Speech Communication 2 (1983) 295–303
North-Holland

A STUDY OF TWO-FORMANT MODELS FOR VOWEL IDENTIFICATION

K.K. PALIWAL

Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India

W.A. AINSWORTH and D. LINDSAY

Department of Communication and Neuroscience, University of Keele, Keele, Staffs., ST5 5BG, England.

Received 7 December 1982
Revised 14 June 1983

Abstract. An experiment has been performed where various two-formant models reported in the literature were assessed as to their ability to predict the formant frequencies obtained in a vowel identification task. An alternative model is proposed in which the auditory processing of vowel sounds is assumed to take place in two stages: a peripheral processing stage and a central processing stage. In the peripheral stage the speech spectrum is transformed to its auditory equivalent and the formant frequencies are extracted from this spectrum using a peak-picking mechanism. The central stage performs a two-formant approximation on the results of the first stage operation, and it is this formant pair that vowel identification is taken to operate on during vowel perception. The first and second formant frequencies of this two-formant model are taken to be equal to the first and second formant frequencies extracted at the first stage plus a perturbation term which accounts for the interaction effects of the neighbouring formants. The perturbation caused by each of these neighbouring formants is inversely proportional to its separation from the main formants. This model compares favourably with previous models in its prediction of the formant frequencies obtained from the vowel identification task.

Zusammenfassung. In einem Experiment geht es zunächst um die Leistung verschiedener aus der Literatur bekannter Zwei-Formanten-Modelle, in einem Identifikationstest mit Vokalen erhaltene Formantfrequenzen voraussagen zu können. Danach wird ein zweistufiges Alternativmodell der auditiven Verarbeitung von Vokalen mit peripherer und zentraler Stufe entwickelt. Auf der peripheren Stufe, wo das Spektrum des Sprachschalls in seinen entsprechenden Gehörseindruck überführt wird, erfolgt die Formantfrequenzextraktion aus den Spektren nach dem Prinzip des "peak-picking". Die ERgebnisse dieser Operation liefern der zentralen Stufe die Grundlage einer Zwei-Formanten-Approximation. Auf dieses Formantenpaar dürfte sich die Vokalidentifikation während der Vokalperzeption stützen. Die Frequenzen des 1. und 2. Formanten unseres Zwei-Formanten-Modells ergeben sich aus den entsprechenden Formantfrequenzen der ersten Stufe unter zusätzlicher Berücksichtingung eines Korrekturausdrucks, der den Einfluss der Nachbarformanten einfängt. Der Einfluss jedes der Nachbarformanten ist seinem Abstand zu den Hauptformanten umgekehrt proportional. Gegenüber früheren Vorschlägen hat unser Modell den Vorzug, dass sich mit ihm die Formantfrequenzen aus dem Identifikationsexperiment besser voraussagen lassen.

Résumé. Une expérience a été réalisée dans laquelle divers modèles à deux formants, proposés dans la littérature, sont évalués en fonction de leur capacité de prédire les fréquences formantiques obtenues lors d'une tâche d'identification de voyelles. Un modèle concurrent est proposé dans lequel le traitement auditif des voyelles est supposé intervenir en deux étapes. A l'étape périphérique, le spectre de parole est transformé en son correspondant auditif et les fréquences formantiques sont extraites par une procédure de détection de pics. L'étape centrale effectue une approximation à deux formants sur la sortie fournie par le traitement précédent; l'identification de la voyelle s'appuie sur la paire de formants ainsi extraite. Les premières et secondes fréquences formantiques de ce modèle sont obtenues en additionnant un terme correctif aux fréquences des deux premiers formants extraits lors de la première phase. Ce terme correctif est inversément proportionnel à la distance séparant les formants principaux des formants voisins. Confronté aux modèles antérieurs, le nôtre se comporte avantage usement quant à la prédiction des fréquences formantiques obtenues dans une tâche d'identification de voyelles.

Keywords. Vowel identification, two-formant vowel models, auditory processing.

