 30 come from female speakers. That is, for eight different distinct vowel utterances within the vocabulary set, there are total 480 .wav files in the database. Speakers are local Bangladeshi people.

4.1 Preprocessing

4.1.1 Analog to Digital

The input acoustic sound wave is converted into a digital signal which is convenient for speech processing. A microphone is used for converting acoustic sound into an analog signal. Then analog to digital conversion process is performed at sampling frequency of 8 kHz with 16 bits of resolution.

4.1.2 Windowing

Windowing is an important tool in filtering and separating speech from the silence part. The whole speech signal is divided in some overlapping frames. The frame should not be too small, as inadequate frame size would prevent us from extracting valid audio features. Again for catching time varying characteristics of the audio signal, the frame duration cannot be too big. In our experiment we have taken 20 ms or 160 sample points as a frame with overlapping of 10 ms or 80 sample points.

4.1.3 Speech detection

The durations of all the sample speech signals we have worked with are 2 seconds. These 2 seconds contain both the speech and silence portions. We need to take only the speech for the processing. There are a lot of algorithms for speech portion detection. In this experiment we have used an algorithm[10] that calculates the energy and number of zero crossing rate in a frame and then compares these values with a threshold to determine whether there is a possible voice activity or not. The threshold has been set by the following expression:

$$\tau = \sigma + \alpha * \mu$$

Where, σ and μ are the variance and mean of the noise signal respectively. α is a constant which depends on the mean value of the noise signal. If either energy or zero crossing rates exceed the threshold, it continues analyzing frames and start buffering. If the number of the adjacent frames that have exceeded the threshold also exceeds buffer length (12 in this experiment), it will be considered that a word has been detected. Otherwise, the frames would be considered as interference.

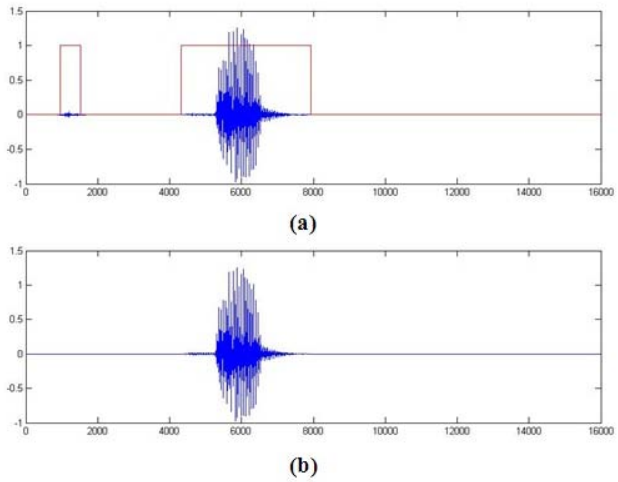

Figure 2: (a) Input speech signal (b) After windowing

4.2 Feature Extraction

- Pitch feature of a speech signal is extracted here implementing the Modified Autocorrelation Function (MACF) algorithm. It is a widely used time domain approach for estimating pitch period of a speech signal. This method [7] is based on detecting the highest value of the autocorrelation function in the region of interest.

This method utilizes short-time autocorrelation function which is given by:

$$R(m)=\frac{1}{N}\sum_{n=0}^{N-1-m}x(n)\ x(n+m),\ 0\le m<M_o \qquad (1)$$

The variable m in (1) is called lag or delay, and the pitch is equal to the value of m which results in the maximum $R(m)$. N is the length of analyzed sequence, $x(n)$ and M_0 is the number of autocorrelation points to be computed.

- In this work, formants are extracted from a speech signal by estimating the power spectral density (PSD) of that signal utilizing the Yule-Walker AR method. Yule-Walker autoregressive (AR) method [8] is a parametric spectral estimation technique. In this method, PSD is estimated from a signal that is assumed to be output of a linear system driven by white noise. In fact, this method estimates the PSD by estimating the parameters of the linear system that hypothetically generates the signal. Yule-Walker method estimates the PSD of a signal by the following equation:

$$P(f)=\frac{1}{f_s}\frac{p}{|1-\sum_{j=1}^{p}\Phi_j e^{-i2\pi j f/f_s}|^2} \qquad (2)$$

In (2), f_s is sampling frequency, p is the order of the associated autoregressive (AR) process and $\{\Phi_j\}$ are the corresponding coefficients of that AR process.

