International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 2
ISSN: 2088-5334

Speaker Independent Speech Recognition of Isolated Words in Room Environment

M. Tabassum[#], M. A. Aziz Jahan[*], M. M. Rahman[$], S. B. Mohamed[&], M. A. Rashid[&]

[#]*Department of EEE, University of Dhaka, Dhaka-1217, Bangladesh*

[*]*Department of EEE, Islamic University of Technology (IUT), Gazipur, Bangladesh*

[$]*Department of CSE, Dhaka University of Engineering and Technology (DUET), Bangladesh*

[&]*FRIT, Universiti Sultan Zainal Abidin (UniSZA), 21300 Kuala Terengganu, Malaysia*
E-mail: marashid@unisza.edu.my

***Abstract*— In this paper, the process of recognizing some important words from a large set of vocabularies is demonstrated based on the combination of dynamic and instantaneous features of the speech spectrum. There are many procedures to recognize a word by its vowel, but this paper presents the highly effective speaker independent speech recognition in a typical room environment noise cases. To distinguish several isolated words of the sound of different vowels, two important features such as Pitch and Formant are extracted from the speech signals collected from a number of random male and female speakers. The extracted features are then analyzed for the particular utterances to train the system. The specific objectives of this work are to implement an isolated and automatic word speech recognizer, which is capable of recognizing as well as responding to speech and an audio interfacing system between human and machine for an effective human-machine interaction. The whole system has been tested using computer codes, and the result was satisfactory in almost 90% of cases. However, the system might get confused by similar vowel sounds sometimes.**

***Keywords*— speech recognition; features extraction; signal processing; sound processing**

I. INTRODUCTION

Speech is the primary means of expressing emotion and communication between mankind. In 1773 a professor of physiology in Copenhagen and a Russian scientist Christian Kratzenstein succeeded in producing vowel sounds by using resonance tubes which were connected to organ pipes. Later in Vienna, Wolfgang von Kempelen constructed an Acoustic-Mechanical Speech Machine in 1791 [1]. In 1881 Alexander Graham Bell with his cousin Charles Sumner Tainter and Chichester Bell invented a recording device, which was the same way as a microphone. Based on this invention, Tainter and Bell formed the Volta Graphophone Co. [2]. In the year 1952, Balashek of Bell Laboratories and Davis, Biddulph, built a system for isolated digit recognition for a single speaker using the measured formant frequencies during vowel regions of each digit. In the 1980's, Speech recognition research was characterized by a shift in methodology from the more intuitive template- based approach towards a more rigorous statistical modelling framework [3].

Although the basic idea of the hidden Markov model (HMM) was understood by a known nearly only in a few laboratories., most speech recognition research, up to 1980, considered the major research problem to be one of converting a speech waveform (as an acoustic realization of a linguistic event) into words (as a best-decoded sequence of linguistic units). The keyword spotting method and its application in AT&T's Voice Recognition Call Processing (VRCP) System, as mentioned earlier, was introduced in response to the first factor while the second factor focused the attention of the research community on the area of dialog Management.

Humans normally express their feelings, ideas, and thoughts orally to another person using a series of complex vocal movements. Speech is an information-rich, frequency modulated, signal exploiting, time and amplitude modulated carriers (e.g. harmonics, noise, power, pitch information, duration, resonance movements) to convey the emotions and information about words, expression, accent, style of the speech, speaker identity, health condition of the speaker and so on [4].

Pitch is one of the important acoustic features in speech recognition process. It is also considered as the fundamental frequency of a complex speech signal. Pitch signal is basically produced due to the vibration of vocal folds. It normally depends on the tension of the vocal folds as well as the subglottal air pressure when speech is generated. Conversely, pitch in a human voice is also dependent on the thickness and the length of the vocal cord, as well as the relaxation and tightening of the muscles surrounding the vocal cord [3]. A system of gender classification can be easily identified based on the pitch, sometimes formants and the combination of both features. Pitch is the perceptual property of speeches which allows the orders on a frequency based scale [4]. Fig. 1 (a) and (b) show the internal structure of a human vocal cord.

