Speaker Specific Formant Dynamics of Vowels

Sharada Vikram Chougule(✉)

Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India
shardavchougule@gmail.com

Abstract. Automatic Speech and Speaker Recognition technology has growing demands in variety of voice operated devices. Although the input for all such systems is speech signal, the features useful for each application/task are different. Of the different speech sounds, vowel sounds spectrally well-defined and well represented by formants. Formants which represent resonances of vocal tract are the result of physiology of individual's speech production mechanism as well as nature of speech (words) being spoken. In this way formants are features of speech as well as of speaker. In this paper significance of formants for speech and speaker recognition is explored through experimental analysis. Formant tracking and estimation is done using adaptive formant filter bank and single pole formant based filter. Twelve vowel sounds represented in ARPABET (Advanced Research Project Agency bet) form are used to estimate the first four formants. The analysis based on extracting and emphasizing speaker specific clues indicates that higher formants carry more speaker specific information than first (lower) formant.

Keywords: Automatic Speech and Speaker Recognition · Vowels · Formants

1 Introduction

Automatic Speech and Speaker Recognition technology has growing demands in variety applications such as automated customer services, healthcare applications, mobile banking, voice operated devices, education etc. Although the input for all such systems is speech signal, the features useful for each application/task are different. For speech recognition the purpose is to emphasize '*what*' is being spoken ignoring who is speaking, whereas in speaker recognition the emphasis is on '*who*' is speaking irrespective of contents of speech. Thus speech signal carries characteristics of both 'speech' as well as 'speaker'.

Human speech is result of physiology of vocal apparatus and articulatory features. During speaking, the short spurts of air are taken in lungs and released in a controlled manner through trachea (windpipe) towards vocal cords. The lungs and associated muscles act as source of air for exciting the vocal mechanism [1]. Voiced speech sounds are produced because of periodic vibration of tensed vocal cords. Unvoiced sounds are the result of relaxed state of vocal cords and constriction of vocal tract. Thus, the manner in which different speech sounds are produced is dependent on vocal structure and position of various articulatory elements (e.g. tongue, teeth, lips etc.).

Human vocal tract is comprised of larynx to the lips (or nose), which takes different length, shapes and cross-sections while speaking. One purpose of vocal tract is 'shape'

S. M. Thampi et al. (Eds.): SIRS 2019, CCIS 1209, pp. 118–129, 2020.
https://doi.org/10.1007/978-981-15-4828-4_11

or 'modulate' the source making perceptually distinct speech sounds [2]. Vowel are generally long induration and can be distinguished better through spectrum (frequency domain analysis). Vowels are characterized by which represent distinct peaks of acoustic energy. Thus, formants or formant frequencies are resonances of vocal tract represented as peaks in the spectrum of vocal tract response, which are well distinguished during vowel sounds. In spite of specific properties of different vowels, there is much variability of vowel properties among the speakers because of articulatory differences [3].

The basic purpose in this work is to explore the characteristics of vowel sounds representing both speech as well as speaker in a distinct manner. Studies in literature [6–8] used short time power spectrum, Linear Prediction (LP) spectrum, mel-spectrum (holomorphic analysis) to extract typically first four to five formant frequencies. Similar spectral and cepstral features like Linear Prediction Cepstral Features (LPCC) and Mel Frequency Cepstral Coefficients (LPCC) are widely used as features for speaker recognition which mainly relate to vocal tract parameters and human perception of speech sounds [9–11].

In this paper, speaker specific formant dynamics of vowel sounds is studied for Twelve vowel sounds represented in ARPABET (Advanced Research Project Agency bet) form generated by [13]. Formant tracking and estimation is done using adaptive formant filter bank and single pole formant based filter. Section 2 discusses the methodology used for formant tracking and estimation. Results and discussion are given in Sect. 3 and conclusion in discussed in Sect. 4.

