ISSN 1512-1232

Variability of formants for Vowels using Autoregressive Filter Model

J. T. Pramod[#1], G. N. Kodandaramaiah[*2], S.A.K. Jilani[#3]

[#]Madanapalle Institute of Technology & Science, Madanapalle, India
[#]Kuppam Engineering College, Kuppam, India
[1]jpramodvarma@yahoo.co.in, [2]kodandaramaiah@yahoo.com
[3]jilani_s_a_k@yahoo.com

Abstract

This paper presents Formant estimation for non-contextual vowels based on Autoregressive Filter Model for different recording conditions. Most of the speech specific information for vowels uttered by an individual speaker is contained in the lower Formants F_1 and F_2. It is also probable that the higher Formants F_3 and F_4 carry more speaker specific information. We propose to use frequency response model of vocal tract and Formants, using vowel utterances from south Indian speakers taken at different times, in different recording conditions. The condition used here is to record speech samples by consuming ice cold water and its effect at the very instant and after a time lapse of 5 minutes is checked. Formant values are obtained for the above 2 conditions and are compared to the normally recorded speech samples. The vowel utterances from 20 South Indian male speakers are recorded 40 times in different recording conditions and the variability of the resulting formant values are checked for different conditions on intra and inter speaker basis. We propose to look at these Formant values for individual speakers at different occurrences and conditions and investigate the role of this variability as a parameter on speech recognition process. Euclidean distance measure is used here to perform the speech recognition process. This process is implemented using MATLAB.

Keywords: *Formants, AR Model, Vowels & Ice water effect.*

I. INTRODUCTION

Automatic speech recognition (ASR) has made lot of progress with the development of digital signal processing hardware and software, using English and regional languages of choice. In this paper, a modified feature extraction based on frequency response of vocal tract and its resonant frequencies called Formants, using vowel utterances from south Indian speakers is presented. This technique calculates Formants from fixed frequency bands from the utterances of vowels /a/, /e/, /i/, /o/, and /u/. In addition 2 different recording conditions are added, first condition is to record samples right after consuming ice cold water and the second one is to record samples with a time lapse of 5 minutes after consuming ice cold water. The signal analysis involved here is spectral shaping of the recorded audio signal which is the process of converting the speech signal from sound pressure wave to a digital signal using a microphone. Feature extraction is next step where different speech parameters such as energy, pitch, formants and vocal tract area are extracted from the speech signal. Finally, using certain statistical modelling, conversion of parameters into multiple signal vectors is done [1]. Human speech has a complex hierarchical structure starting from sentences, which are divided into words. Words are in-turn built by phonemes (the basic voice elements for construction of voice). Vowels could be defined as phonemes with persistent frequency characteristics. The frequency characteristic represents a better parameter for speech and speaker recognition process [1][2]. This study of Formant frequency variations for different recording conditions serves a better clarity for automatic speech and speaker recognition in various applications.

II. IMPLEMENTATION

A. Autoregressive Model (AR)

A very powerful method for speech analysis is based on Linear Predictive Coding (LPC). It is a fast, simple and effective way of estimating the main parameters of speech signals such as energy, pitch, formants and vocal tract area. The spectral characteristics of speech signal are well defined using the Linear Predictive Coding. The drawback of the LPC is that, to minimize complexity in analysing signals, the speech signal is usually assumed to be the output of an all pole filter model. This means that the assumed model should not have zeros, even though the actual speech spectrum has zeros due to the glottal pulses as well as due to nasals and unvoiced sounds. This requirement of zeros can be accommodated by adding 2 or 4 poles to the existing number of poles in the filter transfer function. A filter model which has just poles is called an autoregressive model (AR), while a model which has just zeros is called a moving average model (MA), and a model which has both poles and zeros is called an autoregressive moving average model (ARMA) [1][2].

The simplest model for the vocal tract, consisting of linked cylindrical tubes, produces an all-pole transfer function. Note that the AR model is based on frequency domain analysis and need to be windowed. A hamming window is used for this purpose. A rectangular window also serves the purpose, but it causes abrupt changes in spectrum magnitude (high frequency noise) at the ends of the window. AR model is also used to determine the characteristics of the vocal system and from this system model evaluate the resonant frequencies of the vocal system. The order of the model is a function of the sampling frequency: fs/1000 + 2 [2]. In general, the z-transform B(z) of a filter's output b(n) is related to the z-transform A(z) of the input by

$$\mathbf{H(z)} = \frac{\mathrm{B(Z)}}{\mathrm{A(Z)}} = \frac{b(1)+b(2)z^{-1}+\cdots+b(mb+1)z^{-m}}{a(1)+a(2)z^{-1}+\cdots+a(na+1)z^{-n}}$$

where H(z) is the filter's transfer function. Here, the constants b(i) and a(i) are the filter coefficients and the order of the filter is the maximum of n and m [1][3].