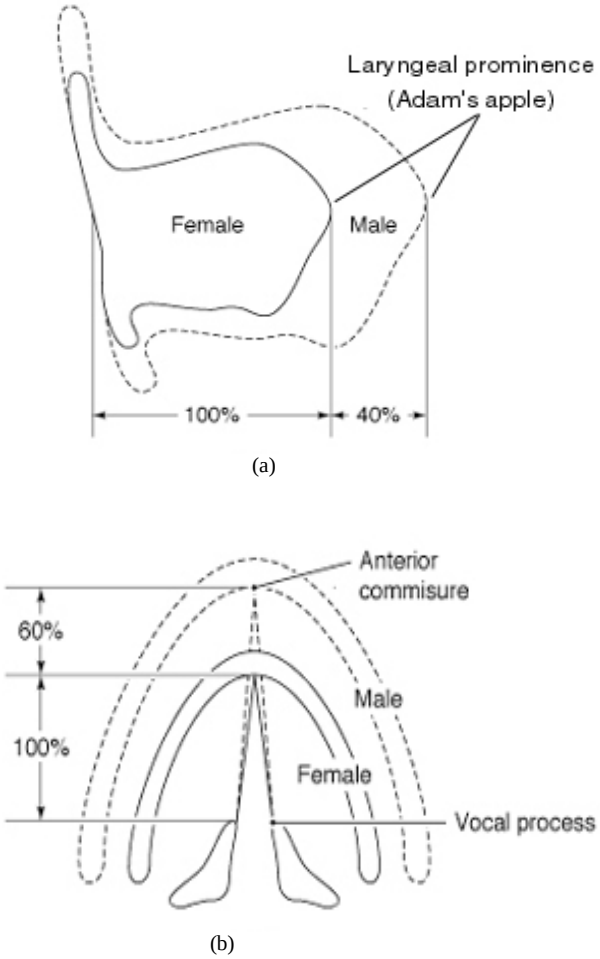


Fig. 1 Internal structure of a human vocal cord (a) side view, (b) front view.

Since women possess shorter vocal cords than of the men, they generally have a higher pitch than men. Thus, pitch in human voice plays a significant role in gender identification [4,5].

In addition, the formant is a harmonic of a note that is augmented by resonance. In acoustics, a very similar definition is widely used: the Acoustical Society of America defines a formant as: "a range of frequencies [of a complex sound] in which there is an absolute or relative maximum in the sound spectrum" [6]. Formants are often measured as amplitude peaks in the frequency spectrum of the sound using a spectrogram or a spectrum analyzer and, in the case of the voice, this gives an estimate of the vocal tract resonances. In vowels spoken with a high fundamental frequency fundamental frequency, as in a female or child voice, however, the frequency of the resonance may lie between the widely spaced harmonics and hence no corresponding peak is visible [6, 7].

A lot of studies have been carried out to investigate the acoustic indicators to detect features in speech. The characteristics that are commonly considered include fundamental frequency, spectral variation, duration, wavelet and intensity-based features [8, 9]. In this paper, linear feature extraction techniques and their extraction algorithms are explained. These features are used to identify the proper language state. The production of those speech signals is considered as the convolution between vocal tracks [10, 11].

II. Material and Method

A. Feature Extraction

The fundamental frequency (Fo) is the main cue of the pitch. However, it is difficult to build a reliable statistical model involving fundamental frequency Fo because of pitch estimation errors and the discontinuity of the Fo space. Thus, a reliable pitch detection algorithm (PDA) is a very important component in many speech processing systems.

By analyzing the power spectral density (PSD) spectra of the sound, formant frequencies of a particular uttered sound can be extracted. The formants are the frequencies corresponding to the peaks in the PSD spectra. In order to obtain the PSD of an utterance, Yule-Walker AR method is used in this work [12].

B. Autocorrelation Method and AMDF

Generally, the pitch detection algorithms use short-term analysis techniques. For every frame $x_{m,}$ we get a score f(T| x_m) that is a function of the candidate pitch periods T. Algorithm determine the optimal pitch by maximizing is given by:

$$Tm = \underset{T}{\operatorname{argmax}} f(T \mid x_m) \quad (1)$$

A commonly used method to estimate pitch is based on detecting the highest value of the autocorrelation function in the region of interest. Given a discrete time signal x(n), defined for all n, the auto-correlation function is generally defined as:

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \quad (2)$$

C. Modified Autocorrelation Method

According to the discussion above, the modified autocorrelation pitch detector based on the center-clipping method and infinite-clipping is used in our implementation. Fig. 4 shows a block diagram of the pitch detection algorithm. The method requires that the speech be low-passed filtered to 900 Hz. The low-pass filtered speech signal is digitized at a 10-kHz sampling rate and sectioned into overlapping 30-ms (300 samples) sections for processing.