2 Methodology for Formant Tracking and Estimation

In this work, Linear Prediction Coefficients (LPC) are most commonly used method to represent the vocal tract parameters (formants) through its spectrum. Vowel parameters are useful for analysis and discrimination of speech patterns for diseases such as neurodegenerative disorder [12]. The disadvantage of using simple LPC to model the speech sounds is that the spectral and temporal characteristics of speech signal varies because of physiological, behavioral characteristics as well as by transducer/channel effects. Such variabilities present a challenge to speech as well as speaker recognition. To minimize the variabilities in LP spectrum and to better track and estimate the formant frequencies, methodology based on work in [14] and [15] is used. The steps in formant estimation algorithm are as follows:

i. Pre-emphasis
ii. Formant tracking filters
iii. Voicing detector and spectral estimation
iv. Decision maker

2.1 Pre-Emphasis

Speech is a quasi-stationary signal, in which high frequency components are having lower energy than low frequency components. Also high frequency speech components

get easily distorted with noise because of low SNR. Pre-emphasis improves the SNR and helps to boost the energy of high frequency speech signals. Pre-emphasis filter is first order Butterworth IIR High pass filter. The real valued pre-emphasized speech signal is converted into analytic signal using Hilbert transformer in order to design complex valued filters.

2.2 Formant Tracking Filters

As speech is quasi-stationary signal and having transients in harmonic frequencies especially during voiced speech, fixed filter bank will not be reasonable. Therefore an adaptive filter bank proposed in [13] and [14] used to track narrowband speech signal. Also in practice, there is leakage or aliasing of neighboring formants, which affects the estimate of desired formant location and creates variability in true estimate. Prior to estimating the formant frequencies, the speech signal passed through a set of adaptive band pass filters, also known as formant filters. The band pass filters are basically complex all-zero filters designed such that complex zeros are placed at the formant location other than single formant which is to be estimated. For example to estimate 4th formant (F4), the band pass filters have complex-zeros at the location of first (F1), second (F2) and third formants (F3) respectively. These filters are designed to have normalized gain and zero phase lag at the filter center frequency [13].

The zeros of all zero filters are given by:

$$z_k = r_z \, e^{\frac{j2\pi F_k}{F_s}} \tag{1}$$

Here r_z indicates distance of zero from origin and angle $\theta = \frac{j2\pi F_k}{F_s}$ decides location of zero in z-plane.

Thus, the transfer function of k^{th} all-zero formant filter for k = 1,2,3,4 is given by [13]:

$$H(z,n) = k_k(n,z) \prod_{l=1, l\neq k}^{4} 1 - r_z e^{-j2\pi Fl(n)} z^{-1} \tag{2}$$

Here $r_z = 0.98$. The location of zeros at formant frequencies ensures minimum response of formant filters except for the k^{th} formant. Unity gain and zero phase lag response at estimated formant frequency at k^{th} component is obtained by the term k_k(n, z).

$$k_k(n) = \frac{1}{\prod_{\substack{l=1 \\ l\neq k}}^{4} 1 - r_z * e^{-j2\pi f_l(n) - f_k(n)}} \tag{3}$$

A single-pole (IIR) filter is cascaded after each all zero FIR filter to estimate peak in the respective filter output. The transfer function of the first order peak tracking IIR filter is given by:

$$H_{DTk}(z) = \frac{1 - b_k}{1 - b_k \, e^{-j2\pi Fl(n)} z^{-1}} \tag{4}$$

where $b_k = 0.9$ is pole radius, $(1 - b_k)$ is DC gain and F_l is l^{th} formant frequency.

The initial set of estimated formant frequencies is updated over time using previous formant frequency estimate and complete set of filters is named as formant tracking filters. The analytic speech signal is filtered by these four formants tracking filters.

2.3 Voicing Detector and Spectral Estimation

Voicing detector based on zero crossing rate (ZCR) [16], [17] is used to detect the voiced frame of speech. Analytic speech signal is sampled at 8 kHz sampling frequency and framed using 20 ms Hamming window (160 samples). The LPC provides good model for voiced regions of speech signal in which all-pole model of LPC provides a good approximation to the vocal tract spectral envelope. Speech sample at a time s(n) can be approximated as linear combination of the past p speech samples such that [1]:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G * u(n) \tag{5}$$

where a_i are weights of samples, G is gain of excitation and u(n) is normalized excitation, leading to p^{th} order all-pole system as:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \tag{6}$$