4.3 Templates Generation

Word templates have been generated here utilizing the formant feature. It is already known to us that each formant (Ist, 2nd & 3rd) extracted from a particular speech signal possesses separate frequency ranges in the audio spectrum. Moreover, our study indicates that among the first three formant ranges, second formant frequency range is the most suitable one to distinguish dissimilar vowel sounds. Hence, we have considered the second formant frequency range of every utterance in our vocabulary set as a template for that utterance and each word template has been built taking into account the second formant frequency range corresponding to that utterance. But due to the difference in male and female formants to some extent, two templates are prepared per utterance-one representing male and another representing female.

4.4 Training

In this work, the developed recognizer has been trained with the reference word templates for every utterance in the vocabulary set. In addition, for the purpose of gender identification of an uttered speech, the recognizer has been also trained with two separate pitch ranges for male and female speech.

4.5 Recognition

Speech recognition operation is performed by the system firstly extracting pitch and formant features from the input speech command. Then according to the extracted pitch value either the male reference templates or the female reference templates are selected for the next operation. Then, second formant frequency value previously extracted from the input speech is compared with each of the reference word templates of corresponding gender. If second formant matches with any of the templates, then recognition is considered successful, otherwise recognition failure occurs. The word of the closest matched word-template is the recognition result. The recognition process is shown below in Figure 3.

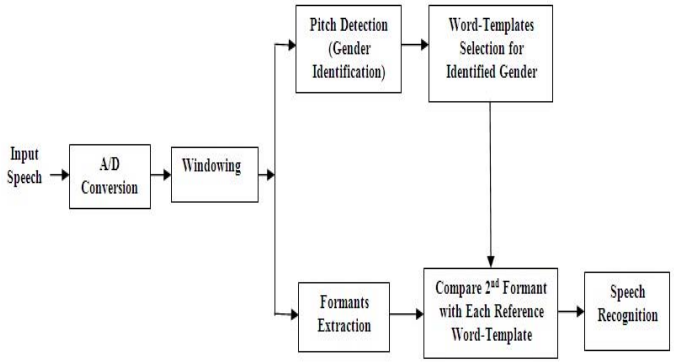

Figure 3: Block diagram for speech recognition

5. Experimental Findings

5.1 Features analysis

In this work, it is noticed that human pitch and vocal formants are useful features for speech recognition. In our experiment, it is seen that the ranges of pitch for male and female are quite different and lies in a particular range. The study provides pitch range for male 100-155Hz and for female 170-250Hz, irrespective of utterances. Although change in emotional state affects pitch value, still it works well to identify gender from the speech of a person.

On the other hand, Analysis of first three formants of several distinct vowel utterances has made it possible to develop a recognizer in order to distinguish them for a particular gender. Our study demonstrates that with the Bengali accented English utterances second formant frequency performs better in making distinction between unlike vowel utterances rather than the first and third formants. The experimental observation is illustrated below.

The following Table 1 presents the first three formants (F1, F2& F3) values for eight dissimilar vowel utterances. In the table, it is seen that F1 and F3 values of all the utterances are almost similar, while F2 values vary between them. That is, these distinct vowel utterances cannot be separated based on F1 and F3 values. Hence, F2 is the right choice as a feature which might be utilized to make distinction between the utterances.

Now, Table 2 below shows the F1, F2 and F3 values for five different samples of a particular utterance. The table indicates that F1 and F2 values remain unchanged between different samples of a particular utterance while F3 values do not remain stable. That is, F2 remain stable for a particular utterance as well as makes distinction between different utterances of unlike vowel. This fact is more clearly represented graphically in Figure 4 and Figure 5.

Table 1: First three formants for distinct vowel utterances (male)

Utterances & associated vowel	Formants		
	F1 (Hz)	F2(Hz)	F3(Hz)
Go /o/	468.8	906.3	2437.5
Right /ai/	593.8	1625.0	2562.5
Left /e/	468.8	1968.8	2593.8
Halt /al/	531.3	1281.3	2556.3
Pick /i/	312.5	2437.5	2937.5
Put /u/	343.8	1093.8	2343.8
Turn /ur/	656.3	1468.8	2531.3
Back /a/	468.8	1812.5	2625.0

Table 2: First three formants for different samples of a single utterance (male)

Samples of an utterance	Formants		
	F1(Hz)	F2(Hz)	F3(Hz)
Go1 /o/	468.8	906.3	2437.5
Go2 /o/	437.5	937.5	2150.0
Go3 /o/	437.5	906.3	2675.5
Go4 /o/	406.3	968.8	2375.0
Go5 /o/	468.8	968.8	2718.8

