Analyzing Vocal Tract Parameters of Speech

Sharada Vikram Chougule

Abstract Speech sounds produced by human depends on movement of various articulators. Dimensions and shape of the various elements of speech production organs also have impact on nature of speech produced. Vocal tract plays a major role like characteristics of linear time invariant system. In this paper, analysis of vocal tract parameters, in terms of resonances of vocal tract, also referred as formants is done. Of the different speech sounds, vowels carry most significant clues of speech. Despite specific properties of different vowels, there is much variability of vowel characteristics among speakers. It gives characteristics of vocal tract related to acoustic resonances. At large, first four formants are useful to categorize vowel sounds. The variation in formants for same speech sounds is a challenge to speech recognition algorithms in which vowel spectral characteristics are assumed to be invariant among speakers. But the same variability of formants among speaker is useful in speaker recognition. The method used for formant estimation is based on designing all-zero filters to track the formants and voicing detection-based formant extraction filters to estimate the first four formants. Vocal tract parameters in terms of formants are analyzed using twelve vowel sounds from different speakers. From the experimental analysis, it is observed that the first formant specifically represents the pertinent characteristics of vowel speech, whereas there is very little consistency of higher-order formants for same speech sounds by different speakers, indicating the impact of physiological nature as well as behavioral aspects of individual on nature of speech produced.

Keywords Resonances · Formants · Speech recognition · Speaker recognition

1 Introduction

Speech signal is a result of variations in articulatory movements and is prone to include variability such as phonetic contents and distressing traits, characteristics

S. V. Chougule (✉)
Finolex Academy of Management and Technology, Ratnagiri, India
e-mail: shardavchougule@gmail.com

S. N. Merchant et al. (eds.), *Advances in Signal and Data Processing*,
Lecture Notes in Electrical Engineering 703,
https://doi.org/10.1007/978-981-15-8391-9_27

of individual human speech production structure, and sometimes behavioral state of speaker while speaking [1]. The wider variability in speech signal of individual's speech leads to automatic speech as well as speaker recognition a challenging. The physiology of human speech production system is the fundamental aspect of characterizing speech sounds. From practical perspective, human speech production starts from vocal cords (vocal folds) and end at mouth (lips) or nose.

Practically, the voice production system can be modeled as connected auditory pipe having some peaks (called as formants) as well as valleys generated based on nature of speech. Formants are basically the perception characteristics of vowels, in which concentration of acoustic energy is observed at certain frequencies, which are called as first formant, second formant, and so on. Formants represent characteristics of speech sound, which involves large dynamics because of physiological as well as behavioral aspects of a person while speaking.

Research have been carried out based on use of formant frequencies for automatic recognition of speech. The main approaches adopted in the literature are linear prediction [1–3] investigation, and analysis of speech using Fourier spectrum [4], and using features such as peaks of homomorphically smoothed cepstrum [5]. Vowels which typically have voiced characteristics are the most useful speech sounds in reliable formant estimation [6].

In this paper, analysis of formant estimation model is investigated for vowel sounds. The algorithm proposed is to track the most prominent formants considering the variabilities in speech during speech production. Following section discusses the methodology to track and estimate the formants from different speech sound.

2 Formant Estimation Algorithm

Input analog speech signal is converted in discrete form using a sampling frequency of 8 kHz and framed using 20 ms Hamming window. Pre-emphasis is performed using Butterworth IIR high pass filter. Pre-emphasis helps to reduce spectral tilt and improves spectral flattening providing more gain for high-frequency components. Further, this pre-emphasized signal is passed through Hilbert transform (all-pass filter) to create an analytic signal from a real signal. A set of adaptive FIR (all-zero) bandpass filters with linear phase characteristics is designed and cascaded with formant filter. Before estimating individual formant, speech signal is filtered out using a set of bandpass filters (filter bank). The most recent formant estimates are used to update the magnitude response of filters. This allows tracking of individual formant frequency over time, and in suppression of nearby formants and intrusion of surrounding noise.