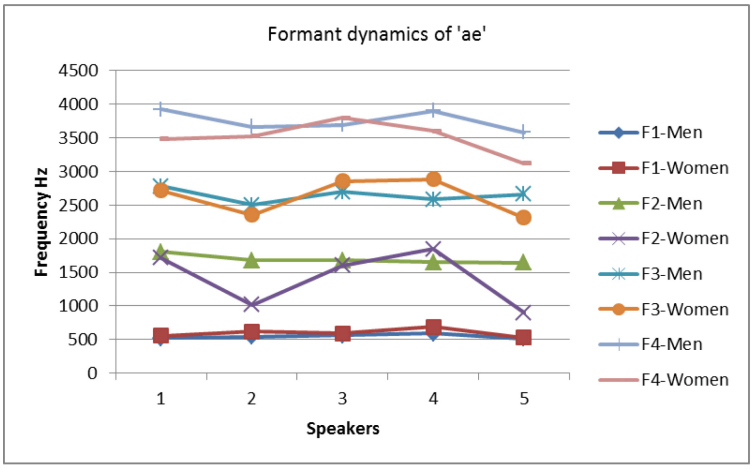
Here a single linear predictive coefficient (LPC) of the windowed frame is calculated for each band using the autocorrelation method [2, 3] for each band. The LPCs are only calculated from the bands if the entire previous 20-ms window of the speech signal is voiced. Moving average value of LPC estimated current formant frequency over the frame is calculated based on decision of voicing detector. Temporal changes in the envelope structure are often derived from estimated formant tracks [18]. Formant based vowel perceptual space classification is possible using LPC spectrum [19]. The correlation of LP analysis with human speech production mechanism is also being used to identify an individual (speaker identification) irrespective of content of speech [20].

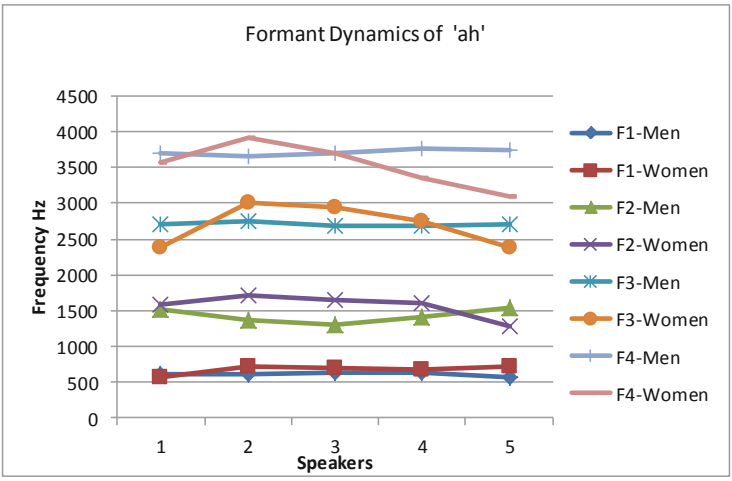
3 Results and Discussion

Speech samples from 10 speakers (five women, five men) are used for experimental analysis. Formant estimation and analysis is carried out on twelve vowel sounds represented in ARPABET form using the method discussed Sect. 2. Table 1 gives the ARPABET

Table 1. Vowel codes used for formant analysis

ae- "had"	ah- "hod"	aw- "hawed"	eh- "head"	ei- "haid"	er- "heard"
ih- "hid"	iy- "heed"	oa- "hoat"	oo- "hood"	uh- "hud"	uw- "who'd"