B. Database Collection

Participants chosen spoke standard Indian English vowels without distinct accents, special speech habits, were between 20 and 21 years of age and did not suffer from any speech hearing disorder. Each speaker was asked to record the required speech as naturally as possible, and their speech recorded individually, in a speech laboratory, with a portable digital recorder via a small collar microphone, the distance between the Microphone and mouth of speakers was approximately 10 cm, and samples are acquired with a sampling rate of 11.025 KHz per second. The speakers were asked to record the samples in 3 different ways. The first one is normal recording, the second recording is done at the very instant of consuming ice cold water and the third recording is done after a time lapse of 5 minutes after consuming the ice cold water.

C. Formant Estimation Algorithm

Figure below depicts the formant estimation process flow for formant estimation. From the speech production model, it is known that the speech undergoes a spectral tilt of –20dB/decade. To counter this, pre-emphasis filter is used to boost the higher frequencies and flatten the spectrum.

The next step involves windowing of the boosted speech signal using the Hamming window followed by calculating the AR Model and Vocal Tract transfer function. From the frequency spectrum of the Vocal Tract transfer function, the formants are extracted using by peak picking [2][3].

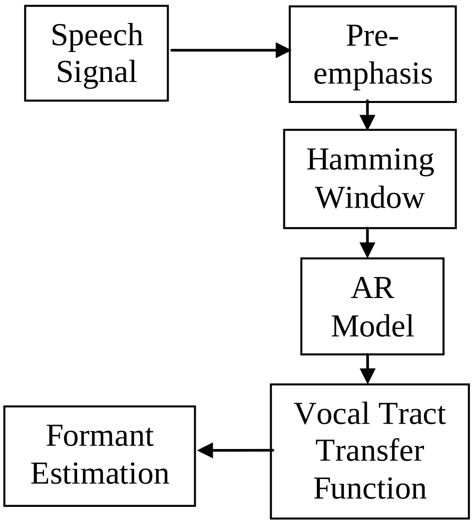

Fig.1 Block diagram for Formant Estimation

The database of 40 sets of samples for different speakers is obtained. This database contains the first four formants of each sample for 20 speakers which are used as references for the speech recognition process. In addition to the normal recording conditions, the recordings are done with ice cold water consumed. Comparing these samples with the reference samples in the database can give the information about how the formant spread varies for different recording conditions. Euclidean distance measure is used to perform the recognition process.

D. Results

Using LPC, the formant spread variations of the subject was found. Study of variability of the above formant spread among 20 different subjects is highlighted to get Intra Speaker Variability. This variability can be used as a cue for personal identification and voice print signature, as well as Vocal Tract Signature of an individual [4][5]. We found the Time averages of the worst and the best patterns of 20 subjects and plotted the resultant worst pattern and resultant best pattern for a subject for the phoneme ‘a’. Further we repeated the above steps for the remaining phonemes ‘e’, ‘i’, ‘o’, & ‘u’.

The variation in formants is identified in the different recording conditions. It is seen that the sample recorded instantly after drinking ice cold water shows much variation in the formants compared to the normal one. We can see that the formants are shifted towards the left i.e., the formant frequencies are reduced when the ice cold water is consumed. The medical reason for this effect is that once ice cold water is consumed, the Vocal Tract muscles tend to get stiff restricting free movement of the muscles. This results in difficulty in pronunciation of the vowels in this context. The stiffness in the muscles results in generation of low frequency resonances as there will be a variation in the shape of the Vocal Tract.

