Chapter 4

Formant Estimation

4.1 Formant Frequencies

The vocal tract is an acoustic transmission system characterized by its natural frequencies, called formants, which correspond to resonances in its frequency response[16]. In general, the formant frequencies for females and children are higher than those of men[19]. For each sound, formant frequencies of different speakers are different, but they lie close to each other. This behavior of formant frequencies can be used for recognition of sounds. Basically, the values of first and second formant frequencies can be used for vowel recognition.

4.2 Formant Estimation

Short-time autoregressive(AR) modeling is used to estimate the parameters of acoustic tube models of the vocal tract based on a linear system framework. The autoregressive model is one of a group of linear prediction formulas that attempt to predict an output y[n] of a system based on the previous outputs (y[n-1], y[n-2], ...) and inputs (x[n], x[n-1], x[n-2], ...). The estimated value of y[n] can be expressed as,

$$y_e[n] = a_1*y[n-1] + a_2*y[n-2]... + b_0*x[n] + b_1*x[n-1] + ...$$

Deriving the linear prediction model involves determining the coefficients a_1, a_2, a_3 ... and b_1, b_2, b_3 ... in the above equation[20]. A model which depends only on the previous outputs of the system is an autoregressive model, where (b_1, b_2, b_3 ...) are all zeros. The AR model has only poles. A model which depends only on the previous inputs of the system is called a moving average model (MA), where (a_1, a_2, a_3 ...) are all zeros. The MA model has only zeros. A model based on both inputs and outputs is an autoregressive-moving-average model (ARMA).

The human vocal system can be modeled as an all-pole system when producing vowel sounds. The formants or the resonant frequencies of the vocal system are the peaks in the frequency response or the above transfer function. The pole locations define the resonance frequencies of the vocal tract. The parameters of the vocal tract model may be estimated using linear predictive analysis. AR modeling yields the all-pole spectrum of the speech waveform, from which one may obtain the frequencies of the individual resonances. After normalizing the sound signal an AR model of the voice is created. Functions available in Matlab are used to estimate formant values. Function 'ar' gives the auroregression model of voice with the required coefficients. Command 'th2tf' is used to define the transfer function of the model. Frequency response of the model can be obtained using 'freqz'. Matlab code presented in *Appendix(G)* shows the estimation of first and second formant frequencies.The AR model is based on frequency-domain analysis and should be windowed. The hamming window is used.
One conjugate pole pair is required to produce each formant and one formant is expected in each 1kHz band or so. Therefore the order of the model(n) is a function of the sampling frequency which is given by

$$n = \text{round}(fs/1000) + 2.$$

The added 2 is the 'empirically' determined adjustment factor[20]. Sampling frequency fs is taken as 11025Hz.

The autoregressive models of vocal tract for each vowel considered in the study are displayed in Figures 4.1-4.5.