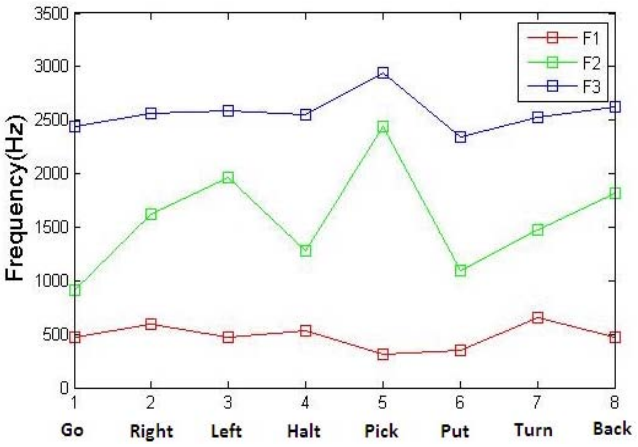

Figure 4: Nature of the first three formants for different vowel sounds (male)

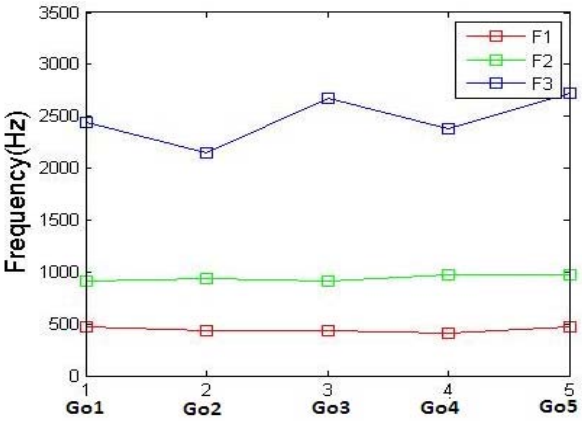

Figure 5: Nature of the first three formants for different samples of the utterance Go/o/.

The above tables, Table 1, Table 2 and graphs Figure 4, Figure 5 are formed based on male speech data. Female utterances provide us with similar outcome except at slightly higher frequencies.

To demonstrate the effectiveness of second formant, F2 in recognition of Bengali-accented English utterances, more data of F2 are represented in the following tables. Table 3 given below presents F2 values of eight distinct vowel utterances uttered by four different male speakers. Again, it is seen that the F2 is different for unlike-vowel utterances whereas it is almost same for a particular utterance uttered by different speakers.

Table 3: 2^{nd} formant values for male utterances

Utterances & associated vowel	2^{nd} Formant frequency,F2 (H_z)			
	Speaker 1	Speaker 2	Speaker 3	Speaker 4
Go /o/	906.3	937.5	968.8	906.3
Right /ai/	1625.0	1656.3	1687.5	1656.3
Left /e/	1906.3	1937.5	1968.8	1937.5
Halt /al/	1218.8	1281.3	1218.8	1218.8
Pick /i/	2437.5	2468.8	2437.5	2468.8
Put /u/	1093.8	1125.0	1093.8	1068.8
Turn /ur/	1437.5	1468.8	1468.8	1437.5
Back /a/	1812.5	1843.8	1843.8	1812.5

From Table 3, it is also clear that F2 corresponding to each distinct-vowel utterance possess a value that lies in a particular range around a certain frequency. Consequently, we have determined F2 ranges for each of the utterances in our vocabulary set. These ranges are non-overlapping and they act as the templates for our system.

Then, F2 values corresponding to the same eight utterances uttered by four different female speakers are given in Table 4. Here we also see different F2 values for different utterances of dissimilar vowels except at slightly higher frequencies.

Table 4: 2nd formant values for female utterances

Utterances & associated vowel	2nd Formant frequency, F2 (H_z)			
	Speaker 1	Speaker 2	Speaker 3	Speaker 4
Go /o/	1000.0	1031.3	1062.5	1031.3
Right /ai/	1937.5	1968.8	1937.5	1968.8
Left /e/	2312.5	2343.8	2375.0	2312.5
Halt /al/	1312.5	1343.8	1375.0	1343.8
Pick /i/	2906.3	2968.8	2968.8	2906.3
Put /u/	1125.0	1187.5	1156.3	1125.0
Turn /ur/	1500.0	1531.3	1562.5	1562.5
Back /a/	2218.8	2250.0	2281.3	2218.8

In this work, formants (F1, F2 &F3) of a particular speech signal have been extracted from the power spectral density (PSD) spectra of that speech. Some of the PSD curves obtained through this experiment are presented in Figure 6, 7, and 8. In these figures, small circles indicate second formant for respective PSD plots.