The center frequency of these bandpass filters are first formant (F1):0.7 kHz, second formant (F2): 1.5 kHz, third formant (F3): 2.2 kHz, and fourth formant (F4):3 kHz, respectively. These four formants are spectrally separated using the adaptive filter bank. To isolate pitch frequency (F0) from the first formant (F1), additional zero is placed at F1 filter transfer function. The Hilbert transformed signal gives complex-valued filter coefficients, which help in designing filters with normalized gain and zero phase characteristics at the center frequency of each filter.

The kth all zero formant filter transfer function for $k = 2, 3, 4$ is given by [7]:

$$H_{Fk}(z, n) = k_K(n, z) \prod_{l=1, l \neq k}^{4} 1 - r_z \mathrm{e}^{-j2\pi Fl(n-1)} z^{-1} \tag{1}$$

Here, $r_\mathrm{z} = 0.98$. Above equation of filter transfer function ensures minimum response of formant filters except for the kth formant. The term $k_k(n, z)$ ensures normalized magnitude response and zero phase characteristics of kth estimated frequency component.

$$k_K(n) = \frac{1}{\prod_{l=1, l \neq k}^{4} 1 - r_z \mathrm{e}^{-j2\pi Fl(n-1) - Fk(n-1)}} \tag{2}$$

An supplementary zero is added in the transfer function of first formant filter, with zero at pitch frequency at 200 Hz. This zero is to prevent interference of pitch frequency to first formant. Thus, transfer function of first formant frequency filter is given by:

$$H_{Fk}(z, n) = k_1(n) \prod_{l=0, l \neq 1}^{4} 1 - r_z \mathrm{e}^{-j2\pi Fl(n-1)} z^{-1} \tag{3}$$

where

$$k_1(n) = \frac{1}{\prod_{l=0, l \neq 1}^{4} 1 - r_z \mathrm{e}^{-j2\pi Fl(n-1) - F1(n-1)}} \tag{4}$$

The signal filtered through all-zero FIR filter is further passed through a set of first-order IIR filter. The pole of each of these filters is updated based on formant frequency estimated in previous frame of that filter. The transfer function of kth single-pole IIR filter at time instant n is as below:

$$H(n, z) = \frac{1 - r_\mathrm{p}}{1 - r_\mathrm{p} \mathrm{e}^{j2\pi Fk(n-1)} z^{-1}} \tag{5}$$

Here, $r_\mathrm{p} = 0.9$ defines radius of pole, which decides the magnitude/gain at a formant frequency, and estimation of kth formant filter at index $(n - 1)$ is given by $F_K(n - 1)$. Equation (5) gives the design of four formant filters having complex-valued coefficients. Thus, these filters divides the spectrum of Hilbert transformed speech signal into four spectrally separated regions and estimate the formant frequencies based on updated filter coefficients [8].

The analytic speech contains all types of phonetic contents, out of which voiced part is most useful in detecting speech specific formants [9, 10]. A voicing detector is used to distinguish voiced and unvoiced frame, as formants are the characteristics related to voicing properties of the speech, generally of vowels. A simple zero cross-rate detector (ZCR) is used to classify voiced and unvoiced part of the speech. A simple measure of zero cross-rate is the count that the speech signal crosses zero (reference) amplitude. As a general observation, unvoiced or noisy speech is having more ZCR than voiced speech. It is calculated using signum function as:

$$z_n = \sum_{n=-\infty}^{\infty} |\mathrm{sgn}[(s(n)] - \mathrm{sgn}[s(n-1)].w(m-n) \tag{6}$$

nn

Here, sgn() is signum function which is 1 for $n \geq 0$ and -1 for $n < 0$, and $w(n)$ is window function of N samples. Thus, the formant frequencies are estimated over each speech frame based on decision of voicing detector. This reduces the redundant estimation of formants during unvoiced or silence part of speech.

3 Results and Discussion

Figure 1 shows the narrowband spectrogram of two different women speech samples of vowel 'ae' extracted from the formant estimation algorithm discussed in Sect. 2. It is observed that variation of first formant is almost similar over the entire time duration, and the position (frequency) of higher formants is different for the same speech sound. Similar analysis is carried of for 12 vowels sounds from speech samples of five female speakers.