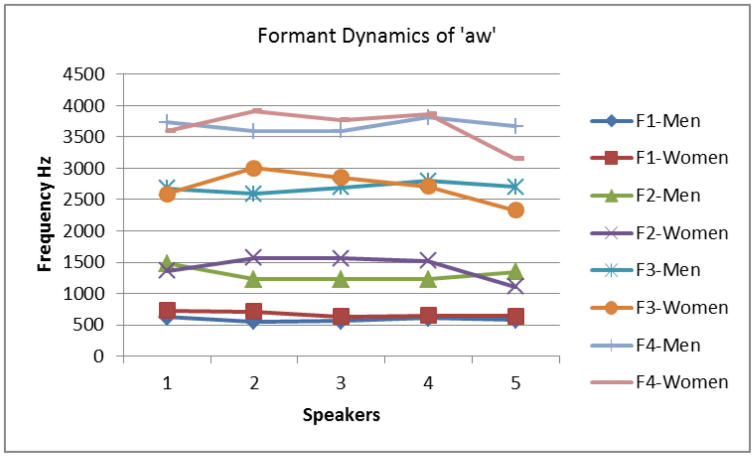
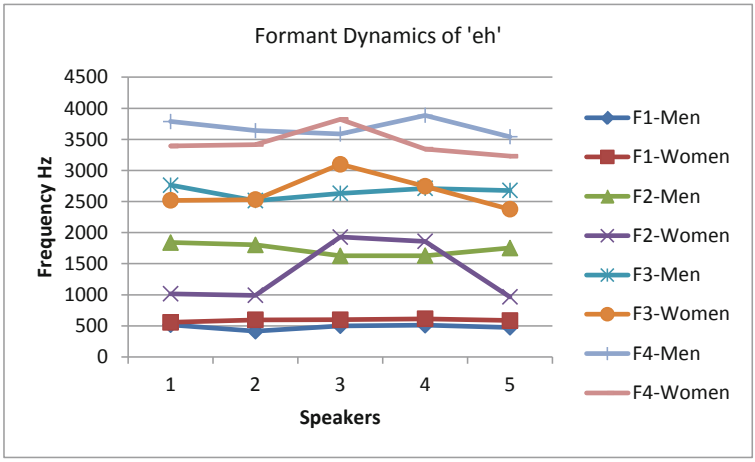
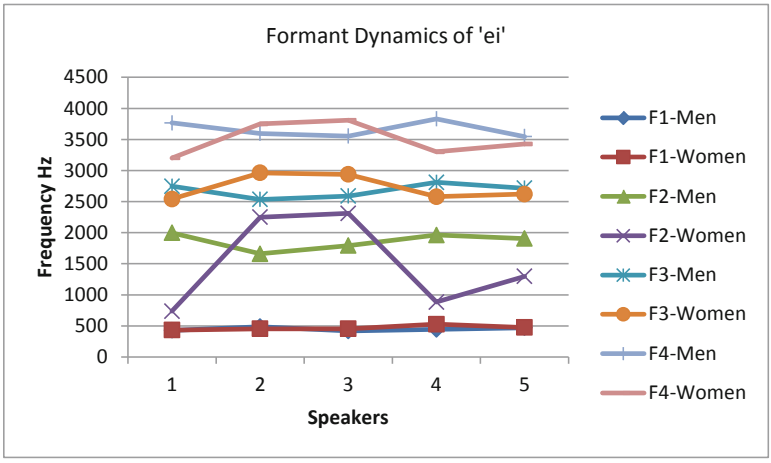


Fig. 1. Plots of Formant dynamics of vowels among Men and Women Speech samples





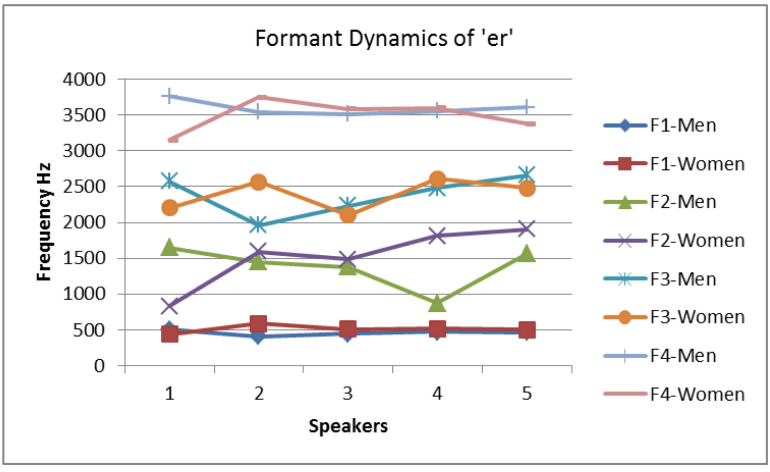
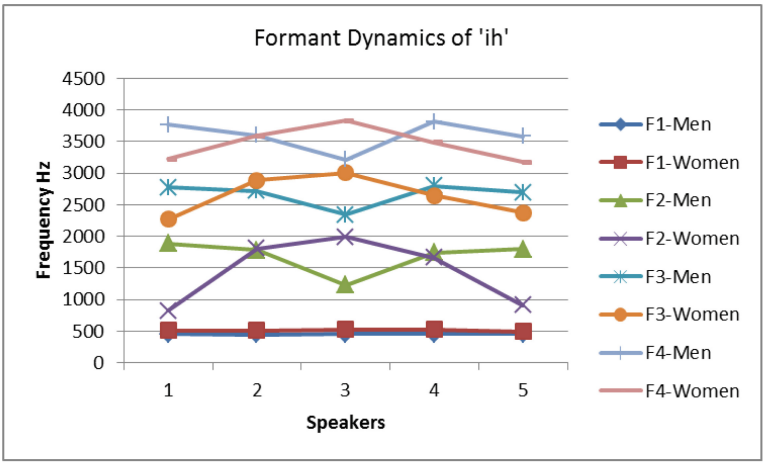
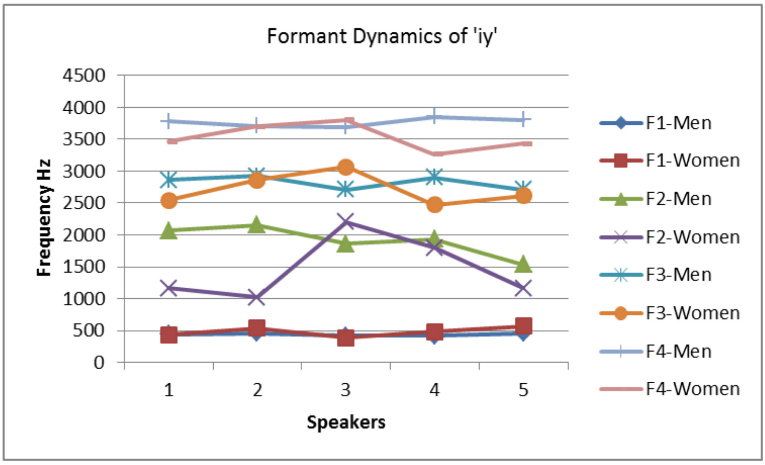