PID control optimization process is done to get the right values in the PID control parameters, so the response generated controllers capable of handling the minimum conditions for the achievement of rapid set points and small overshoot. Fig. 2 shows the block diagram of PDA [14].

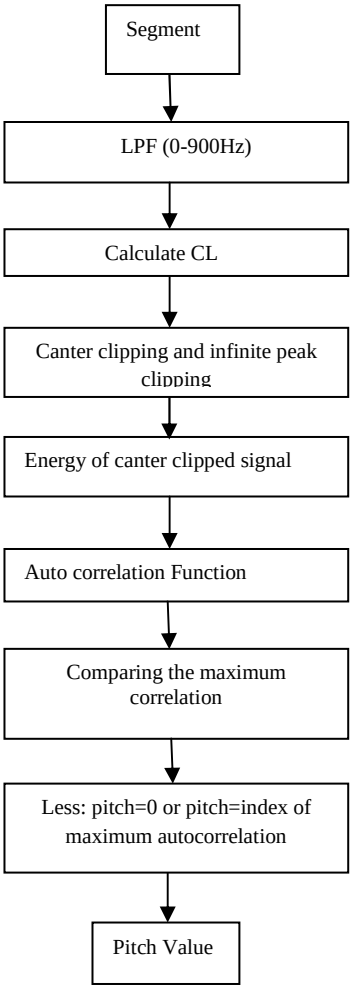


Fig. 2 Block diagram of PDA using modified autocorrelation method

D. Yule-Walker AR Method

Assuming a given zero-mean discrete time series $\{x_i\}_1^N$ is an AR process, the appropriate order p of the AR(p),

$$x_{i+1} = \phi_1 x_i + \phi_2 x_{i-1} + \ldots + \phi_p x_{i-p+1} + \varepsilon_{i+1} \quad (3)$$

and the corresponding coefficients $\{\varphi_j\}$.

E. Speech Recognition Techniques

1) Data Gathering

Various speech samples were taken from 10 female and male subjects. The recording environment was typical Bangladeshi. Audio templates for training were recorded from typical living rooms environments and classrooms.

2) Data Pre-processing

After the collection, data pre-processing is the initial step of the recognition process. Here, some speech commands are taken as inputs by using a microphone. The microphone converts the speech signal into an analog electrical signal. The speech command is recorded in MATLAB with the sampling frequency of 8000 Hz. After sampling, there are some discrete speech signals. Before following the further steps, those discrete speech signals go through some filters and windows for noise cancellation.

3) Pitch and Formant Extraction

After pre-processing the data, feature extraction step begins. Since the pitch of a male and a female speaker lie in two different ranges, formants also differ between them. But this difference is not sharp enough to distinguish between same male and female utterance. Therefore, the pitch is extracted first to detect either the pitch is from a male speaker, or from a female speaker. Then formants are extracted according to that specific pitch.

4) Preparing Sample Templates

In this paper, data for four specific speech commands ‘Go’, ‘Right’, ‘Left’ and ‘Halt’ have been collected from 10 male and 10 female speakers. Each command has uttered two times. Then the pre-processing and feature extraction steps are followed on the collected data. The obtained values of extracted features are analyzed, and then templates are prepared through determining two things if the ranges of the pitch are for male or female utterances. And two different set of ranges of formants for each of the specific speech commands, where one of the set is for male utterances and another one is for female utterances.

5) Analyzing

The whole speech recognition system must be trained with enough feature data to make it much capable of recognition. The system has been trained with pitch and formant data ranges by obtaining through the previous steps. After training the system, this is now ready to recognize the specific input speech commands by random speakers including both male and female through the microphone. As the results of testing will evaluate the system performance rate, the trained system has been tested many times by many speakers in order to find out the accuracy of this system.

III. Result and Discussion

Gathering and analyzing the pitch and formant values of different speakers, there is a clear distinction between male and female voices.

A. Pitch Calculation for Male and Female Utterances

Pitch readings of 10 male and 10 female voices for the utterances Go, Right, Left and Halt are collected, and 10 of the Pitch readings for male voices are summarized. Table 1 and Table 2 represent the Pitch values for 10 female voices.