/ a: / sound

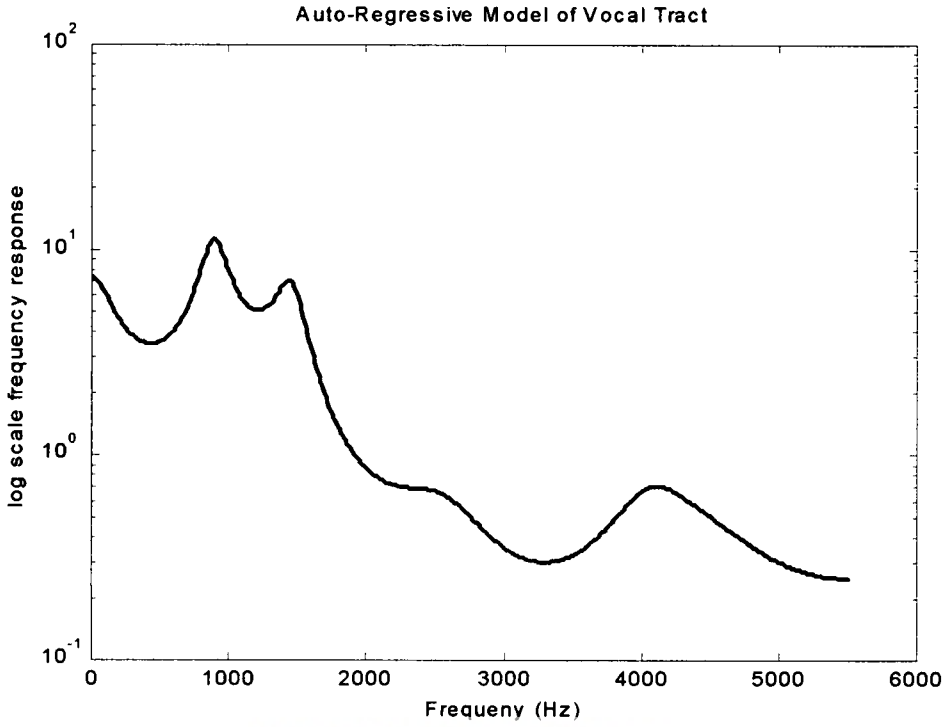

Figure 4.1 – Auto-regressive model of vocal tract for /a:/ sound

/ æ / sound

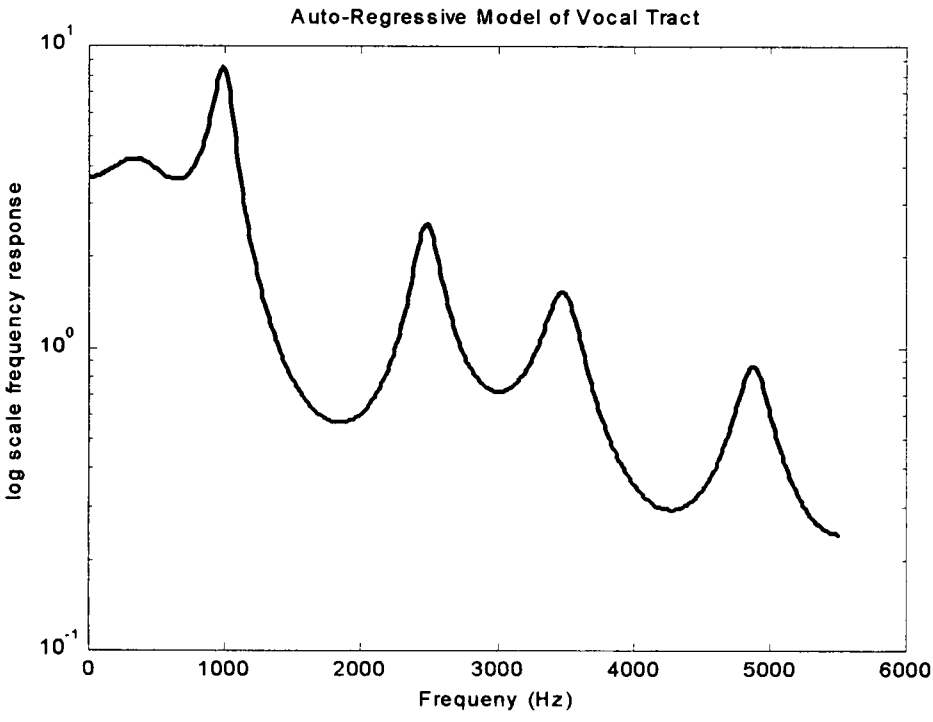

Figure 4.2 – Auto-regressive model of vocal tract for / æ / sound

/ **u:** / sound

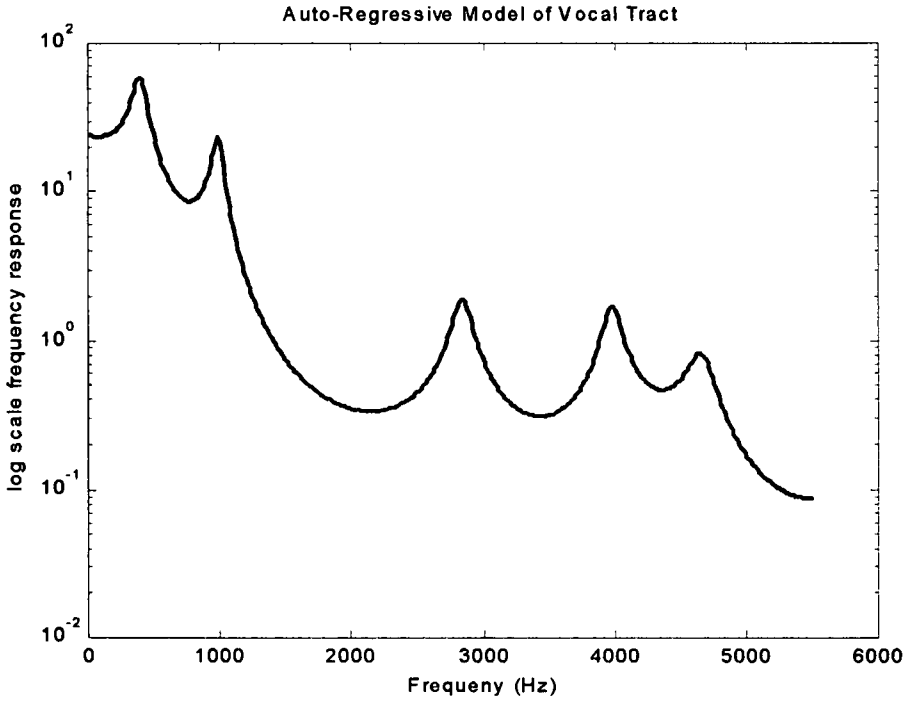

Figure 4.3 – Auto-regressive model of vocal tract for /u:/ sound

/ **o** / sound

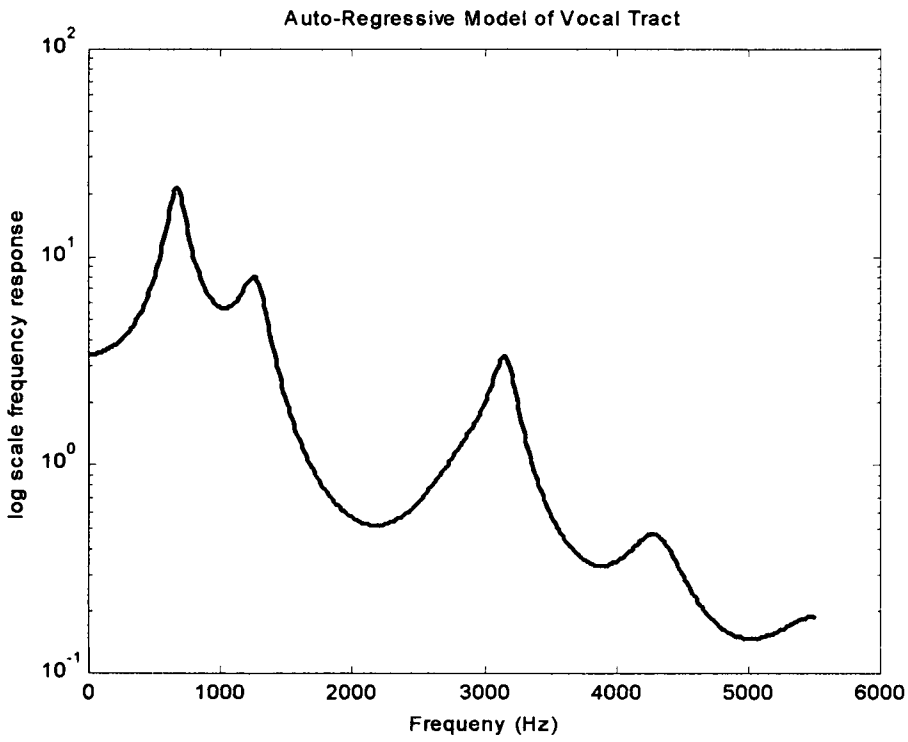

Figure 4.4 – Auto-regressive model of vocal tract for /o/ sound

/ eI / sound

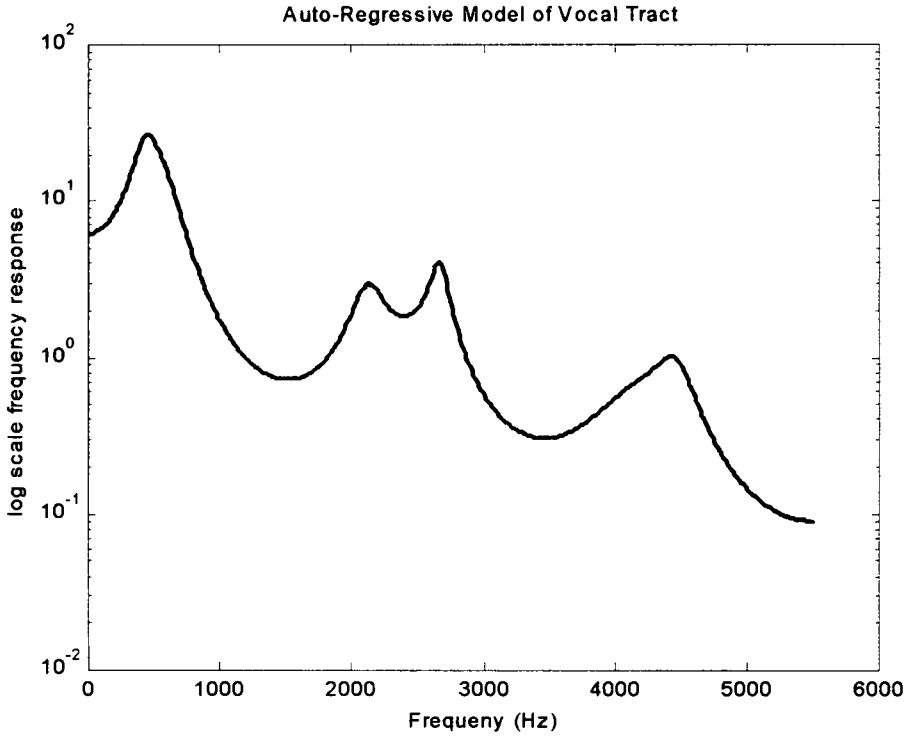

Figure 4.5 – Auto-regressive model of vocal tract for /eI/ sound

4.3 Vowel Recognition Using Formants

Having estimated the first and second formant frequencies, the next task is to derive the vowel. The second formant frequency is displayed along the X-axis, and the first formant frequency is along the Y-axis. The result is displayed as a 'vowel triangle'. Vowels are described by the following factors[5].

1. Tongue high or low
2. Tongue front or back
3. Lips rounded or unrounded
4. Nasalized or unnasalized

As the first formant frequency increases, the 'position' of the vowel in the mouth goes from low to high. As the second formant frequency increases, the 'position' of the vowel moves from the back of the mouth to the front[19]. High or low and front or back refer roughly to the position of the highest part of the tongue. Front is toward the lips and back is toward the pharynx. In nasalized vowels, the velum is open, so that sound passes through the nasal cavity as well as through the mouth. In unnasalized vowels, the velum is shut and the sound passes through the mouth only[5].

For each vowel, the values of formants for about 40 normal speaker samples are found and the locations are shown in Figures 4.6 - 4.10.