0167-6393/83/$3.00

1. Introduction

It has often been suggested that the frequencies of the first two formants are the most important factors in the recognition of vowel sounds [1–3]. Moreover, some perceptual studies have been reported in the literature [4–7] which suggest that a two-formant model of a vowel is a valid representation at some level of auditory processing. These studies have proposed different methods of constructing the two-formant model from the speech spectrum. Plomp [4], for example, computed the first two principal components (suggesting an equivalence to the first two formants) from the entire speech spectrum measured in one-third octave bands. Karnickaya et al. [5] converted the speech spectrum into an auditory equivalent by a transformation which involved critical-band filtering and a spectral sharpening process; the values of the model's two formants were obtained by picking the two largest peaks from the equivalent auditory spectrum. Carlson et al. [6] reported an empirical relation using a spectral prominence model to compute the second formant, $F2'$, of the two-formant model from the frequencies, $F1$, $F2$, $F3$ and $F4$, of the first four formants of the speech spectrum, leaving the first perceptual formant, $F1'$, to be equal to $F1$. They showed the validity of $F2'$-formulae for nine Swedish vowels by conducting an experiment where two-formant synthetic vowel stimuli were matched against the standard four-formant stimuli for minimal difference in vowel colour. Bladon and Fant [7] repeated this matching experiment for a full set of 18 IPA cardinal vowels and found the performance of the $F2'$ model given by Carlson et al. [6] to be unsatisfactory. Using the spectral prominence model, they derived a new formula for $F2'$ calculation based on the acoustic theory of speech production [8]. This formula could explain the results of a matching experiment for 17 of the 18 cardinal vowels.

In the present paper our aim is to study the various two-formant models of vowel perception reported in the literature. To do this the perceptual formant frequencies of different vowels were estimated for a group of subjects through a vowel identification experiment. As well as this, acoustic measurements were made on utterances containing these vowels made by the same subjects.

An earlier vowel identification experiment by Ainsworth [9] showed relatively lower second perceptual formant frequency values for the vowel /i/ than those predicted by the models of Carlson et al. [6] and Bladon and Fant [7]. Carlson et al. [6] tried to explain these relatively lower values for the vowel /i/ using two arguments. Firstly, in Ainsworth's experiment [9] the upper limit of the stimulus second formant frequency available to the subjects in the identification task was 2440 Hz; this may have led to low second perceptual formant frequency values for /i/. Secondly, the relative 'crowdedness' of the vowel system in the identification task might have biased the second formant frequency measures. In the vowel identification experiment reported in the present paper, the first point is answered by increasing the upper limit of the second formant frequency values available to the subjects to 2560 Hz. Also, to increase the confidence in the first and second perceptual formant frequency values obtained from the vowel identification experiment, the quantisation step size along the $F1$ and $F2$ dimensions was reduced, and the number of subjects and repetitions was increased. Details of the vowel identification experiment are given in the following Section.

The perceptual data with which to test the adequacy of two-formant models was obtained by an identification experiment. This experimental paradigm differs from the matching experiment used in most previous work. In matching the subjects are presented with both a full vowel and a two-formant construct; the second formant of the two-formant vowel is varied until the subject is satisfied with the match in quality. It has been reported [7] that subjects quickly atune to the experimental task and produce consistent results; Bladon and Fant give a measure of the subjects' performance. In the present experiments, the subjects heard a single two-formant vowel-like sound, and were required to identify the intended vowel.

2. Experiment

A two-formant model transforms the formants of a vowel spectrum into a pair of frequencies of a phonologically similar two-formant vowel. In order to assess the performance of such a model we need two sets of data: the formant frequencies $F1-F4$ measured from the speech spectrum, and the equivalent perceptual values $F1^p$ and $F2^p$ derived from the results of a two-formant synthetic vowel identification experiment. The procedures used to obtain these data are described below. The frequencies of the first four formants present in the speech spectrum were measured from the utterances of ten subjects for eleven English vowels. The frequency values of the perceived formant pairs were measured by presenting a set of synthetic vowel stimuli covering the entire $F1-F2$ plane to the same subjects for identification. The ten subjects were all unpaid volunteers: nine postgraduate students at the University of Keele studying topics unrelated to speech research and one of the authors (WAA).

2.1. Measurement of the vowel formants from the identification experiment

The stimulus sounds were generated by a parallel-formant, terminal analogue speech synthesiser similar to that described by Mattingly, Holmes and Shearme [10], controlled by a synthesis-by-rule program described by Ainsworth [11] running on a small digital computer. Two steps were taken in order to increase the consistency of the subjects' responses. Firstly, an /hVd/ context was chosen for the generation and presentation of the vowel stimuli on the grounds that the stimuli would be made more speech-like. Of the eleven syllables generated in this way, ten are familiar words of conversational English. The co-articulation effects within the generated syllable are relatively simple, especially with regard to the /h/-vowel assimilation. Secondly, a linear, falling intonation contour was produced through the syllable, with the fundamental frequency determined by the initial and final values of 169 Hz and 109 Hz, respectively. This made the stimulus sound like spoken syllables, but may have introduced slight diphthongisation.