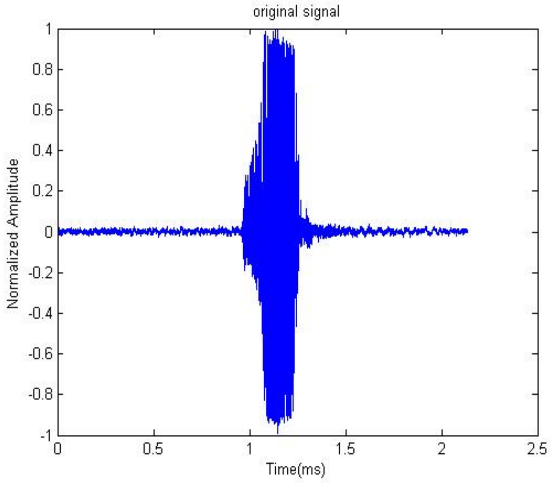

Fig.2 Recorded speech sample for vowel a

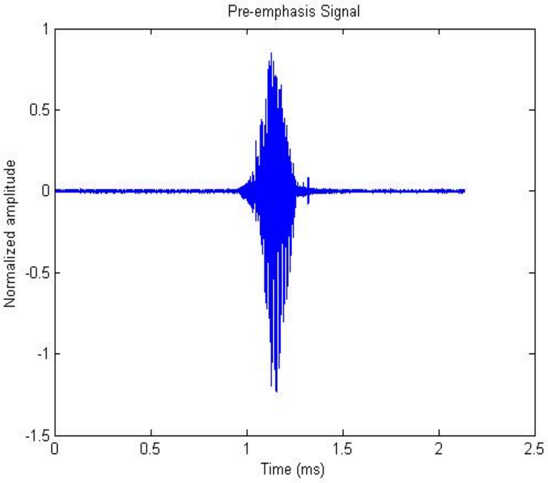

Fig.3 Pre-emphasized signal for vowel ‘a’

Figures 1 to 5 show the original signal, pre-emphasized signal, frequency domain representations of the signal obtained for vowel ‘a’. The figures 6 to 10 show the comparison of Formant spread for the different recording conditions for vowels ‘a’, ’e’, ’i’, ’o’, ‘u’ respectively.

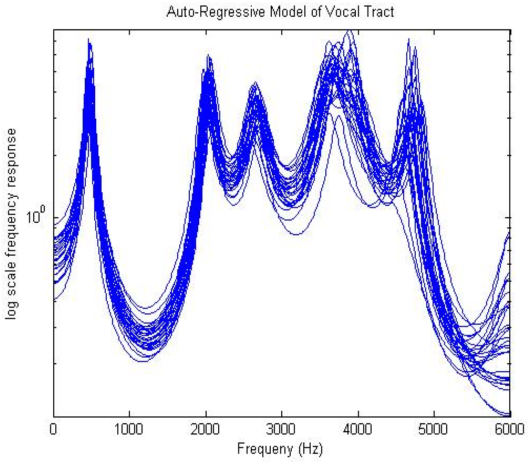

Fig.4 Frequency domain representation of the vocal tract shape for 40 samples taken together for vowel ‘a’

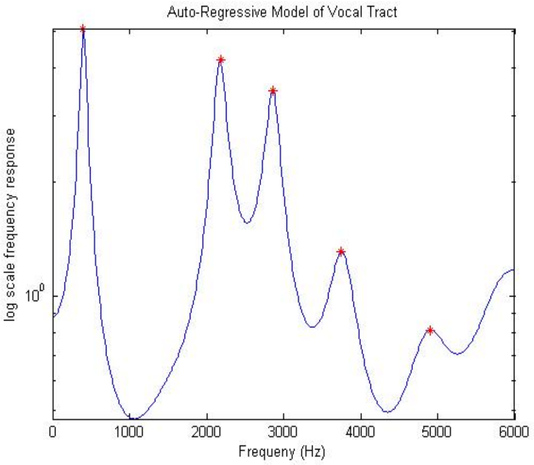

Fig.5 Peak picking to get Formant center frequencies for vowel ‘a

The different recording conditions are represented in three different colours as shown in figure 6; the red coloured signal shows the normal recording, and the blue coloured signal shows the sample recorded at the very instant of drinking ice cold water and the black coloured signal shows the sample recorded after a time lapse of 5 minutes after drinking ice cold water. We can see that for the 2nd case, the formant frequencies are shifted to low values because of the muscle stiffening and the reasons already mentioned above. This shift is significant at the higher formant frequencies. This suggests that recognition becomes difficult at the higher formant frequencies as there are larger variations in the formant center frequencies. The 3rd case where the sample is recorded after 5 minutes after drinking ice cold water shows that the effect of ice cold water on the Vocal Tract muscles gets decreased due to time lapse. The Formant center frequencies for this case comes closer to the normal recording condition and the effect further gets reduced after further time lapse. Using the Euclidean distance measure, the recognition of vowels is carried out for the 3 different conditions. It is observed that the recognition of vowels is not precise due to the shift in Formant frequencies with the ice water consumed.