/ a: / sound

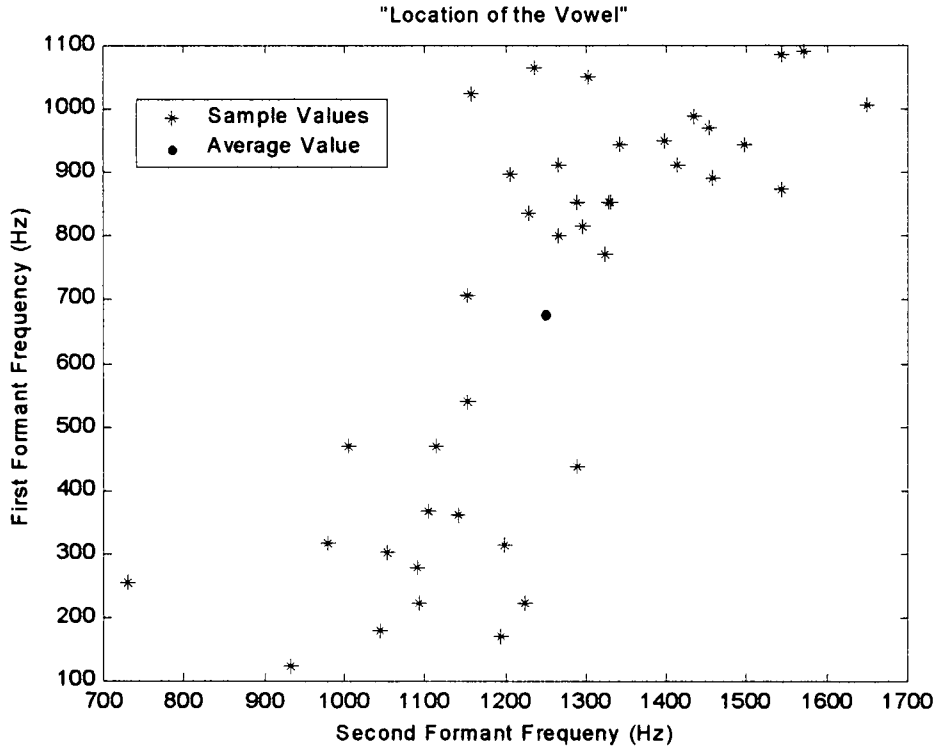

Figure 4.6 – Location of Formants for /a:/ sound

/ æ / sound

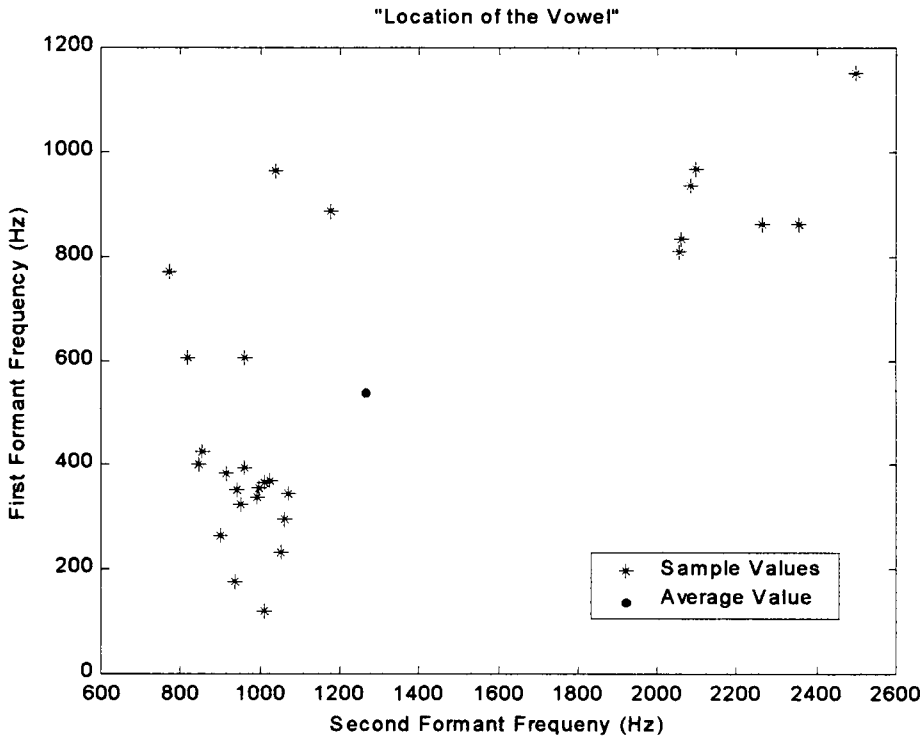

Figure 4.7 – Location of Formants for / æ / sound

/ u: / sound

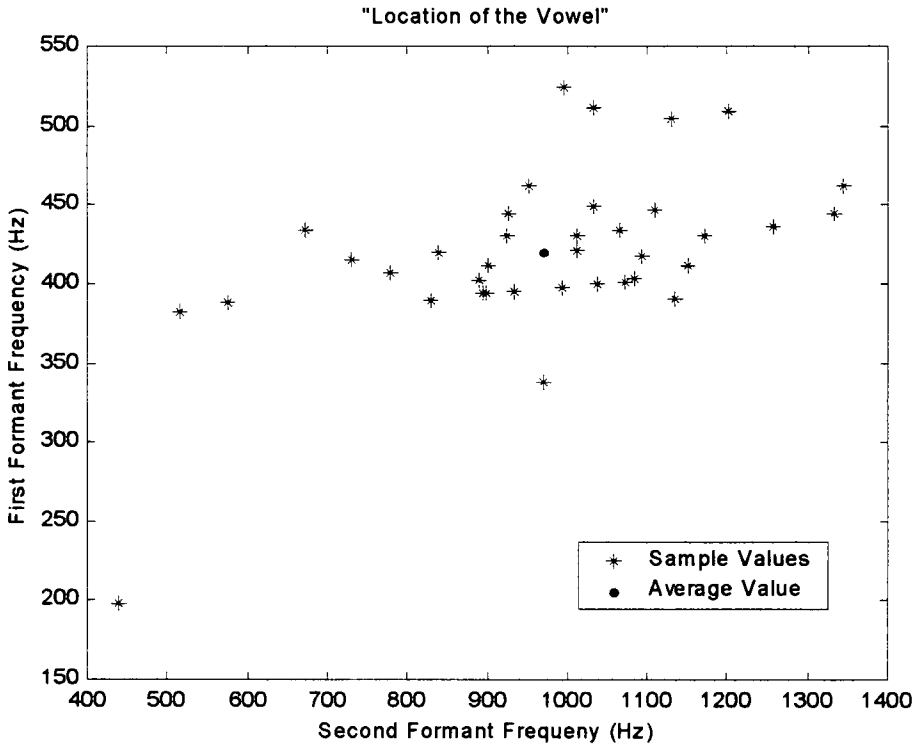

Figure 4.8 – Location of Formants for /u:/ sound

/ o / sound

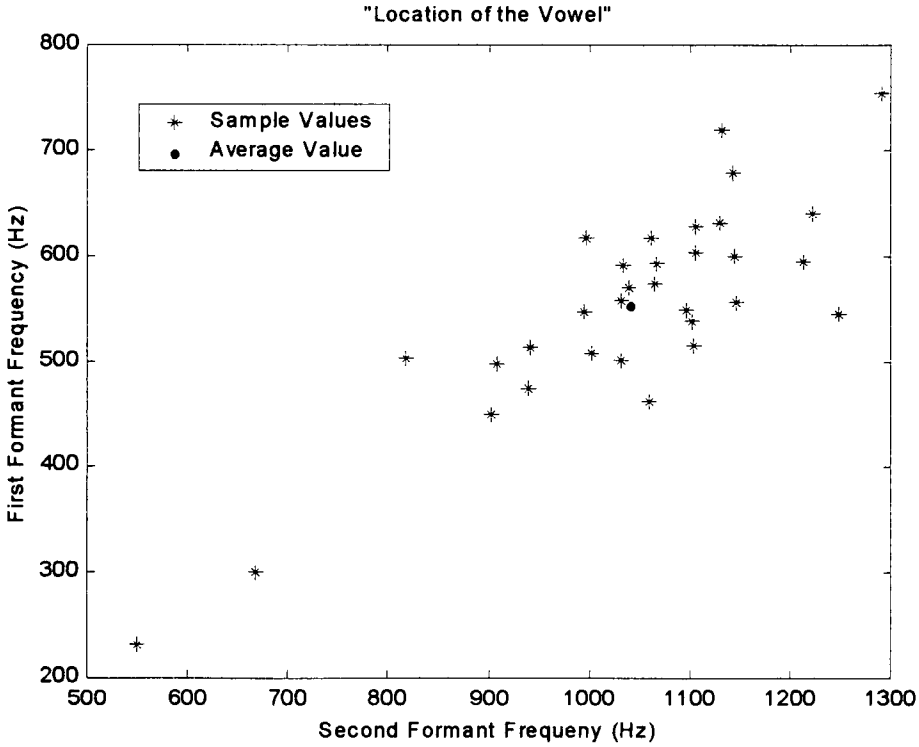

Figure 4.9 – Location of Formants for /o/ sound

/ eI / sound

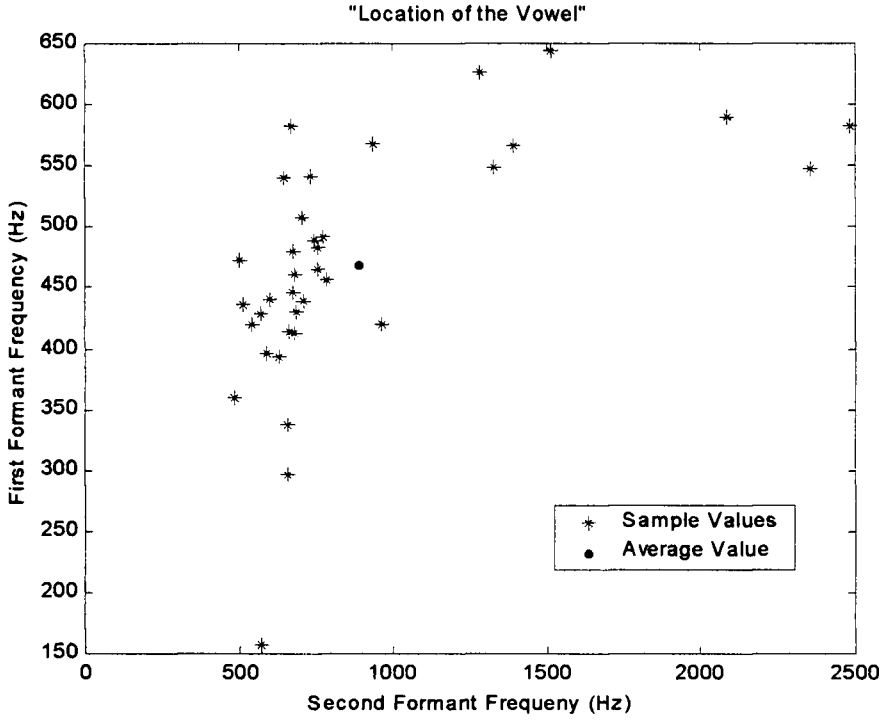

Figure 4.10 – Location of Formants for /eI/ sound

4.4 Average formant values

Average formant values for each vowel are also found to check whether they are close to standard formant values. The standard values are found to be not unique. The values given by different authors [4], [5], [19] and estimated average values in the study are compared in Table 4.1.

Vowel	First and Second Formant Frequencies for Men			
	[4]	[5]	[19]	Estimated Values
/a:/	640, 1190	-----	-----	678, 1251
/æ/	660, 1720	690, 1660	735, 1625	529, 1281
/u:/	300, 870	310, 870	290, 940	420, 971
/o/	570, 840	590, 880	610, 865	552, 1041
/eI/	-----	-----	-----	469, 889

Table 4.1 – Comparison of Formant Frequencies

The above results show that the average values do not give a good estimation of formant frequencies.