The stimuli generated had two formants: $F1$ spanned the range 220–880 Hz in 12 steps of 60 Hz and $F2$ the range 760–2560 Hz in 16 steps of 120 Hz. The formant bandwidths were fixed, because of the limitations of the synthesiser employed, at 80 and 100 Hz, respectively. The amplitude of $F2$ was attenuated 6dB with respect to $F1$. In each experimental session a set of 192 stimuli were generated corresponding to a 12×16 matrix in the above ranges. The steady-state part of the vowel was fixed in duration to 240 ms. This value was chosen as previous experiments by Ainsworth [12] had shown it to lie approximately midway between the duration most appropriate for long and short vowels in isolated /hVd/ syllables.

The set of 192 stimuli was produced in pseudo-random order by the computer and were presented to all of the ten subjects via Sennheiser HD414 headphones. The subjects were asked to identify the stimulus heard as one of the words in the group: heard, hud, who'd, hood, hoard, hod, hard, had, head, hid, heed; and to press the appropriately labelled switch in front of them. An extra switch was provided in case the sound they heard did not correspond to any of the vowels in the response set of words. The computer did not generate the next stimulus until the responses from all the subjects participating in the session had been recorded; in this way each listener could take as long as he liked over the identification of the vowels. Each set of 192 synthetic stimuli was presented to the ten subjects in five repetitions; each time the order of presentation was changed. For a given subject and for a given repetition of the stimulus set, the responses of that subject defined different areas in the $F1 - F2$ plane for the eleven vowels. The centroid of each area gave the values of the perceptual formants $F1^p$ and $F2^p$ of the corresponding vowels. These perceptual formant pairs were obtained for each of the five repetitions for all the ten subjects and all eleven vowels.

2.2. Acoustic measurement of the vowel formants

The same ten subjects who participated in the identification experiment were used for the acoustic measurement of the vowel formants. The subjects were asked to read from a list of /hVd/ words: heard, hud, who'd, hood, hoard, hod, hard, had, head, hid, heed. Five repetitions were recorded in an ordinary office on a Revox A77 tape recorder and microphone. The utterances of these syllables were digitised at a sampling rate of 10 kHz by a Computer Automation Alpha computer. A semiautomatic analysis program was used for the spectral analysis. The selection of a 26.5 ms segment of the portion of maximum energy of the vowel was performed manually with the help of a CRT display cursor; then the log-power spectrum was obtained from a 256-point Fourier transform of the vowel segment, and this was smoothed using Cepstral smoothing [13]. The first four peaks in the spectrum were selected manually using a CRT display cursor and then the peak location was calculated by a 3-point parabolic interpolation. The measurements on each of the five repetitions of each vowel were obtained without reference to the intended vowel target. For each vowel a set of four formant frequencies was derived for all ten subjects.

3. A comparison of two-formant models

The data obtained from the experiment described in the preceding Section consist of the frequencies of the first four formants measured from the vowel spectrum for 11 different vowels and 10 different subjects (denoted by ($FN(j, k)$; $N = 1, 4; j = 1, 11; k = 1, 10$)), and the frequencies of the two formants measured from the identification of synthetic stimuli (denoted by [$FN^{\mathrm{P}}(j, k)$; $N = 1, 2; j = 1, 11; k = 1, 10$]). We have shown in an earlier paper [14] that subject-to-subject differences in formant frequencies measured from the vowel spectra do not show any correlation with differences in the corresponding formant frequencies derived from the identification of synthetic stimuli. So, for the purpose of exploring various two-formant models, we can reduce this data by averaging the formant frequencies over the ten speakers; i.e.,

$$FN_{\mathrm{av}}(j) = 1/10 \sum_{k=1}^{10} FN(j, k) \quad \text{for } N = 1, 4 \text{ and } j = 1, 11 \tag{1}$$

and

$$FN_{\mathrm{av}}^{\mathrm{P}}(j) = 1/10 \sum_{k=1}^{10} FN^{\mathrm{P}}(j, k) \quad \text{for } N = 1, 2 \text{ and } j = 1, 11. \tag{2}$$

These speaker-averaged formant frequencies are shown in Fig. 1 for all the eleven vowels.