Fig. 1. (*continued*)





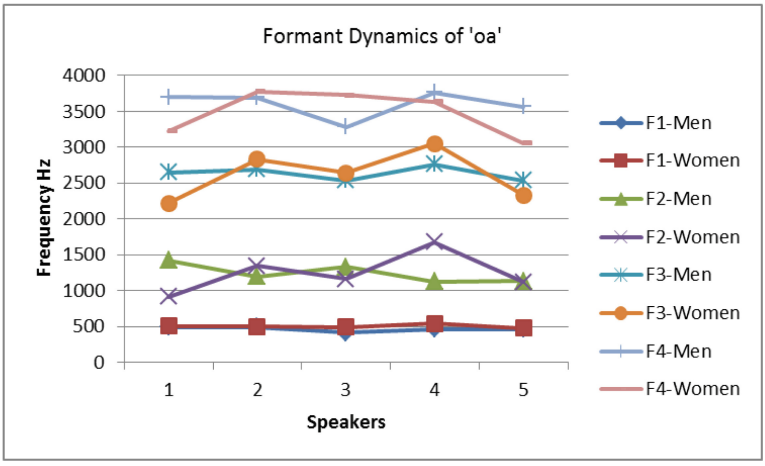
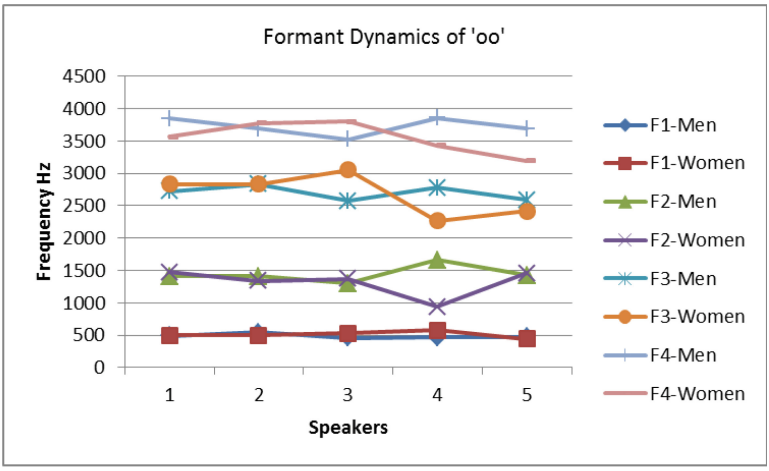
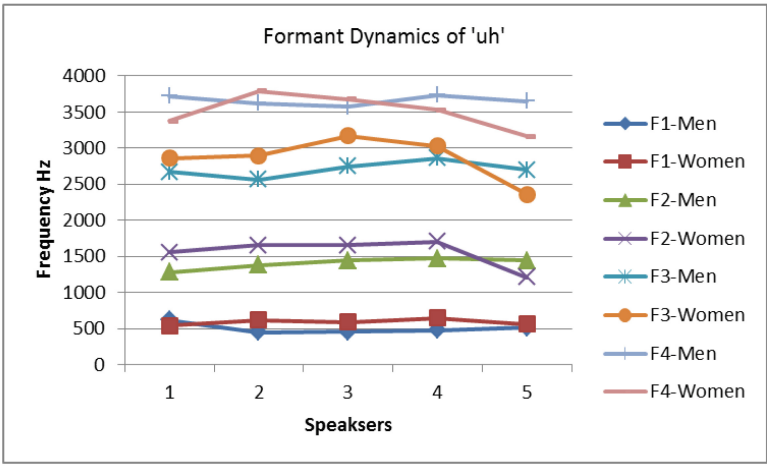