4.5 Specific regions of vowels

As the next step, possible area for each vowel is found using the graphs in Figures 4.6-4.10. Results for normal and hearing impaired speakers are calculated by changing the area to maximize the correct number of decisions. The number of correct decisions is

defined as the following equation and the results for each vowel are shown in Tables 4.2 – 4.6.

$$\text{No. of correct decisions} = \begin{cases} \text{No. of samples; Pronunciation good \&} \\ \text{Locate within the selected area} \\ + \\ \text{No. of samples; Pronunciation bad \&} \\ \text{Locate outside the selected area} \end{cases}$$

First and Second formant frequencies are represented by f1 and f2.

/ **a:** / sound

Area	Percentage of correct decisions	
	Normal Speakers	Hearing Impaired Speakers
124<f1<1091, 729<f2<1651	100	37.78
150<f1<1080, 950<f2<1550	87.50	37.78
150<f1<550, 950<f2<1250 750<f1<1100, 1150<f2<1550	85.00	60.00
150<f1<400, 900<f2<1300 700<f1<1100, 1100<f2<1600	82.50	71.11
120<f1<475, 725<f2<1400 708<f1<1091, 1100<f2<1651	**97.50**	**75.56**

Table 4.2 – Possible area for /a:/ sound with % of correct decisions

/ **æ** / sound

Area	Percentage of correct decisions	
	Normal Speakers	Hearing Impaired Speakers
120<f1<1153, 816<f2<2498	100	15.91
100<f1<1000, 800<f2<1200 800<f1<1200, 2000<f2<2500	96.55	43.18
100<f1<610, 800<f2<1100 800<f1<1000, 2000<f2<2400	86.21	61.36
100<f1<450, 800<f2<1100 800<f1<1200, 2000<f2<2500	82.76	79.55
106<f1<427, 816<f2<1072 809<f1<1180, 1038<f2<2498	**89.66**	**81.82**

Table 4.3 – Possible area for / æ / sound with % of correct decisions

/ u: / sound

Area	Percentage of correct decisions	
	Normal Speakers	Hearing Impaired Speakers
198<f1<524, 438<f2<1345	100	62.79
335<f1<524, 516<f2<1345	97.37	67.44
195<f1<435, 435<f2<900 335<f1<525, 900<f2<1345	100	69.77
381<f1<435, 516<f2<900 337<f1<524, 900<f2<1345	**97.37**	**76.44**

Table 4.4 – Possible area for /u:/ sound with % of correct decisions

/ o: / sound

Area	Percentage of correct decisions	
	Normal Speakers	Hearing Impaired Speakers
231<f1<755, 549<f2<1291	100	44.19
450<f1<755, 817<f2<1291	93.75	48.84
300<f1<755, 608<f2<1291	96.88	37.21
461<f1<755, 667<f2<1291	90.63	51.16
474<f1<755, 804<f2<1291	87.50	60.47
474<f1<755, 804<f2<1291 225<f1<350, 549<f2<668	**93.75**	**67.44**

Table 4.5 – Possible area for /o/ sound with % of correct decisions

/ **eI** / sound

Area	Percentage of correct decisions	
	Normal Speakers	Hearing Impaired Speakers
157<f1<645, 486<f2<2485	100	34.88
296<f1<645, 504<f2<2485	97.22	37.21
296<f1<600, 504<f2<1000 530<f1<645, 1200<f2<2485	97.22	62.79
296<f1<585, 504<f2<787 548<f1<645, 1279<f2<2485	88.89	76.74
296<f1<549, 504<f2<967 547<f1<645, 1279<f2<2485	88.89	67.44
296<f1<585, 486<f2<784 547<f1<645, 1279<f2<2485	**91.67**	**76.74**

Table 4.6 – Possible area for /eI/ sound with % of correct decisions

The identification of correct pronunciation was done by optimizing the possible area for each vowel. A sound falling within the identified area of the formants plot was taken as a correct utterance. Those identified areas are shown in Figures 4.11-4.15.