3.1. Comparison procedure

In order to measure the performance of different two-formant models, we use the root-mean-

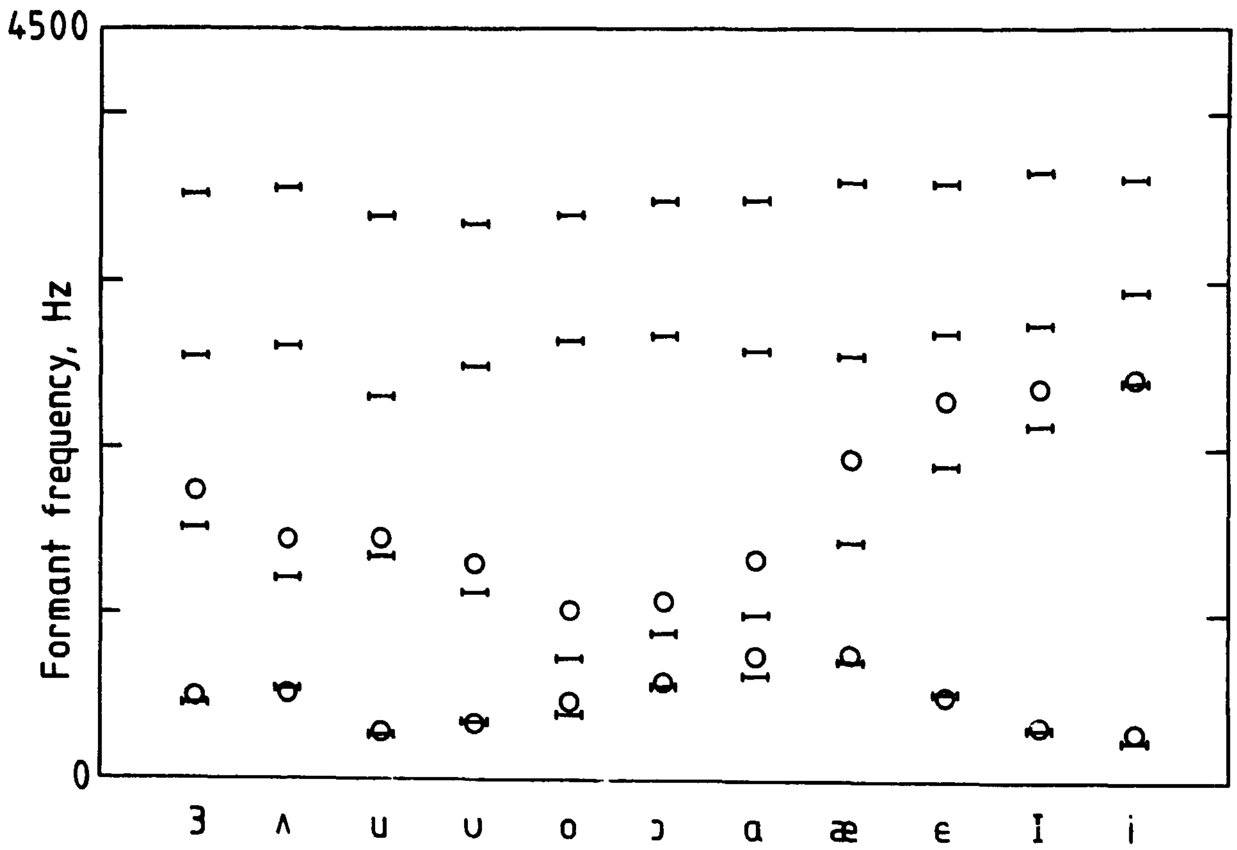

Fig. 1. Formant frequencies of elevent English vowels; solid bars showing $F1$, $F2$, $F3$ and $F4$ measured from speech spectra, circles showing $F1^{\mathrm{P}}$ and $F2^{\mathrm{P}}$ measured from the identification of synthetic vowel stimuli.

square (RMS) errors made by the model in predicting the first and second formant perceptual frequencies. We denote these errors by $E1$ and $E2$ and compute them using the relations

$$E1=\left[\sum_{j=1}^{11}\left(F1^{\mathrm{p}}_{\mathrm{av}}(j)-F1'_{\mathrm{av}}(j)\right)^2\right]^{1/2}/11, \qquad (3)$$

and

$$E2=\left[\sum_{j=1}^{11}\left(F2^{\mathrm{p}}_{\mathrm{av}}(j)-F2'_{\mathrm{av}}(j)\right)^2\right]^{1/2}/11, \qquad (4)$$

where $F1'_{\mathrm{av}}(j)$ and $F2'_{\mathrm{av}}(j)$ are the speaker-averaged frequencies of the first and second formants, respectively, predicted by the two-formant model for the jth vowel.