It is clearly seen from Table 1 and Table 2 that the most of the pitch values for male utterances lie within the range of 100-170 Hz, and pitch for women vary from 180 Hz to 290 Hz. The conclusion can be drawn that the range of male and female pitch are located far away from one another. There is

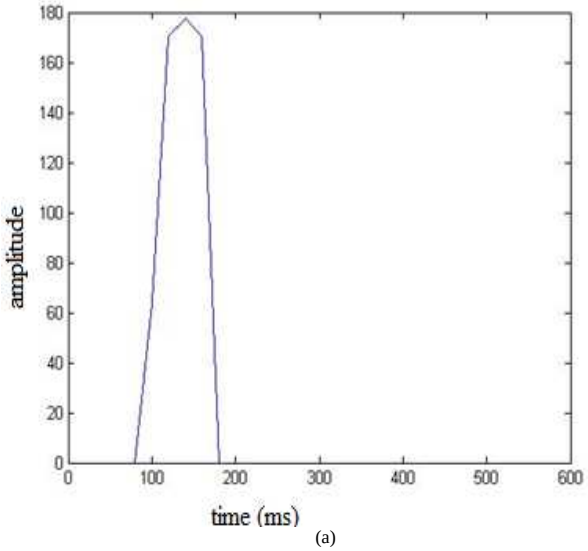
no overlapping between the two pitches. Figs. 3(a), (b) indicate the difference between male and female pitch values.

TABLE I
PITCH VALUES FOR MALE UTTERANCES IN HZ

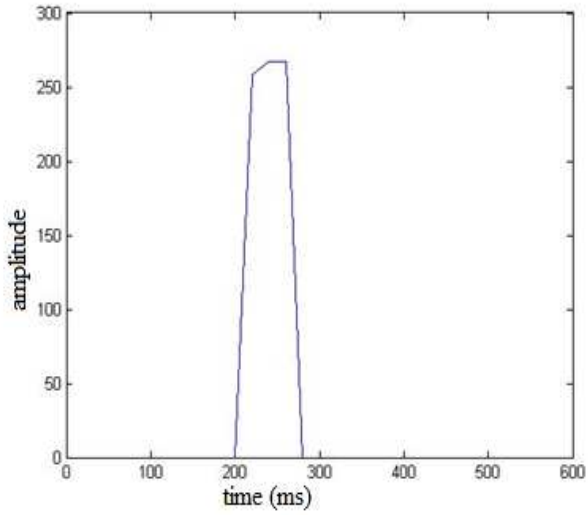
No. of male speakers	GO	Right	Left	Halt
M1	145.5508	126.5881	144.1151	161.1069
M2	140.3509	131.2169	95.7133	139.7911
M3	166.3186	143.9632	167.9957	162.0968
M4	157.8947	152.8786	166.6667	163.2653
M5	166.9246	145.5526	157.9196	106.7304
M6	154.9663	151.5446	149.0909	154.4195
M7	151.0383	132.7829	146.2799	155.1577
M8	123.8059	112.6761	112.9366	115.6380
M9	110.9263	153.7109	161.7654	165.8723
M10	143.5562	142.6367	149.0909	157.2852

TABLE II
PITCH VALUES FOR FEMALE UTTERANCES IN HZ

No. of Female Speaker	GO	Right	Left	Halt
F1	263.7993	184.1414	193.2252	252.9923
F2	262.5483	252.1383	255.5781	241.5541
F3	216.5242	179.2717	214.3196	198.3740
F4	285.7143	2699.929	289.4849	282.4382
F5	242.8571	228.5714	247.4747	240.0475
F6	232.0646	242.0909	269.6029	230.8649
F7	219.9488	237.8066	235.0929	256.5609
F8	248.9588	252.3719	285.7143	253.7292
F9	286.3905	221.5893	267.9535	276.9468
F10	245.7359	235.9648	245.6943	211.4878



(a)



(b)

Fig. 3 (a) PSD of 'go' for male 1, and (b) PSD of 'go' for female 1 indicating pitch

B. Formant Calculation for Male and Female Utterances

Formant analysis actually assists us to distinguish between various isolated words containing a different vowel. The 2nd formant is very effective in the recognition process among the first three formants. Table 3 is representing the normalized frequency scale of the 2nd formant values of 10 male speakers. And immediately after that, there is Table 4 which represents the normalized frequency scale of the 2nd formant values of 10 female speakers.