Fig. 1. (*continued*)





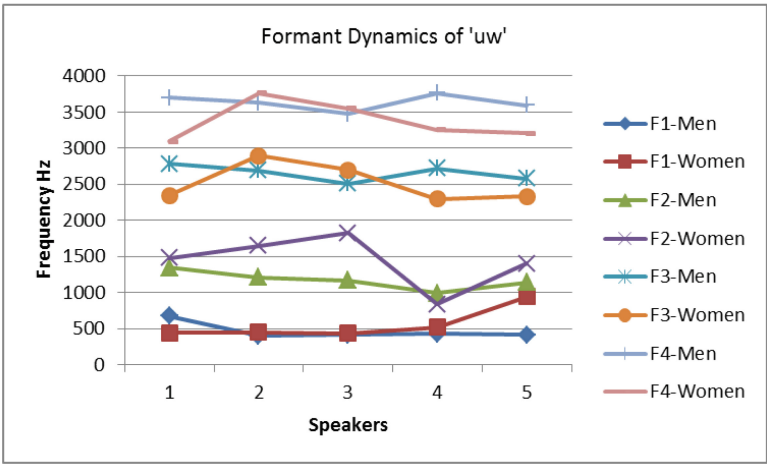
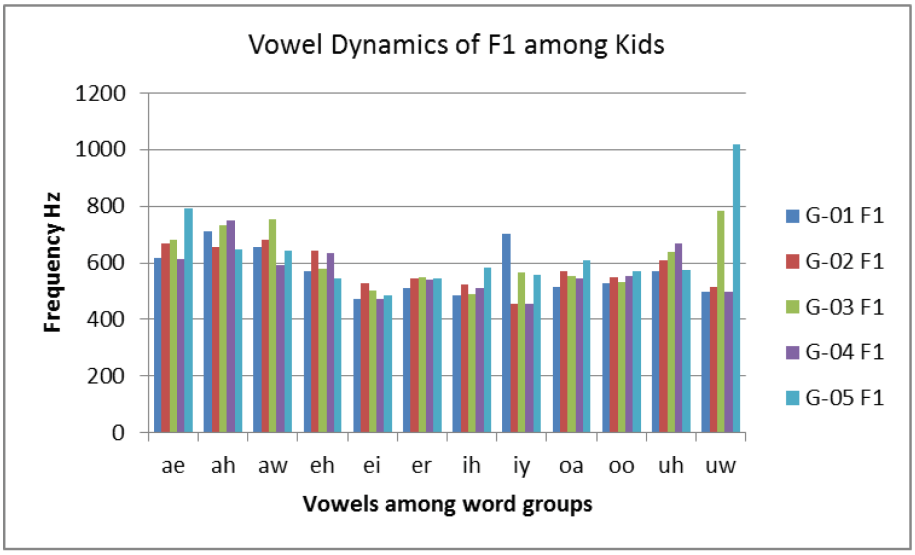


Fig. 1. (*continued*)