All the models described in the first Section, except the one given by Plomp [4], were investigated. Plomp's model uses the entire speech spectrum measured in one-third octave bands and computes the first two principal components as a weighted linear combination of the spectral amplitutes in all the one-third active bands. Two experiments reported in the literature [15–16] indicate that such a process is not likely to take place in the auditory system. Ainsworth and Millar [15] explored the perceptual effects as the second formant amplitude of synthetic stimuli was systematically reduced. They observed that the identity of each stimulus remained constant until the second formant amplitude was reduced to the level of noise, then changed abruptly to that of the corresponding back vowel. They concluded that the formants are not being calculated as the weighted linear sum of spectral amplitudes, but are more likely derived by a peak-picking mechanism. Sapozhkov [15] reported high intelligibility for clipped speech in spite of broad-band spectral distortions whilst the peaks corresponding to the original formant structure are retained. These experiments strongly suggest that the model given by Plomp is not always an appropriate model of the auditory processing of vowel sounds, and, hence, has not been included in the comparison study in the present paper. We now investigate the three models given by Carlson et al., Bladon and Fant, and Karnickaya et al.

3.2. The Carlson model

The two-formant model of Carlson et al. [6] takes the first formant frequency, $F1'$, of the model to be the same as that extracted from the speech spectrum; i.e., $F1' = F1$. The second formant frequency, $F2'$, of the two-formant model is derived from the second, third and fourth formants ($F2$, $F3$ and $F4$) without taking into account their associated amplitudes by employing the spectral prominence model. The empirical relation for $F2'$ in their model is given by

$$F2' = \left(F2 + C(F3F4)^{1/2}\right)/(1+C), \qquad (5)$$

where

$$C = (F1/500)^2((F2-F1)/(F4-F3))^4 \times ((F3-F2)/(F3-F1))^2.$$

The RMS errors $E1$ and $E2$ are computed from Eqs. (3) and (4) by using the values $F1'$ and $F2'$ given by this model. These errors were found to be 53.3 and 199.2 Hz.

3.3. The Bladon and Fant model

The two-formant model of Bladon and Fant [7] uses the formant frequencies of $F2$, $F3$ and $F4$ together with their associated amplitudes to locate the position of spectral dominance. The first and second formant frequencies $F1'$ and $F2'$, are given by

$$F1' = F1 \qquad (6)$$

and

$$F2' = F2 + C^2(F3F4)^{1/2}/(1+C^2), \qquad (7)$$

where

$$C = \left[K(f)67F2(1-F1^2/F2^2)(1-F2^2/F3^2) \times (1-F2^2/F4^2)\right] \times \left[(F4-F3)^2(F3F4/F2^2-1)\right]^{-1}$$

and

$$K(f) = 12F2/1400.$$

The RMS errors $E1$ and $E2$ computed for this model are 53.5 Hz and 287.4 Hz, respectively. The RMS error $E2$ for this model is larger than that for the Carlson model, though for a matching experiment this model resulted in a better performance.

3.4. The Karnickaya model

In the two-formant model of Karnickaya et al. [5], the locations of the first two peaks in the equivalent auditory spectrum are considered as the frequencies of $F1'$ and $F2'$ of the two-formant model. The equivalent auditory spectrum is constructed by applying four levels of physiologically-motivated transformations on the vowel waveform: filter bank analysis, auditory threshold logic, a mapping into a loudness density function and a process of lateral inhibition. As an approximation to auditory spectral processing, we calculated the locations of the peaks in the auditory equivalent spectrum from the first four formant frequencies, $F1$, $F2$, $F3$ and $F4$, as follows. If the formants were separated from each other by more than the critical bandwidth, then they remained as peaks in the auditory equivalent spectrum. However, if these formants, say $F2$ and $F3$, were close to each other so that their separation, $F3 - F2$, was less than the critical bandwidth, $0.10(F2 + F3)/2$, then these two formants were considered to be merged into a single peak at location $(F2 + F3)/2$. Evans [17] has reported that critical-bandwidths are of the order of 10% of the frequency for the range of the first four formants of speech spectra. (It may be noted here that for the formant data obtained in this experiment, the formants $F1 - F4$ were found to be separated from each other by more than the critical bandwidth for all the vowels and for all the subjects. Effectively, then, for the data used here there would be no peak-merging.)

The locations of peaks in the auditory equivalent spectrum were obtained from $F1$, $F2$, $F3$ and $F4$ using this procedure. Taking the locations of the first two peaks as the frequencies $F1'$ and $F2'$ of the two-formant model, RMS errors $E1$ and $E2$ were determined using (3) and (4). The errors $E1$ and $E2$ for this model were found to be 53.5 Hz and 275.8 Hz, respectively.