TABLE III
THE 2ND FORMANT VALUES FOR MALE UTTERANCES (NORMALIZED FREQUENCY SCALE)

No. of male speakers	GO	Right	Left	Halt
M1	.1328	.1602	.2461	.1406
M2	.1101	.1788	.2540	.1406
M3	.1132	.1758	.2000	.1450
M4	.1000	.1758	.2292	.1567
M5	.1171	.1523	.2579	.1501
M6	.1328	.1650	.2656	.1562
M7	.1101	.1680	.2774	.1528
M8	.1101	.1663	.2649	.1415
M9	.1143	.1632	.2042	.1453
M10	.1242	.1758	.2322	.1533

TABLE IV
THE 2ND FORMANT VALUES FOR FEMALE UTTERANCES (NORMALIZED FREQUENCY SCALE)

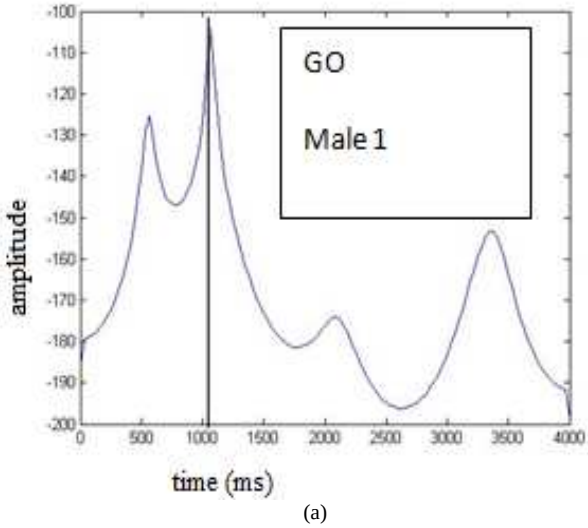
No. of female speakers	GO	Right	Left	Halt
F1	.1211	.1638	.1428	.1640
F2	.1328	.1679	.1421	.1640
F3	.1367	.1605	.1501	.1640
F4	.1211	.1584	.1501	.1681
F5	.1289	.1543	.1489	.1691
F6	.1100	.1565	.1450	.1719
F7	.1100	.1514	.1450	.1691
F8	.1367	.1602	.1484	.1736
F9	.1132	.1634	.1421	.1736
F10	.1123	.1578	.1450	.1681

After analyzing the '2nd formant values' of male and female speakers for the utterances 'Go', 'Right', 'Left' and 'Halt', it is seen that for a specific person like Male 2, values of 2nd formant are different for different words. Thus, the system can easily recognize these words of the dissimilar vowel. On the contrary, Table 3 and Table 4 also indicate that second formant of a specific utterance such as Left remains almost identical irrespective of speakers. But, due to the variation of accent and recording environment, it slightly varies from one another. Hence, we find the range of variation for a specific utterance regardless of speakers.

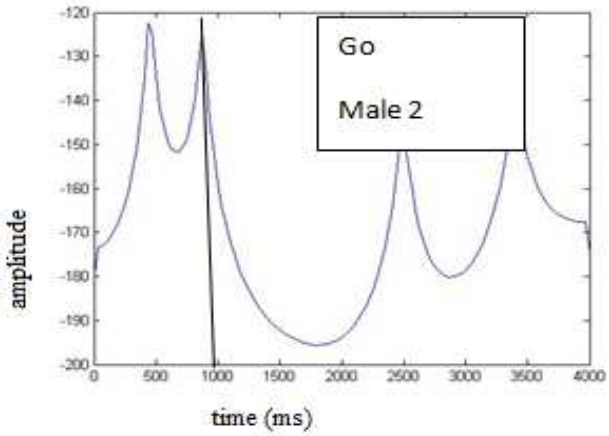
C. Simulation Results

In the simulation experiment, the obtained power spectral density, PSD spectrum for different utterances uttered by random speakers using Yule walker AR method. Peak locations in those spectrums correspond to the desired formant frequencies.

Figs. 4 (a), (b), 5 (a), (b) demonstrate the PSD spectra of two male, Male 1 and Male 2 for the word "GO" and "Halt". The locations of 2nd formants are indicated in the PSD curves. From the graphs, it is clear that locations of 2nd formants are different for different utterances.