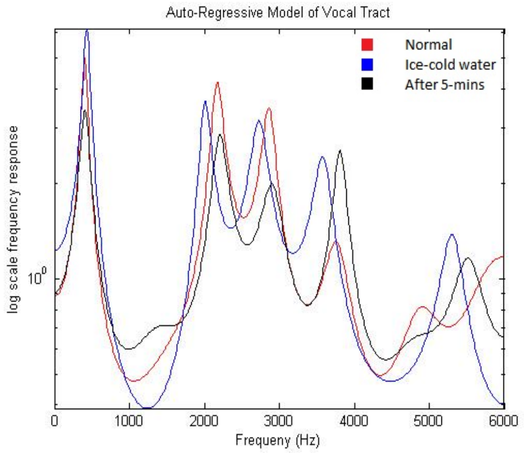

Fig.6 Comparison of the Formant spread for the 3 recording conditions for vowel ‘a’

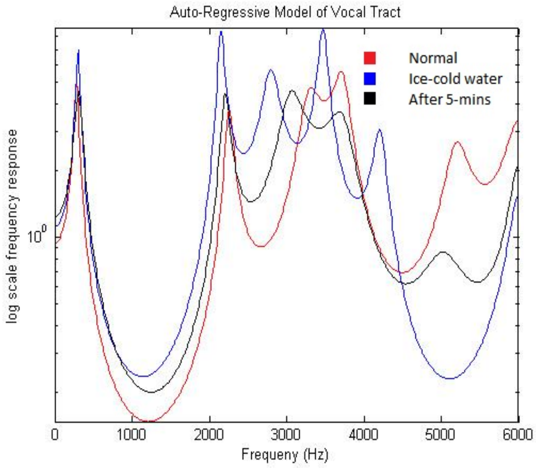

Fig.7 Comparison of the Formant spread for the 3 recording conditions for vowel ‘e

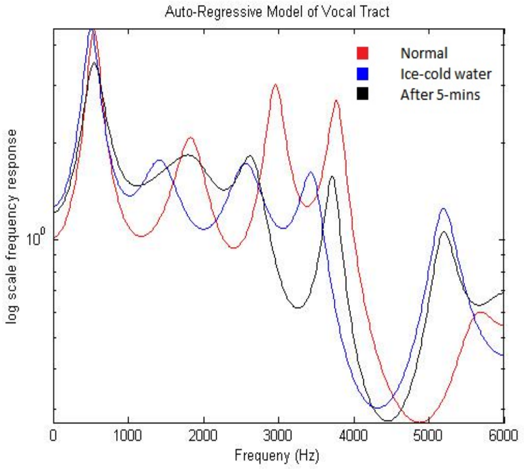

Fig.8 Comparison of the Formant spread for the 3 recording conditions for vowel ‘i’

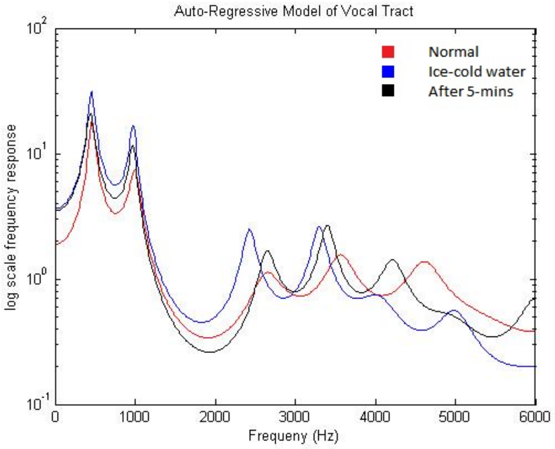

Fig.9 Comparison of the Formant spread for the 3 recording conditions for vowel ‘o’

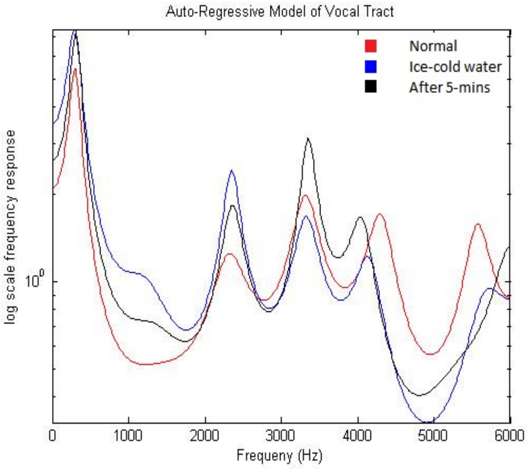

Fig.10 Comparison of the Formant spread for the 3 recording conditions for vowel ‘u’

ISSN 1512-1232

III. RECOGNITION

The tables 1, 2 & 3 give the results for vowel recognition based on Euclidean Distance method for the 3 different recording conditions i.e., normal condition, ice-cold water condition and after 5-minutes condition. Vowels /a/ and /o/ has achieved perfect classification compared to other vowels for the normal and ice-cold water conditions and vowel /e/ has achieved perfect classification for normal condition.