The Carlson model gives a better performance than either the Bladon and Fant or the Karnickaya models when we compare the RMS errors $E2$. However, even with this model, the error $E2$ is approximately 200 Hz which is large in comparison with the error resulting from the vowel matching experiment reported by Carlson et al. Also, these models consider the frequency $F1'$ to be the same as $F1$. However, it can be seen from Fig. 1 that the first formant frequency obtained from the vowel identification experiment is higher than $F1$ for vowels having the first two formants close to each other. This strongly suggests the need for an alternative two-formant model which can better explain the results of vowel identification experiments. In the following section we present such a model.

4. A two-stage model

The two-formant model we propose to develop in this paper is broadly based on the results of perception experiments with split-formant vowel sounds [6,18,19]. In these perception experiments, one or more formants of the vowel sound were presented to one ear and the remaining formants in the other ear. Results of these experiments have shown that subjects could identify the vowel sound in this split mode of presentation, indicating the integration of timbre at a central level of the auditory system [6,18]. This suggests that vowel identification is two-stage process.

The present model handles the auditory processing of vowel sounds in two stages: a peripheral processing stage and a central processing stage. In the first stage, the speech spectrum of the vowel sound is transformed to its auditory equivalent as proposed by Karnickaya et al. [5]. The formant frequencies are extracted from the auditory equivalent spectrum (in which formants within a critical band have been combined into a single peak) by a peak-picking mechanism. The peaks obtained are labelled $P1-P4$. In the second stage, the two-formant approximation is performed: this stage combines the effects of all the $P1-P4$ to obtain the frequencies $F1'$ and $F2'$ of the model.

It can be observed from Fig. 1 that the difference between the first formant frequency values obtained through the vowel identification experiment and from the vowel spectrum depends on the separation of other formants from the first formant, the dependence being inversely proportional to this separation. A similar observation can be made from Fig. 1 for the second formant. Thus, the frequency of $F1'$ in the present model is obtained as a sum of $P1$ and a perturbation term accounting for the interaction with the other formants $P2$–$P4$; each interaction being inversely proportional to the separation between this formant and $P1$. We can write $F1'$ as a function of $P1$–$P4$ as follows:

$$F1' = P1 + C1P2/(P2 - P1) + C2P3/(P3 - P1) + C3P4/(P4 - P1), \tag{8}$$

where $C1$, $C2$ and $C3$ are constants determined from the experimental data by the method of least squares fitting. Similarly, the frequency $F2'$ is obtained from the relation:

$$F2' = P2 + D1P1/(P2 - P1) + D2P3/(P3 - P2) + D3P4/(P4 - P2), \tag{9}$$

where $D1$, $D2$ and $D3$ are constants determined from the experimental data by the method of least-squares fitting.

In order to study the effectiveness of the present model described by (8) and (9), were require the formant frequencies P_1–P_4. These are obtained in the present model from the auditory equivalent spectrum using a peak-picking mechanism. The procedure of obtaining the locations of the peaks in the auditory equivalent spectrum from the frequencies $F1$–$F4$ has already been described in the preceding Section. Furthermore, if we use the experimental data obtained in Section 2 to determine the constants $C1$, $C2$, $C3$, $D1$, $D2$, and $D3$ using the least-squares fitting method and then use the same experimental data for evaluating the performance of our model, we may get an (optimistically) biased estimate. In order to achieve an unbiased performance estimate, we used the data from nine subjects to calculate $F1'$ and $F2'$ for the remaining tenth subject. This procedure was repeated ten times, each time calculating $F1'$ and $F2'$ for a new subject by using the constants $C1$, $C2$, $C3$, $D1$, $D2$ and $D3$ determined from the data from the remaining nine subjects. The values of $F1'$ and $F2'$ so obtained were averaged over the ten subjects and the RMS errors $E1$ and $E2$ were calculated by using (3) and (4). The errors $E1$ and $E2$ for the present model were found to be 37.8 Hz and 125.2 Hz, respectively.

It has been mentioned already that the interaction of other formants with the main formant is inversely proportional to their separation from the main formant. From this we expect the interaction of non-neighbouring formants to be small. In order to ascertain that this interaction is negligible, we excluded this interaction term from Eqs. (8) and (9) and rewrote them as follows:

$$F1' = F1 + C1F2/(F2 - F1) \tag{10}$$

and

$$F2' = F2 + D1F1/(F2 - F1) + D2F3/(F3 - F2). \tag{11}$$

The RMS errors $E1$ and $E2$ were calculated for the model given by (10) and (11) using the procedure described above. These errors were found to be 38.3 Hz and 126.1 Hz. It can be seen that the RMS errors resulting from the model by (10) and (11) are comparable with the corresponding errors resulting from the model given by (8) and (9). Thus, we can assume the interactions due to non-neighbouring formants to be negligible and, hence, take (10) and (11) to define the present two-formant model. The constants $C1$, $D1$, $D2$, of the current model are 16.7, 164.9 and 33.3, respectively.