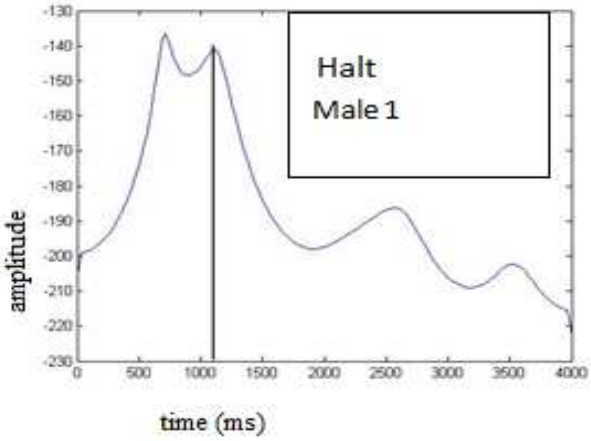


(a)

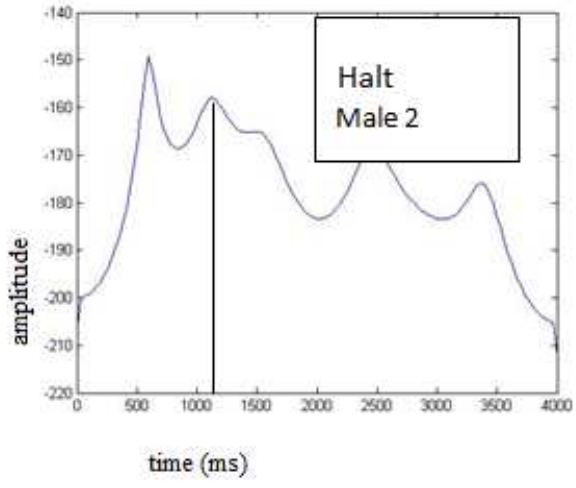


(b)

Fig. 4 (a) PSD of 'go' for male1 and (b) PSD of 'go' for male 2 indicating 2nd formant



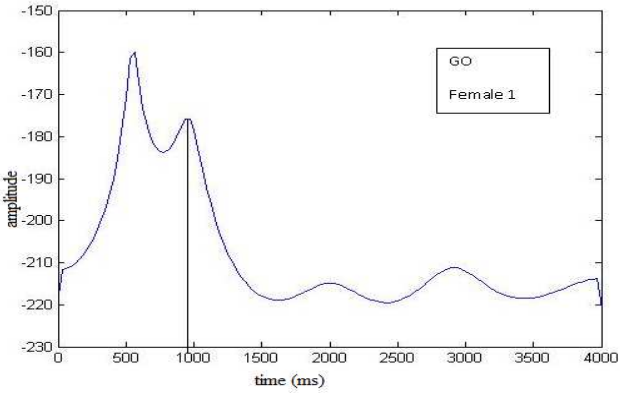
(a)



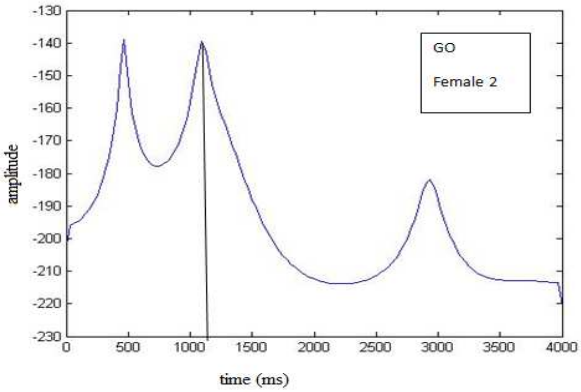
(b)

Fig. 5 (a) PSD of 'halt' for male 1 and (b) PSD of 'halt' for male 2 indicating 2nd formant

Figs. 6 (a), (b), 7 (a), (b) demonstrate the PSD spectra of two female, Female 1 and Female 2 for the word "GO" and "Halt". The locations of 2nd formants are indicated in the PSD curves. From the graphs, it is clear that locations of 2nd formants are different for different utterances.