/ **a:** / sound

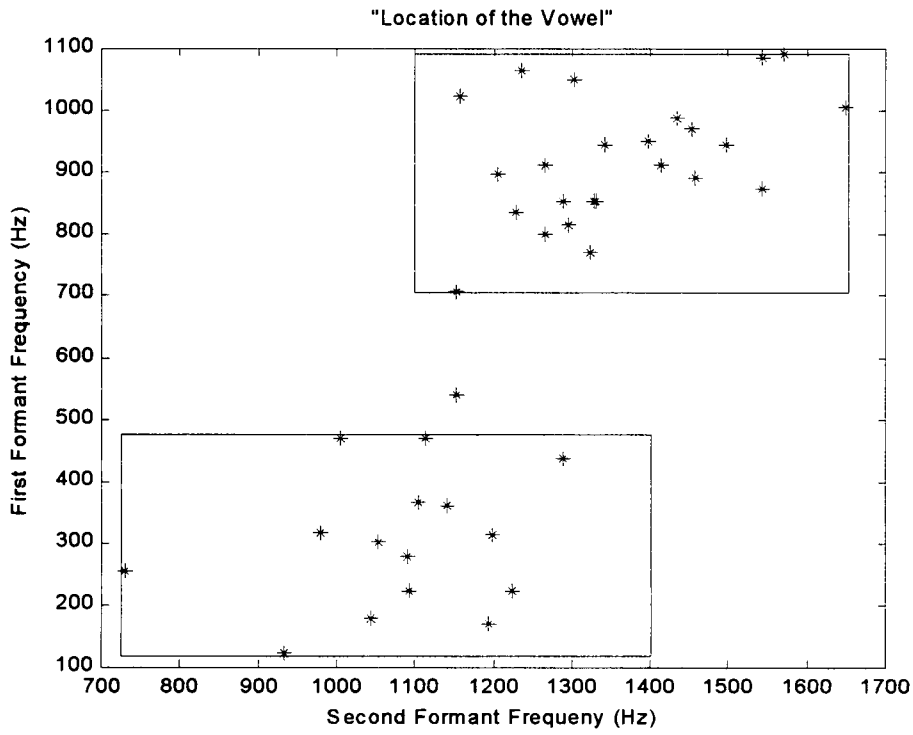

Figure 4.11 – Areas for the Location of /a:/ sound

/ æ / sound

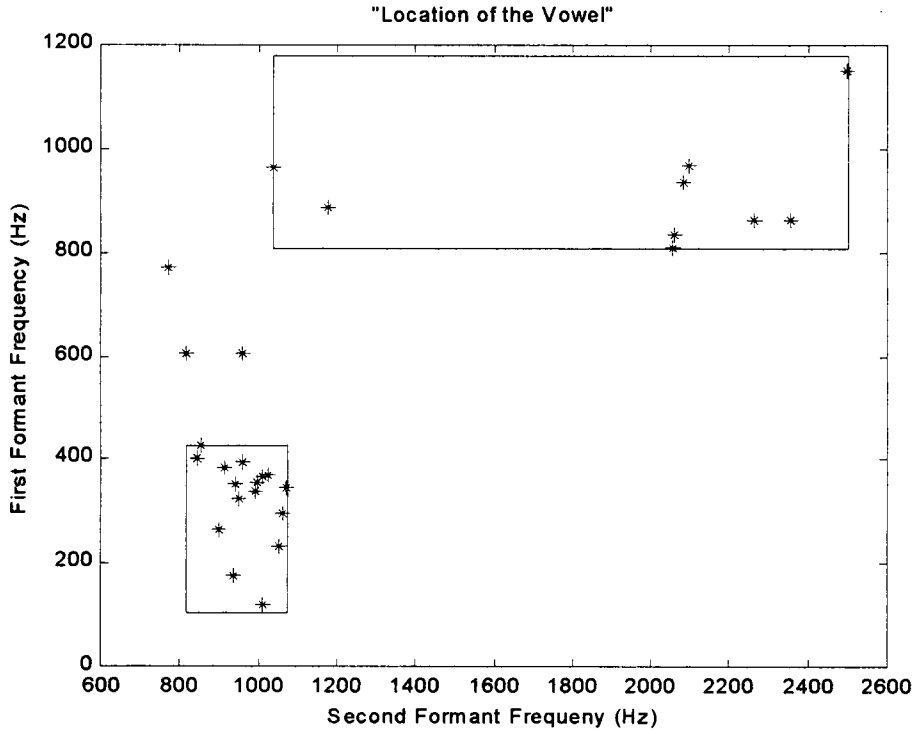

Figure 4.12 – Areas for the Location of /æ/ sound

/ u:/ sound

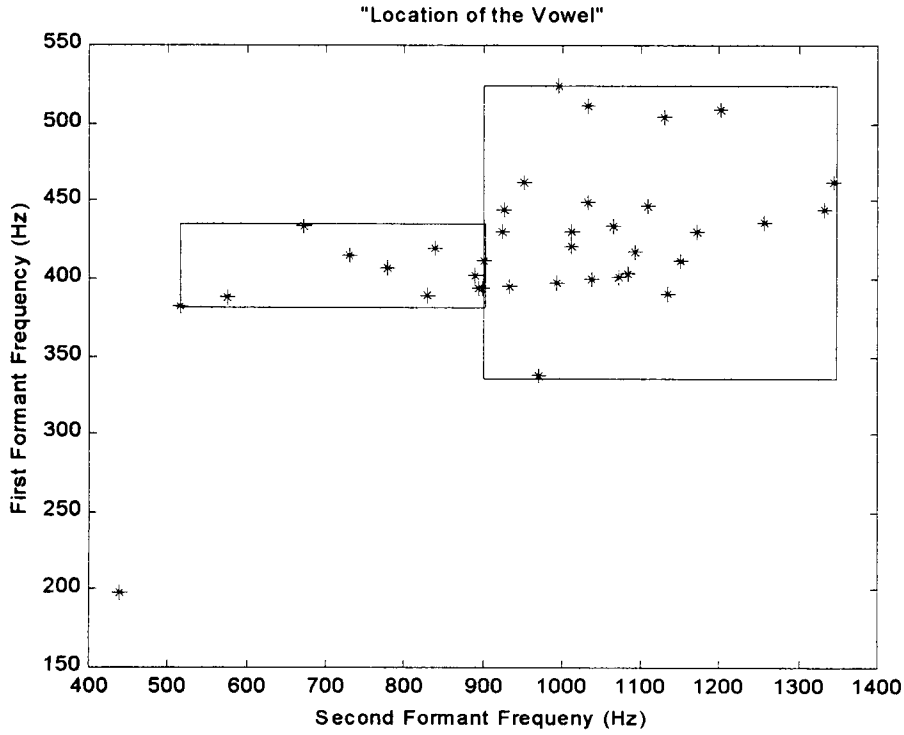

Figure 4.13 – Areas for the Location of /u:/ sound

/ o / sound

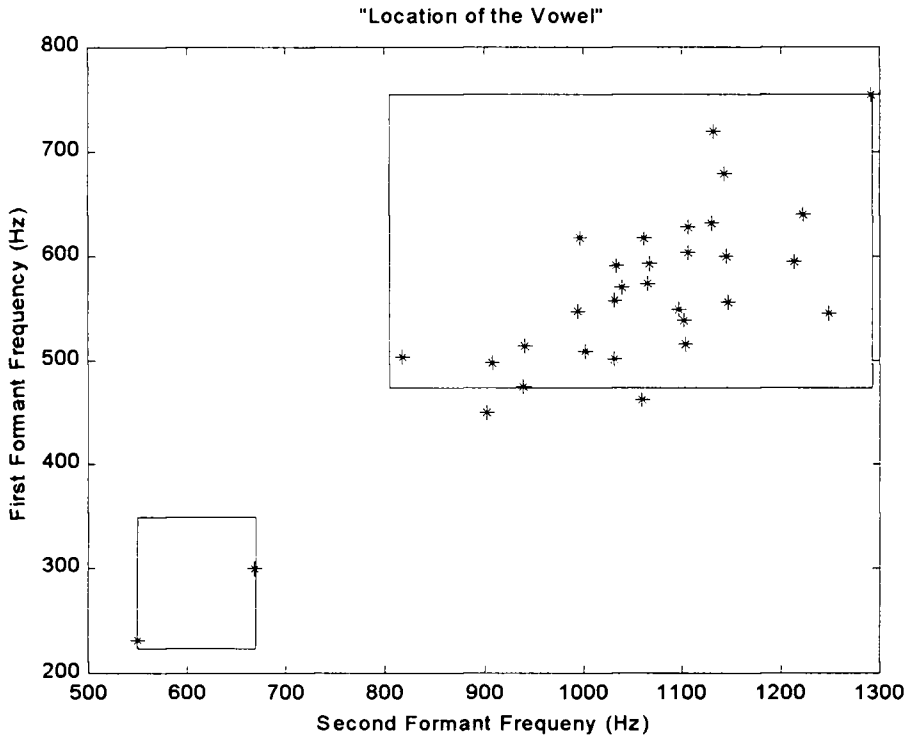

Figure 4.14 – Areas for the Location of /o/ sound

/ eI / sound

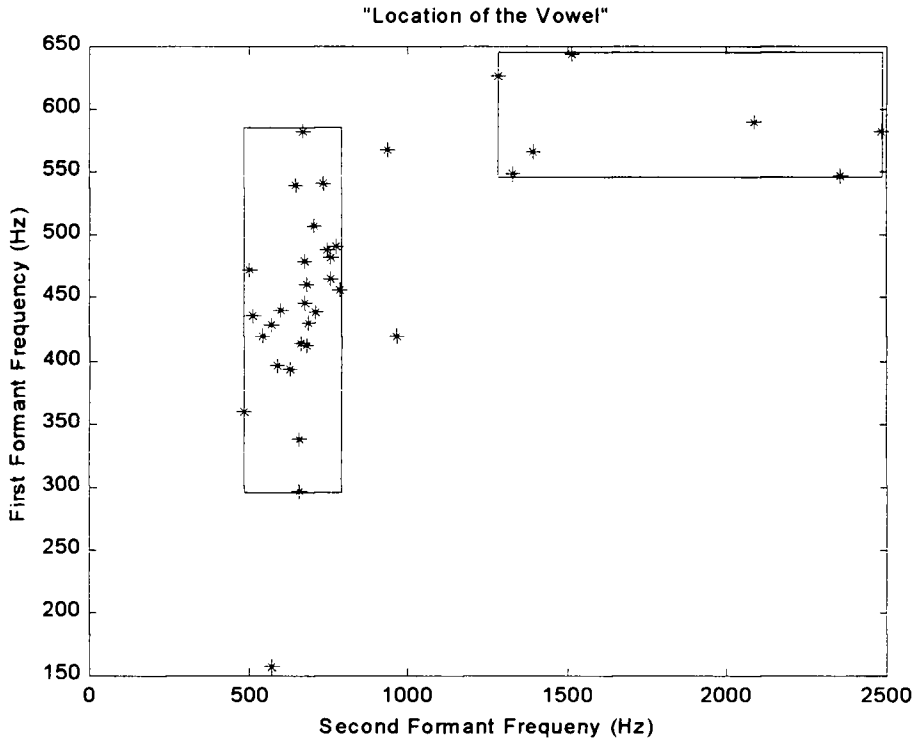

Figure 4.15 – Areas for the Location of /eI/ sound

Using formant analysis method success rates of 90 – 98% were achieved for the normal speakers, while the corresponding rates from 67 – 82% were obtained for hearing-impaired speakers as shown in Table 4.7.

Vowel	Area	Percentage of correct decisions	
		Normal Speakers	Hearing Impaired Speakers
/a:/	120<f1<475, 725<f2<1400 708<f1<1091, 1100<f2<1651	**97.50**	**75.56**
/æ/	106<f1<427, 816<f2<1072 809<f1<1180, 1038<f2<2498	**89.66**	**81.82**
/u:/	381<f1<435, 516<f2<900 337<f1<524, 900<f2<1345	**97.37**	**76.44**
/o/	474<f1<755, 804<f2<1291 225<f1<350, 549<f2<668	**93.75**	**67.44**
/eI/	296<f1<585, 486<f2<784 547<f1<645, 1279<f2<2485	**91.67**	**76.74**

Table 4.7 – Summary of Results for % of correct decisions