In order to compare the performance of this model with those mentioned in Sections 3.1 to 3.3 above, the RMS errors $E1$ and $E2$ are listed in Table I. Data taken before averaging over speakers is used here to enable comparison with the breakdown for individual vowels given in Table II. The

Table 1
Summed RMS errors $E1$ and $E2$ for different two-formant models on data before averaging over speakers using Hertz, mel and Bark frequency scales

	Hertz		Mel		Bark	
Model	$E1$	$E2$	$E1$	$E2$	$E1$	$E2$
Carlson	95	343	90	183	0.83	1.40
Bladon	95	359	90	198	0.83	1.49
Karnickaya	95	352	90	202	0.83	1.53
Two-stage	90	244	85	133	0.78	1.02

Table 2
Comparison between the Bladon and Fant model and the two-stage model. RMS errors $E1$ and $E2$ for individual vowels using data before averaging over speakers calculated for a mel frequency scale

Vowel	$E1$ Bladon and Fant	$E1$ Two-stage	$E2$ Bladon and Fant	$E2$ Two-stage
3	112	104	199	126
ʌ	107	119	187	122
u	56	49	150	127
ʊ	56	71	151	87
o	97	70	264	143
ɔ	85	87	175	170
a	113	81	244	150
æ	81	74	269	165
e	100	109	121	120
I	90	90	130	100
i	75	59	232	126

error summation for individual vowels requires to be calculated on data averaged only over repetitions, not speakers, to increase the error value significance. It can be seen that the two-stage model results in smaller errors than the other models. The present model performs better in predicting the formant frequency values obtained from the identification experiment.

The frequency values obtained in the two experiments were measured in the usual physical units of Hertz. It is likely, however, that some psychophysical unit of frequency, such as mel or Bark, is more appropriate for the perceptual measurements. Accordingly all the measurements were transformed into mels and Barks, and the calculation of $E1$ and $E2$ were repeated. The transformation from Hertz to mel was calculated using the equation given by Fant [20]

$$m = 1000 \log_2(1 + f/1000) \qquad (12)$$

where f is the frequency in Hz and m in mels. The transformation from Hertz to Bark was calculated using the equation given by Schroeder et al. [21]

$$f = 650\sinh(x/7) \qquad (13)$$

where f is the frequency in Hz and x in Barks. These results are also shown in Table 1. The same pattern of results emerged with each of the frequency units.

It is interesting to examine the performance of the models with the individual vowels. Table II shows $E1$ and $E2$ (in mels) for each of the eleven vowels on non-averaged data. The results obtained with the two-stage model are compared with those from the Bladon and Fant model. It will be seen that the values of $E1$ obtained with the present model are less for six vowels (and equal for one) than the values obtained with the Bladon and Fant model. The values of $E2$ are less for all the eleven vowels.

In particular it will be seen that the value of $E2$ for /i/ using the two-stage model is about equal to the average of $E2$ for all vowels. This suggests that the value of $F2^p$ obtained for /i/ was not an artefact of the experimental technique employed.

5. Conclusion

In the present paper we have compared various two-formant models reported in the literature in their prediction of vowel formant frequencies obtained from a vowel identification task. Formant frequencies for eleven English vowels were measured for a group of subjects in two ways: firstly from the utterances of the vowels spoken in a /hVd/ context by these subjects and, secondly, by presenting a set of synthetic /hVd/ stimuli spanning the entire $F1$–$F2$ plane for natural vowels to the same subjects for identification. It was found that existing models could be improved in their prediction of vowel identification results. We propose an alternative two-formant model which performs the auditory processing of vowel sounds in two stages: a peripheral and a central processing stage. In the first, peripheral, stage the vowel spectrum is transformed to its auditory equivalent and the formant frequencies are extracted from this spectrum using a peak-picking mechanism. In the second, central, stage a two-formant approximation is performed in which the first and second formant frequencies of the model are taken to be equal to the valus obtained from the peripheral processing together with a perturbation term which depends on neighbouring formants. The perturbation caused by each of these neighbours is proportional to its separation from the main formant. The model gives a more satisfactory account of the experimental results than the other models we have discussed.