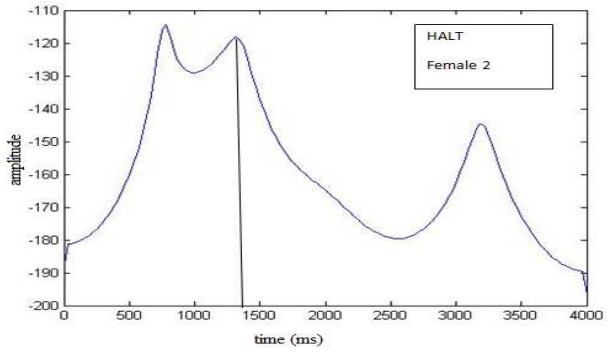


(a)

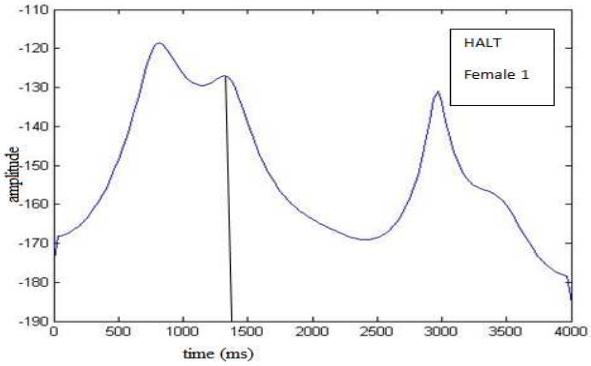


(b)

Fig. 6 (a) PSD of 'go' for female1 and (b) PSD of 'go' for female 2 indicating 2nd formant



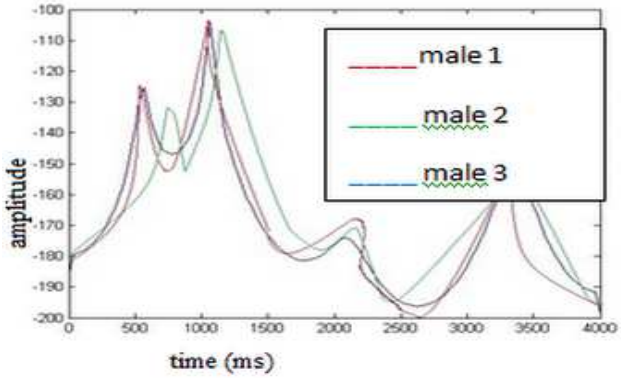
(a)



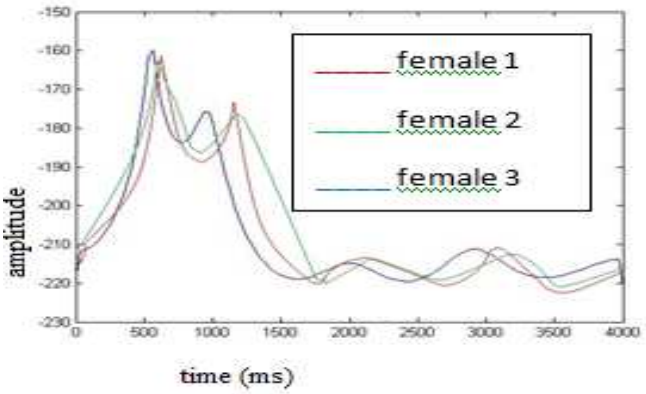
(b)

Fig. 7 (a) PSD of 'halt' for female 1 and (b) PSD of 'halt' for female 2 indicating 2nd formant

Three PSD spectra uttering 'Go' by three different male and female speakers are displayed in Figs. 8 (a) and (b) respectively. Here, these two figures are also representing almost the same value of 2nd formant for those three PSDs of those speakers.



(a)



(b)

Fig. 8 (a) PSDs of three samples of 'go' uttered by different persons male and (b) female

D. Simulation Environment

- Subjects' age 18-50 years
- 16 bits per sample speech resolution
- 8000 sampling frequency
- MATLAB is used as the simulation platform

E. Some Important Observations

We found this automatic speech recognition system as the percentage of recognition.

- Single-speech recognition which is uttered by different speakers.
- Single-speech recognition which is uttered by one speaker at a different time.
- Speech recognition by one or different speaker in a different environment.

For performance evaluation, the recognizer has been tested many times by inputting the same speeches command which is uttered by different speakers. Both male and female speaker have been given voice, sometimes a single speaker at different times. From various testing and implementation, it was found that this recognizer recognized almost 18 inputs out of 20 inputs successfully. So it can be said that it provides about 90% of the accuracy of specific speech commands in the recognition process.