The PSD plots of three dissimilar vowel utterances, 'go/o/', 'left/e/' and 'pick/i/' uttered by three male speakers are shown in Figure 6.

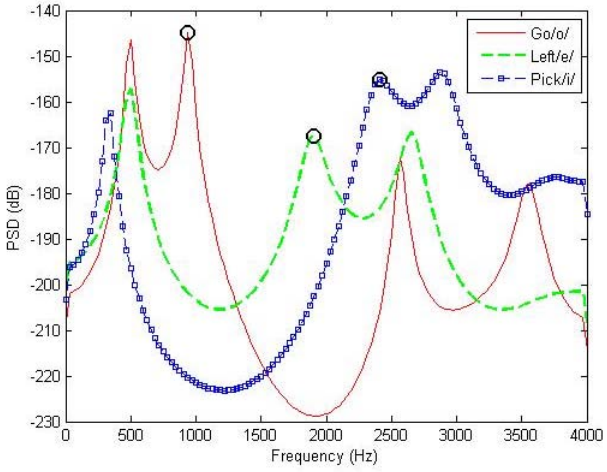

Figure 6: PSD of 'go/o/', 'left/e/' & 'pick/i/' by three male speakers

Figure 6 shows the peaks of the PSD curves for three distinct vowel utterances selected from the vocabulary set. From the figure it is also clear that the 2nd peak of each PSD curve better describes the isolation between them. Thus, all the dissimilar vowel utterances in our vocabulary set can be distinguished based on the differences in second formant frequencies.

Figure 7 given below manifests the PSD plots of the utterance 'go/o/' uttered by three different male speakers. From the figure it is observed that second formant frequencies are almost same for the utterance 'go/o/' by three different speakers, although PSD values slightly vary between them. That is, for a particular vowel utterance, second formant value is almost constant irrespective of speakers. More accurate consistency is observed in the values of second formants while uttering the same speech by a single speaker at different instants and this fact is shown in Figure 8.

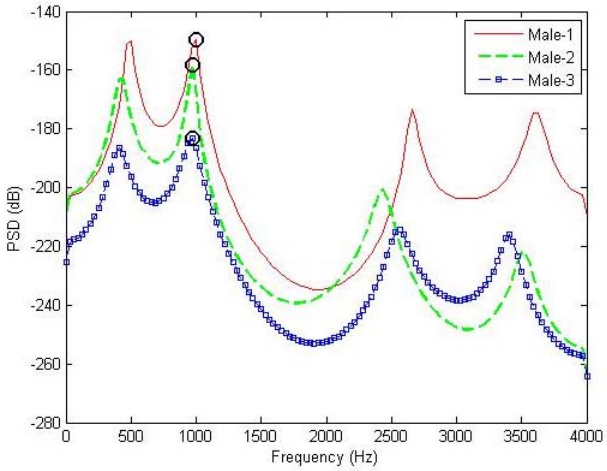

Figure 7: PSD of 'go/o/' by three different male speakers.

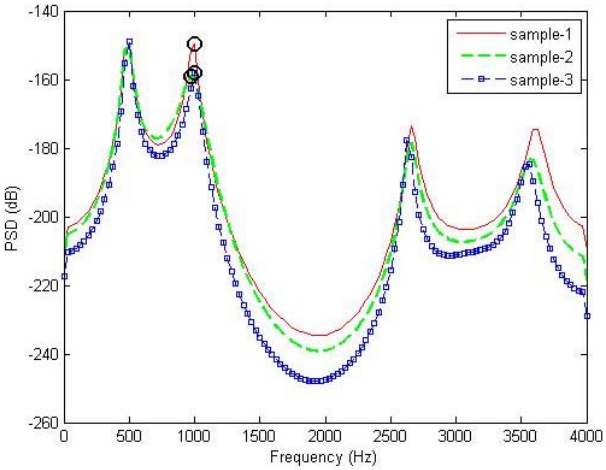

Figure 8: PSD of three samples of 'go/o/' by same person (male)

The above findings for 'go/o/' shown in Figure 7, and 8 are also true for other utterances.

Likewise, female speeches provide us with similar graphs as in Figure 6, 7, and 8, despite slightly higher values of formant frequencies in PSD spectra.