Acknowledgements

K.K. Paliwal wishes to thank the Indian National Sciences Academy (INSA) for awarding the fellowship under the INSA-Royal Society, London, Exchange Program. K.K. Paliwal was on deputation from the Tata Institute of Fundamental Research, Bombay. D. Lindsay was supported by the Medical Research Council, London. We wish to thank all the subjects for their patience throughout the experimental sessions.

References

[1] P. Delattre, A.M. Liberman, F.S. Cooper and L.J. Gerstman, "An experimental study of the acoustical determinants of vowel colour', *Word*, Vol. 8, 1952, pp. 195–210.

[2] G.E. Peterson and H.L. Barney, "Control methods used in the study of vowels", *J. Acoust. Soc. Amer.* Vol. 24, 1952, pp. 175–184.

[3] L.C.W. Pols, L.J.Th. van der Kamp and R. Plomp, "Perceptual and physical space of vowel sounds", *J. Acoust. Soc. Amer.*, Vol. 46, 1969, pp. 458–467.

[4] R. Plomp, "Auditory analysis and timbre perception", in: G. Fant and M.A.A. Tatham, eds., *Auditory Analysis and Perception of Speech*, Academic Press, London, 1975, pp. 7–22.

[5] E.G. Karnickaya, V.N. Mushnikov and N.V. Slepokurova, "Auditory processing of steady-state vowels", in: G. Fant and M.A.A. Tatham, eds., *Auditory Analysis and Perception of Speech*, Academic Press, London, 1975, pp. 37–53.

[6] R. Carlson, G. Fant and B. Granstrom, "Two-formant models, pitch and vowel perception", in: G. Fant and M.A.A. Tatham, eds., *Auditory Analysis and Perception of Speech*, Academic Press, London, 1975, pp. 55–82.

[7] R.A.W. Bladon and G. Fant, "A two-formant model and the cardinal vowels", *Speech Transmission Laboratory Quarterly Progress Resport*, Stockholm, KTH, Jan. 1978, pp. 1–8.

[8] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[9] W.A. Ainsworth, "Perception of synthesised vowels and h-d vowels as a function of fundamental frequency", *J. Acoust. Soc. Amer.* Vol. 49, 1971, pp. 1323–1324.

[10] J.N. Holmes, I.G. Mattingly and J.N. Shearme, "Speech synthesis by rule", *Language and Speech*, Vol. 7, 1964, pp. 127.

[11] W.A. Ainsworth, "Performance of a speech synthesis system", *International Journal of Man-Machine Studies*, Vol. 6, 1974, p. 493.

[12] W.A. Ainsworth, "Duration as a cue in the recognition of synthetic vowels", *J. Acoust. Soc. Amer.*, Vol. 51, 1972, pp. 648–651.

[13] R.W. Schafer and R.W. Rabiner, "System of automatic formant analysis of voiced speech", *J. Acoust. Soc. Amer.* Vol. 47, 1970, pp. 634–648.

[14] K.K. Paliwal, D. Lindsay and W.A. Ainsworth, "Correlation between production and perception of English vowels", *J. Phonetics*, Vol. 11, 1983, pp. 77–83.

[15] W.A. Ainsworth and J.B. Millar, "The effect of relative formant amplitude on the perceived identity of synthetic vowels", *Language and Speech*, Vol. 15, 1972, pp. 328–341.

[16] G. Fant, "Vowel perception: Chairman's review", in: G. Fant and M.A.A. Tatham, eds., *Auditory Analysis and Perception of Speech*, Academic Press, London, 1975, pp. 1–6.

[17] E.F. Evans, "Functions of the auditory system", in: H.B. Barlow and J.D. Mollon, eds., *The Senses*, Cambridge University Press, Cambridge, 1982, pp. 307–332.

[18] D.E. Broadbent and P. Ladefoged, "On the fusion of sounds reaching different sense organs", *J. Acoust. Soc. Amer.*, Vol. 29, 1957, pp. 708–710.

[19] W.A. Ainsworth, "Perception of speech sounds with alternate formants presented to opposite ears", *J. Acoust. Soc. Amer.*, Vol. 63, 1978, pp. 1528–1534.

[20] G. Fant, "Acoustic analysis and synthesis of speech with applications to Swedish", *Ericsson Technics*, Vol. 1, 1959, pp. 3–108.

[21] M.R. Schroeder, B.S. Atal, and S.L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception", in: B. Lindblom and S. Ohman, eds., *Frontiers of Speech Communication Research*, 1979, London, Academic Press, pp. 217–229.