Vowels	**Predicted**					
Actual	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**% correct**
/a/	40	0	0	0	0	100
/e/	0	40	0	0	0	100
/i/	2	1	37	0	0	92.5
/o/	0	0	0	40	0	100
/u/	2	5	1	0	32	80

Table.1 Percentage of vowel recognition (normal)

Vowels	**Predicted**					
Actual	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**% correct**
/a/	40	0	0	0	0	100
/e/	1	38	0	0	1	95
/i/	8	2	27	0	3	67.5
/o/	0	0	0	40	0	100
/u/	1	12	0	0	27	67.5

Table.2 Percentage of vowel recognition (ice-cold water)

The detection rate for vowels /i/ and /u/ are better for normal and after 5-mins conditions and not good for the ice-cold water condition. This is due to the fact that, recognition is affected either in positive or negative manner for different vowels due to the ice-cold water consumption. The Fig.11 shows comparison of vowel percentage vowel recognition for the 3 conditions for 40 samples.

Vowels	**Predicted**					
Actual	**/a/**	**/e/**	**/i/**	**/o/**	**/u/**	**% correct**
/a/	38	0	2	0	0	95
/e/	2	38	0	0	0	95
/i/	6	1	32	0	1	80
/o/	2	0	0	38	0	95
/u/	5	3	0	0	32	80

Table.3 Percentage of vowel recognition (after 5-minutes)

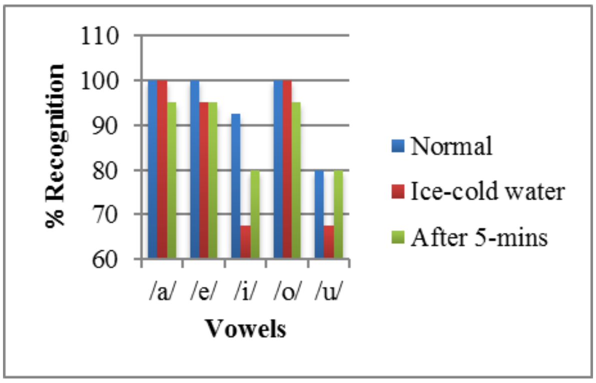

Fig.11 Vowel vs % vowel recognition for three recording conditions

IV. CONCLUSION AND FUTURE SCOPE

The Formant spread varies with the 3 different recording conditions. It is observed that when ice cold water is consumed, the effect felt is that the Formant center frequency values get reduced i.e., the formants shift towards the left side in the frequency band. This is because of the reasons mentioned in the above section ***II.C*** As a result, the speaker or speech recognition becomes little difficult due to this formant shift. The same effect is felt for the samples recorded for various geographical or cultural populations. This work can be extended by considering additional recording conditions for example the effect of inhaling of helium and consuming spicy food, as these conditions also result in variation of the Vocal Tract configuration and ultimately the Formant frequencies. In addition, other speech parameters such as Formant Bandwidths, Pitch and Energy variations can be checked for the different recording conditions.

REFERENCES

[1] Paulraj M.P. Sazali Yaacob, Shahrul Azmi M.Y “vowel classification based on frequency response of vocal tract,” proc. International conference on communication engineering, 2008.

[2] Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Pearson Education, 1993.

[3] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, Englewood Cliffs, NJ, 1978.

[4] Prof. G.N. Kodanda Ramaiah, Dr. M.N. Giri Prasad, Prof. R.B. Kulkarni, Dr. Mukunda Rao “Intra-Speaker Tract Shape Variability Estimation for Indian Males for Non-Contextual Vowel /a/ using LPC coding”, international conference on systemic, cybernetics and informatics.

[5] J. Makhoul, “Linear Prediction: A Tutorial Review,” Proc. IEEE, Vol. 63, pp 561-580, 1975.

Article received: 2015-11-26