IV. Conclusion

The values of pitch and formants of the individual voice samples can determine the gender of the speaker. Also, the values of the 2nd formants can determine the different words. Variation of accent and the environment of the lab or working station also vary the observation of the experiment of this speech recognition system. Different people from the different area has a different accent, and that make a lot of difference in the values of the second formants of the speech. Conversely, the 2nd formant reading for male and female utterances are sometimes slightly different. That is the reason of slightly overlapping of utterances of formants of different speeches.

One lacking of this work is that this speech recognition system cannot make a difference between words with same vowels. That is the reason for choosing the words with different vowel sounds for a better result. If there are two words with the same vowel sounds like “Right” and “Light”, the system will get confused, and the result might not be satisfactory. In future, this speech recognition system can do better with a large range of vocabulary in a more advanced way that is the hope.

Acknowledgment

Authors would like to express the deepest appreciation to A.H.M. Asadul Huq, Ph.D., Professor and Chairman, Electrical and Electronic Engineering, University of Dhaka, for the professional guidance. We also take the opportunity to express our heartiest thanks to some of the students of Electrical and Electronic Engineering Department of University of Dhaka and Islamic University of Technology, who provided us with the attention and support to collect the sample of data regarding this thesis. Finally, we like to thank UniSZA for financial support to publish this paper.

References

[1] B. Gold, N. Morgan, D. Ellis, “Speech and audio signal processing” Perception of Speech and Music, 2nd Ed., 2011 .

[2] M. Sahidullah, T. Kinnunen, "Local spectral variability features for speaker verification", Digital Signal Processing, 50:1–11, 2016. doi:10.1016/j.dsp.2015.10.011.

[3] E. Yucesoy, V. V. Nabiyev, “Gender identification of a speaker from voice source”, IEEE Proc. of 21st Signal Processing and Communications Applications Conference, vol. 4, no. 26, 2016.

[4] S. Nafisah, O. Wahyunggoro, L.E. Nugroho, “Mel-frequencies Stochastic Model for Gender Classification based on Pitch and Formant”, vol.6, no. 2, 2016. DOI: 10.18517/ijaseit.6.2.615.

[5] L. R. Rabiner, R. W. Schafer, “Digital Processing of Speech Signals.” Pearson Education, vol. 12, 2005.

[6] J. Kollewe, "HSBC rolls out voice and touch ID security for bank customers Business". The Guardian, (February, 2016.

[7] E. S. Gopi, “Digital Speech Processing Using Matlab”, Springer-Verlag, vol. 10.1007/978-81-322-1677-3, (2014.

[8] V. Evaldas, V. Antanas, G. Adas et.al. “Fusion of voice signal information for detection of mild laryngeal pathology”, Elsevier, vol. 18, 2014.

[9] K. Constantine, S. Stamatios, “Mobile phone identification using recorded speech signals”, IEEE International Conference on Digital Signal Processing, 2014.

[10] W. F. Lee; L. Jingsheng, L. Rynson et al. “Design and implementation of voice conversion system based on GMM and ANN”, Communications in Computer and Information Science Multimedia and Signal Processing, vol. 346, 2012.

[11] Y. Xu-Kui, H. Liang, Q. Dan, Z. Wei-Qiang, “Voice activity detection algorithm based on long-term pitch information”, EURASIP Journal on Audio, Speech, and Music Processing, vol. 1, 2016.

[12] F. Müller; A. Mertins, “Contextual invariant-integration features for improved speaker-independent speech recognition”, Elsevier, Speech Communication, vol. 53 , 2011.

[13] J. Gao, X. D. Chongqing, “Noise-robust pitch detection algorithm based on AMDF with clustering analysis picking peaks”, IEEE Information Technology, Networking, Electronic and Automation Control Conference , vol. 7560544, 2016.

[14] B. D. Argo, Y. Hendrawan, D. F. Al-Riza et al. “Optimization of PID Controller Parameters on Flow Rate Control System Using Multiple Effect Evaporator Particle Swarm Optimization, International Journal on Advanced Science, Engineering and Information Technology, vol.5, no. 2, 2015. DOI:10.18517/ijaseit.5.2.491 .