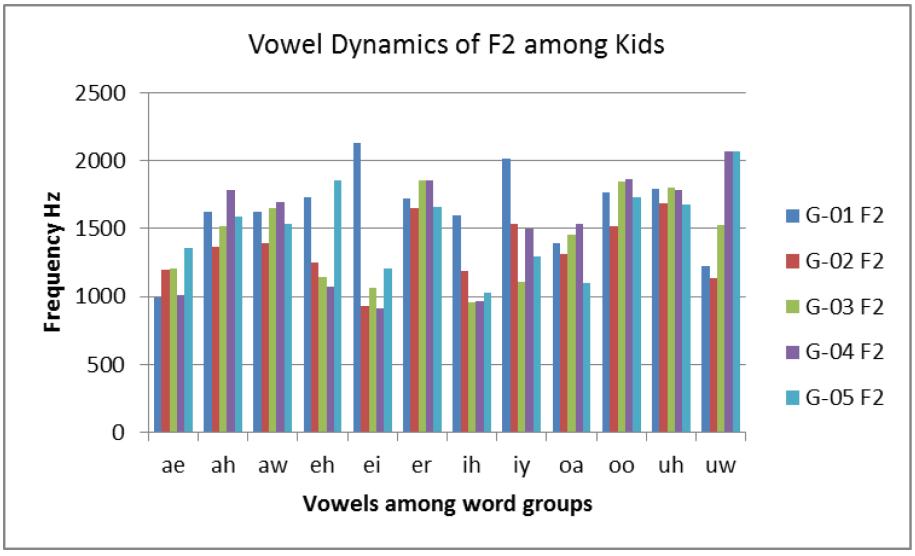
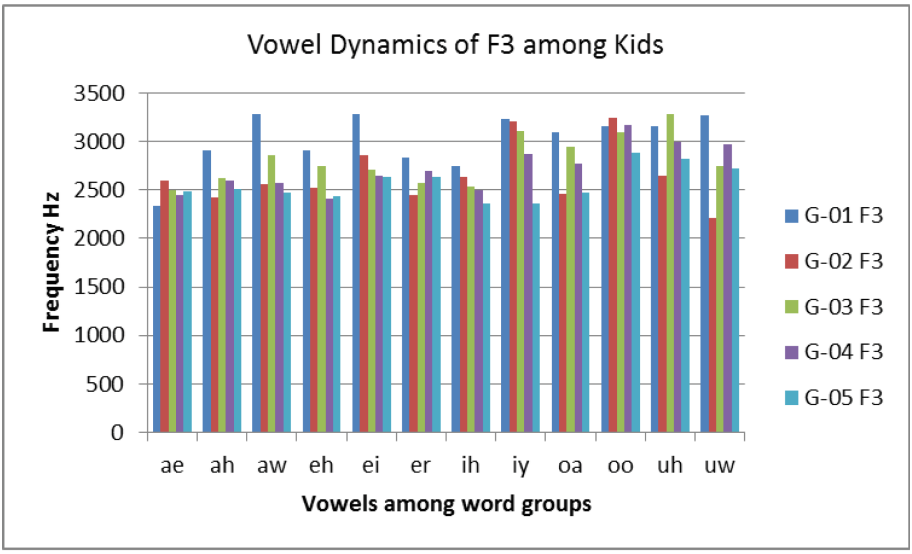


Fig. 2. Plots of Formant dynamics of vowels among Kids (Girls) Speech Samples

symbols of 12 vowels and vowel code used for each group [13]. The plats shows the estimated formants (F1 to F4) for ten of speakers for each vowel code. For comparison of characteristics of male and female speakers, formant plots of speech of both men and women are shown in same plot.

Figure 1 shows the plots of first four formants estimated for 05 male and 05 female speakers for 12 code (word) groups of vowel data. It was observed from the plots that, there is less deviation of first formant (F1) among the speakers for all vowels. Also, the



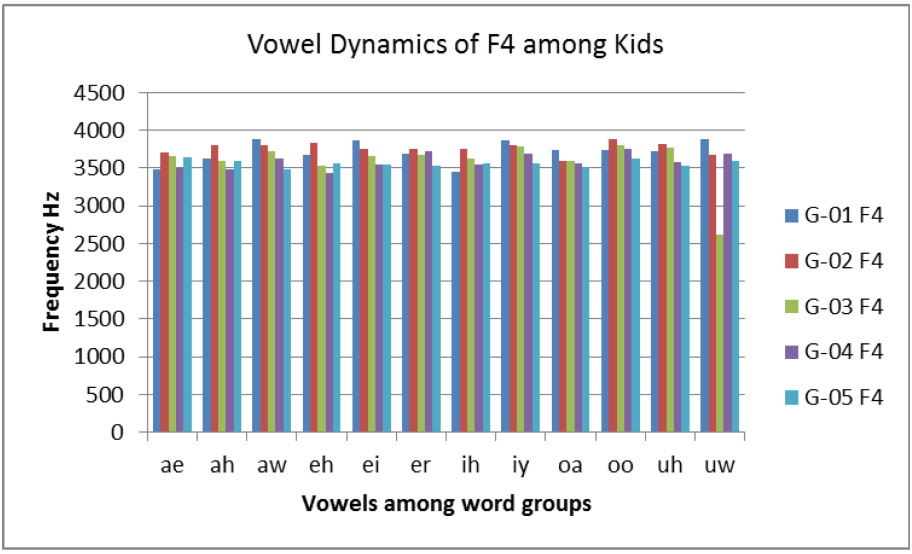


Fig. 2. (*continued*)

frequency of first formant for female speakers is much greater than that of male speakers. In general, female speaker formants are higher than male speakers. There is substantial variation in second formant except vowels codes ah, aw and uh. Higher formants of all vowels are incomparable for different speakers, i.e. there is speaker specific variation of higher formants.