5.2 Performance

The performance of a speech recognition system is evaluated in terms of accuracy it provides. Performance of most of the speech recognition systems is strongly influenced by the environmental noise. The mismatch between the training conditions and the testing conditions has a deep impact on the accuracy of these systems and represents a barrier for their operation in noisy environments. As a result, we have tested our system at both high-noise and low-noise environments. Performance test has been performed on 50 people including male & female speakers. Each utterance was tested two times by a single speaker. Accuracy is calculated for each of the utterances in the vocabulary set. Recognition accuracy is presented in Table 5 and its analysis is presented in graph form in Figure 9. From the graph, it is obvious that percentage of accuracy for a particular utterance is significantly higher at low-noise environment (SNR between 30-35 dB) compared to the high-noise one (SNR between 10-15 dB).

Table 5: Recognition accuracy

Utterances & associated vowel	Recognition in accuracy (%)	
	Low-noise	High-noise
Go /o/	99.2%	96.3%
Right /ai/	98.5%	95.5%
Left /e/	98.8%	95.0%
Halt /al/	99.0%	95.3%
Pick /i/	97.8%	93.7%
Put /u/	99.1%	96.0%
Turn /ur/	97.8%	94.8%
Back /a/	98.0%	94.5%

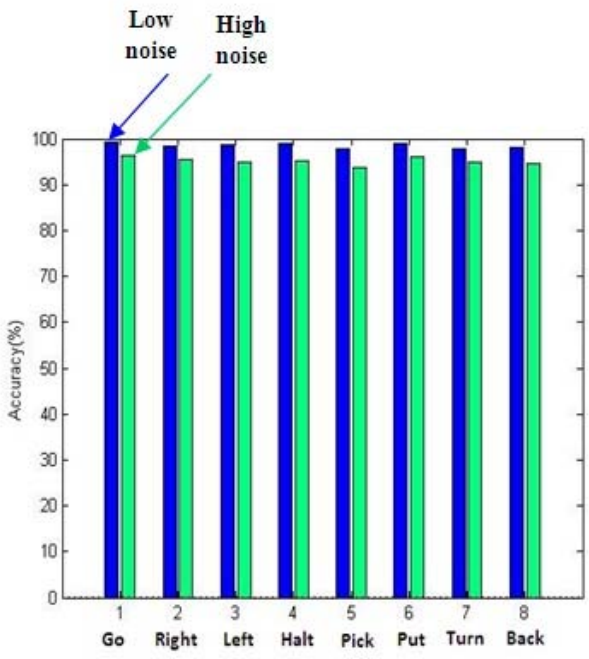

Figure 9: Recognition accuracy of dissimilar vowel utterances at low-noise & high-noise environments.

In addition, it is already mentioned that training data for our system has been provided from the Bengali speaking people who are non-native English speakers. As a result, same utterance may be uttered by different person with a slightly different pronunciation. The pronunciation variation between training and testing data sometimes causes recognition confusion between the utterances. This phenomenon is shown at Figure 10 through a confusion matrix.

The matrix represents the performance of our system in distinguishing between the distinct-vowel utterances of our vocabulary set.

	Go/o/	Right/ai/	Left/e/	Halt/al/	Pick/i/	Put/u/	Turn/ur/	Back/a/
Go/o/	**99.00%**	0.00%	0.00%	0.00%	0.00%	1.00%	0.00%	0.00%
Right/ai/	0.00%	**98.25%**	0.00%	0.00%	0.00%	0.00%	1.75%	0.00%
Left/e/	0.00%	0.00%	**98.00%**	0.00%	0.00%	0.00%	0.00%	2.00%
Halt/al/	1.50%	0.00%	0.00%	**98.50%**	0.00%	0.00%	0.00%	0.00%
Pick/i/	0.00%	0.00%	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
Put/u/	1.25%	0.00%	0.00%	0.00%	0.00%	**98.75%**	0.00%	0.00%
Turn/ur/	0.00%	1.75%	0.00%	0.00%	0.00%	0.00%	**98.25%**	0.00%
Back/a/	0.00%	0.00%	2.00%	0.00%	0.00%	0.00%	0.00%	**98.00%**

Figure 10: Confusion matrix showing both correct and wrong identification by the system

Each row of the matrix corresponds to a correct class and each column corresponds to a predicted class. The diagonal elements represent correctly classified compounds while the cross-diagonal elements represent misclassified compounds. From the figure we see that 99.00% of the test-utterances 'go' has been correctly identified as 'go' while 1.00% of them has been incorrectly identified as 'Put'. Again, 98.25% of the test-utterances 'right' has been correctly identified as 'right' while 1.75% of them have been misidentified as 'Turn'. Similarly, other utterances have both proportions of correct identification and wrong identification except 'Pick', which has 100.00% correctly classified compounds.