The results in Fig. 2 show the analysis plots of first four formants of 12 different vowel sounds and its variation among five female speakers notes as W1 to W5.

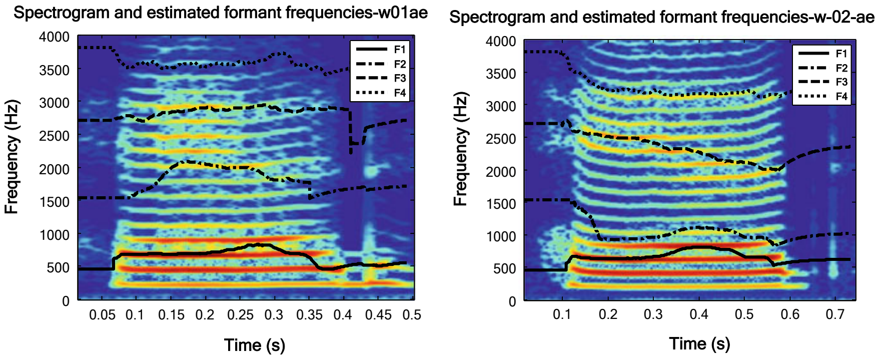

Fig. 1 Narrowband spectrogram of two female speakers for vowel sound 'ae'

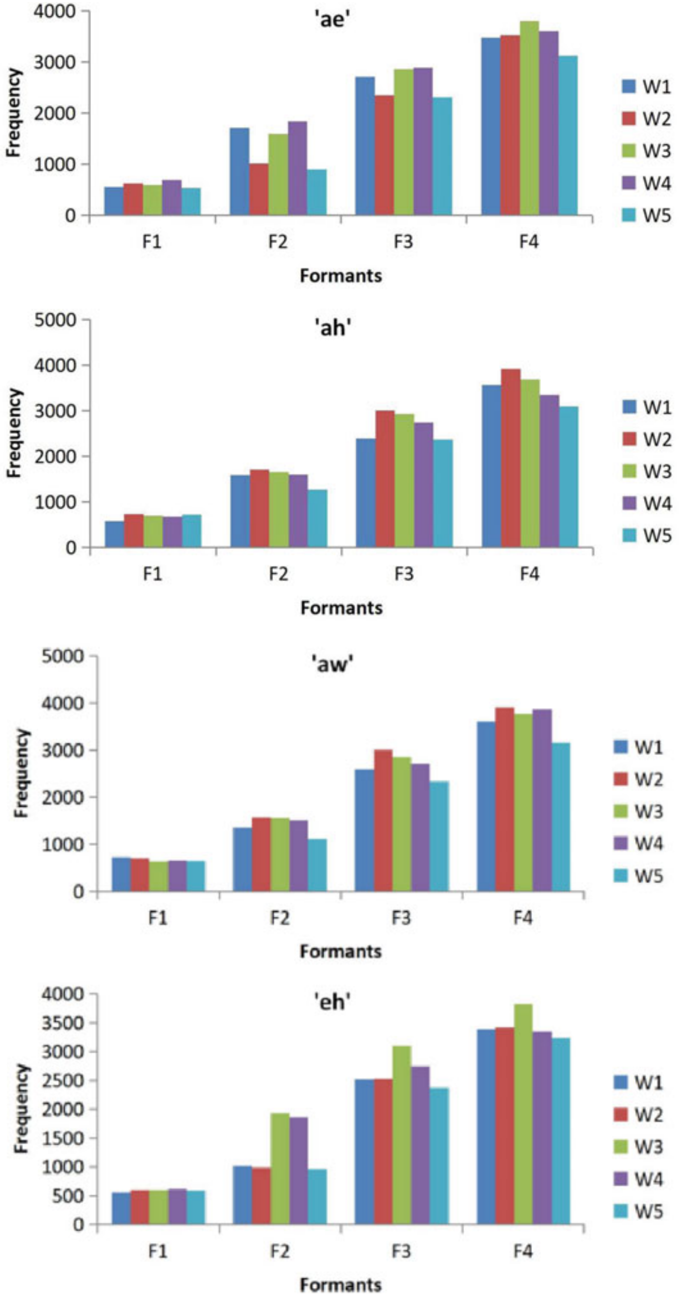

Fig. 2 Formant analysis of vowel sounds of five female speakers

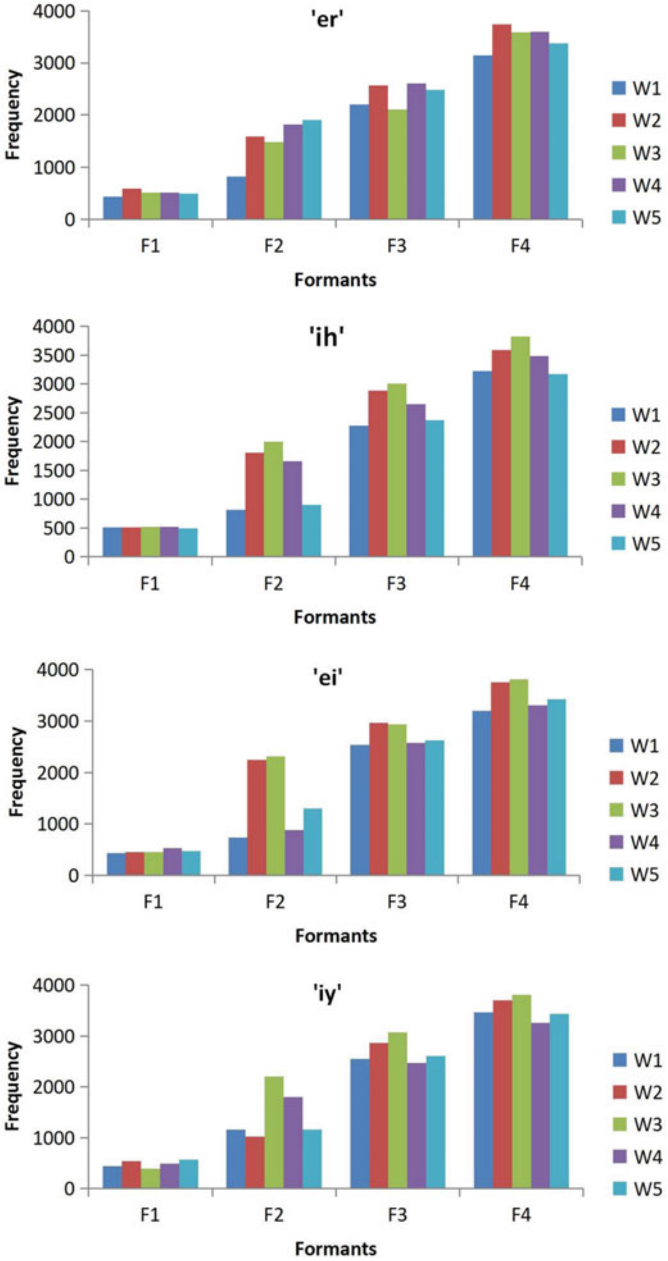

Fig. 2 (continued)