Figure 2 shows the plots of formant dynamics (F1 to F4) of speech samples of five kids (girls). As compared to results in Fig. 1, there is more variation of first formant

frequency of kids speaker for all vowel types, whereas first formant is observed to be nearly constant among men and women speakers. Also average formant frequencies among children is higher that of men and women.

4 Conclusion

The purpose of the work carried in this work is to analyze speaker specific formant dynamics of vowel sounds. For the experimental work a set words having of twelve vowel sounds are used. The reason behind using vowel speech data is, vowels carry most significant information related to speech patterns as well as properties of vocal mechanism of individual. In order to get true estimate of the formant frequencies, formant filters and voicing detector are used prior to LPC analysis. The first formant consistently carries the information of speech phoneme, over the speakers. The higher formants vary with respect to speaker for each vowel. The deviation is larger for third and fourth formant (F3 and F4). This variability in higher formants can be used as key point for speaker or voice recognition. Thus, similarity in estimated formants for a particular vowel can be used as feature for speech recognition, whereas variation or differences in estimated formants for same vowel sounds as a feature for speaker recognition.

References

1. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River (1993)
2. Deller, J., Hansen, J., Proakis, J.: Discrete-Time Processing of Speech Signals. IEEE Press, New York (2000)
3. Quatieri, T.F.: Discrete-Time Speech Signal Processing: Principles and Practice. Pearson Education, Third Impression (2007)
4. Flanagan, J.L.: Speech Analysis, Synthesis, and Perception, 2nd edn. Springer-Verlag, New York (1972). https://doi.org/10.1007/978-3-662-01562-9
5. Welling, L., Ney, H.: A model for efficient formant estimation. In: Proceedings of the IEEE ICASSP, Atlanta, pp. 797–800 (1996)
6. Welling, L., Ney, H.: Formant estimation for speech recognition. IEEE Trans. Speech Audio Process. **6**(1), 36–48 (1998)
7. Holmes, J.N., Holmes, W.J.: The use of formants as acoustic features for automatic speech recognition. In: Proceedings IOA, vol. 18, part 9, pp. 275–282, November 1996
8. Rao, P., Das Barman, A.: Speech formant frequency estimation: evaluating a nonstationary analysis method. Sign. Process. **80**, 1655–1667 (2000)
9. Anusuya, M.A., Katti, S.K.: Speech recognition by machine: a review. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **6**(3), 501–531 (2009)
10. Kinnunen, T., Li, H.: An overview of text independent speaker recognition: from features to supervectors. J. Speech Commun. **52**(1), 12–40 (2010)
11. Reynolds, D.A.: An over view of automatic speaker recognition technology. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4 (2002). MIT Lincoln Laboratory, Lexington, MA USA
12. Mirarchi, D., Vizza, P.: Signal analysis for voice evaluation in Parkinson's disease. In: 2017 IEEE International Conference on Healthcare Informatics (2017)

13. Vowel Data: James M. Hillenbrand, Speech Pathology and Audiology, Western Michigan University
14. Rao, A., Kumaresan, R.: On decomposing speech into modulated components. IEEE Trans. Speech Audio Process. **8**(3), 240–254 (2000)
15. Mustafa, K., Bruce, I.C.: Robust formant tracking for continuous speech with speaker variability. IEEE Trans. Audio Speech Lang. Process. **14**(2), 435–444 (2006)
16. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signal, 1st edn. (2003)
17. Yang, X., Tan, B., Ding, J., Zhang, J., Gong, J.: Comparative study on voice activity detection algorithm. In: IEEE International Conference on Electrical and Control Engineering (2010)
18. Craciu, A., Paulus, J., Sevkin, G., Backstrom, T.: Modeling formant dynamics in speech spectral envelopes. In: 25th European Signal Processing Conference (EUSIPCO) (2017)
19. Dey, S., Ashraful Alam, Md.: Formant based Bangla vowel perceptual space classification using support vector machine and k-nearest neighbor method. In: 21st International Conference of Computer and Information Technology (ICCIT) (2018)
20. Almaadeed, N., Aggoun, A., Amira, A.: Text-independent speaker identification using vowel formants. J. Sign. Process. Syst. **82**, 345–356 (2016)