6. Conclusion and Future Scopes

Speech recognition technology has improved a lot in the last few decades. It is still considered as an attractive field for conducting research. In our study we have worked on isolated English words recognition. This type of speech recognition is an older one, but here we have approached a recognition system that works very successfully to recognize isolated English words uttered with Bengali accent. Local Bangladeshi people are the subjects of our experiment. English is considered to be the second language in this country and widely used in government, law, business, media (newspapers) and education. Hence, a speech recognition system has been developed with the speech samples from the local people. Analyzing the collected speech samples, we have found 2nd formant frequency to be very effective in making distinction between dissimilar vowel utterances. By adding the pitch feature with 2nd formant, a very simple as well as successful system has been prepared to recognize various isolated English words by our people. The utterances in our system vocabulary are the useful commands to drive a voice-operated vehicle or a simple voice-driven home appliance. Besides, the utterances chosen for our vocabulary possess dissimilar vowel sounds. As a future extension of this work, similar vowel sound utterances would also be added in our vocabulary and the system will be made capable of identifying them. Moreover, the recognition accuracy of this system is found quite satisfactory at less-noise environment but it is not as much successful while noise is very high. Hence, noise cancellation algorithm applied in this will be further

improved to provide better accuracy at high-noise environment too.

References

[1] Y. R. O. a. H. K. K. Mina Kim "Non-native Pronunciation Variation Modeling for Automatic Speech Recognition," 1Mobile Communication Department, LG Electronics; 2 School of Information and Communications, Gwangju Institute of Science and Technology,Korea

[2] S. K. K. M.A. Anusuya, "Speech recognition by machine: A Review,"*International Journal of Computer Science and Information Security*,vol. 6, 2009

[3] M. M. R. D. Ms. Arundhati S. Mehendale, "Speaker identification," *An International Journal (SIPIJ)*, vol. 2, 2011.

[4] A. R. P. Bageshree V. Sathe-Pathak, "Extraction of Pitch and Formants and its Analysis to identify 3 different emotional states of a person," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.

[5] D. A. Kocharov, "Automatic vowel recognition in fluent speech," in *Proceedings of the 9th Conference of Speech and Computer*, St. Petersburg, Russia, Sept. 2004.

[6] S. s. I. c. Biljana Prica, "Recognition of vowels in continuous speech by using Formants," vol. 23, 2010

[7] L. R. RABINER, "On the use of Autocorrelation Analysis for Pitch detection," *IEEE Transaction On Acoustics, Speech, and Signal Processing*, 1977.

[8] G. Eshel, "The Yule Walker equations for the AR coefficients."

[9] E. PEPIOT, "Voice, speech and gender: male-female acoustic differences and cross language variation in English and French speakers,"Paris, 2012.

[10] E. H. Bourouba, M. Bedda and R. Djemili, "Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM", Informatics, vol. 30, (2006), pp. 373–384, Algeria.

Author Profile

Tanjin Taher Toma received the B.Sc (Hons.)in Applied Physics, Electronics & Communication Engineering from University of Dhaka, Dhaka, Bangladesh (2012). She is currently pursuing her M.sc degree in the same department.

Abu Hasnat Md. Rubaiyat received the B.Sc in Electrical & Electronic Engineering from Bangladesh University of Engineering & Technology, Dhaka, Bangladesh (2012). He is currently pursuing his M.sc degree in the same department.

A.H.M Asadul Huq received the B.Sc. (Hons.) and M.Sc. degrees in Applied Physics and Electronics from the University of Dhaka in 1980 and 1981, respectively. Since M.Sc. studies, he is working in the field of Electronic communications and DSP. He received the Doctor of Engineering degree in 1994 from the Graduate School of Natural Science and Technology, Kanazawa, Japan in the field DSP. He was with Bangladesh Atomic Energy Commission, Dhaka, Bangladesh from 1985 to 1996. He worked in Oakridge National Laboratory (ORNL) in Tennessee, USA in 1991 as IEAE fellow. In 1996, he joined the Department of Applied Physics, Electronics and Communication Engineering as Assistant Professor. In 1997 he went to the Communication Research Laboratory (CRL) of Japan to do Post-doc research in the field of the mobile telecommunication systems engineering. Presently, he is Professor in the same department of University of Dhaka and continuing research in the field of DSP and Electronic Communications.