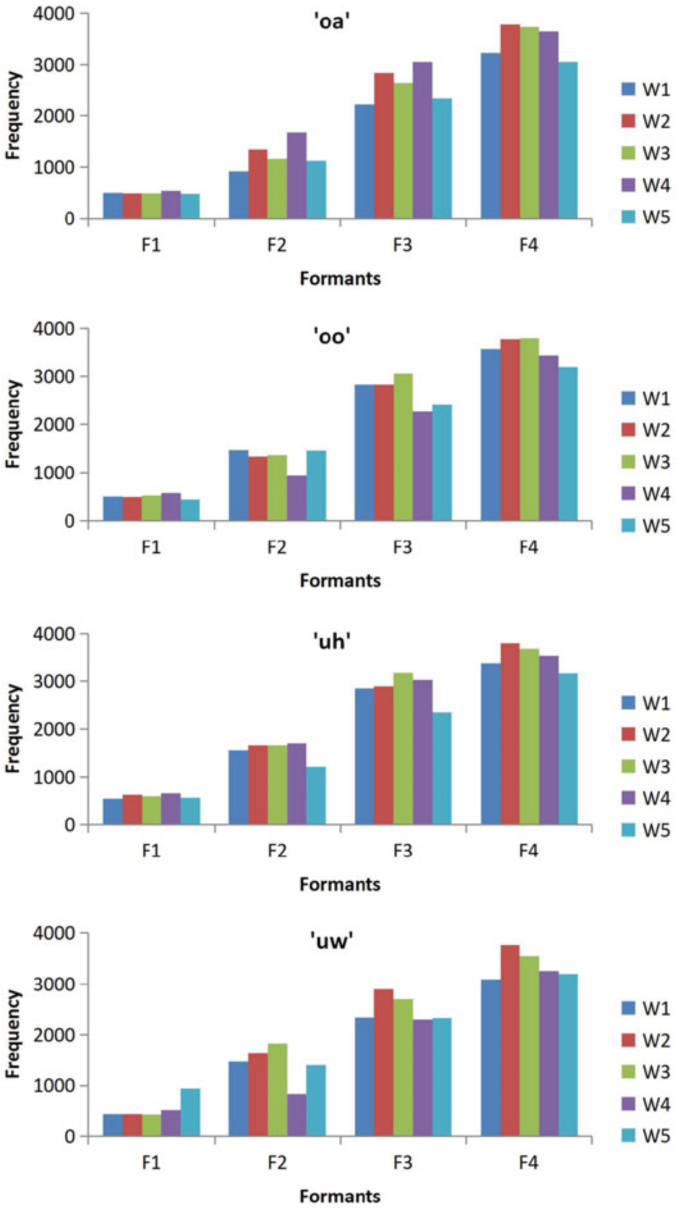

Fig. 2 (continued)

The above plots show the variation of formants for 12 different vowel sounds by five female speakers. From the experimental results plotted, it is observed that the first formant, i.e., F1 (excluding the fundamental frequency), for each vowel sound is almost constant over the speakers. All higher formants (F2, F3, F4) for a vowel single sound vary considerably, over the speakers. These characteristics of formants can be used for formant as a feature for speech as well as speaker recognition. Thus, from experimental analysis, it can be concluded that first formant is most appropriate feature of speech sounds, whereas higher formants represent the characteristics of speaker.

4 Conclusion

In this paper, formant analysis of vowel sounds is done to explore the significance of formants for speech as well as speaker recognition. A set of twelve vowel sounds is used for experimental analysis. The purpose of the experimental analysis is to study and explore the significance of formants in relation to characteristics of speech sounds and speakers. As an initial study, experimentation is carried out on five female speech samples. Using the methodology described, the work will be extended for continuous speech, for a large speaker database and in a variety of real-world dynamic conditions for applications such as speech recognition and speaker recognition.

References

1. Rabiner LR, Juang BH (1993) Fundamentals of speech recognition. Prentice-Hall (1993)
2. Deller J, Hansen J, Proakis J (2000) Discrete-time processing of speech signals. IEEE Press
3. Quatieri TF (2007) Discrete-time speech signal processing: principles and practice, Third impression. Pearson Education
4. Welling L, Ney H (1996) A model for efficient formant estimation. In: Proceedings of IEEE ICASSP, Atlanta, pp 797–800
5. Holmes JN, Holmes WJ (1996) The use of formants as acoustic features for automatic speech recognition. In: Proceedings of IOA, vol 18, part 9, pp 275–282
6. O'Shaughnessy LDD. Speech processing, a dynamic and optimization- oriented approach. Marcel Dekker Inc. New York, NY, USA
7. Mustafa K, Bruce IC (2006) Robust formant tracking for continuous speech with speaker variability. IEEE Trans Audio Speech Lang Process 14(2)
8. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48:1162–1181
9. Craciu A, Paulus J, Sevkin G, Backstrom T (2017) Modeling formant dynamics in speech spectral envelopes. In: 25th European signal processing conference (EUSIPCO)
10. Dey S, Alam MA (2018) 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)
11. Hamzenejadi S, Yousef SA, Goki H. Extraction of speech pitch and formant frequencies using discrete wavelet transform. In: 2019 7th Iranian joint congress on fuzzy and intelligent systems (CFIS)