Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

A two-formant model and the cardinal vowels

Bladon, A. and Fant, G.

journal: STL-QPSR
volume: 19
number: 1
year: 1978
pages: 001-008

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

I. SPEECH ANALYSIS

A. A TWO-FORMANT MODEL AND THE CARDINAL VOWELS

R.A.W. Bladon and G. Fant*

Abstract

An improved version of the F_2' formula for calculating an "effective" upper formant from a set of F_1, F_2, F_3, F_4 values has been developed. It has been tested with a set of cardinal vowels. The relations within the system come out as expected with the exception of [ɑ] where the calculated F_2' was found to be too large. The general limitations of the F_2' concept are discussed.

The notion that vowel quality can be satisfactorily approximated by an acoustically-derived representation in two dimensions has held a long-standing attraction for phoneticians, in the hope that such acoustic data might provide correlates for the articulatory dimensions of tongue height and tongue retraction, which form the basis of conventional phonetic vowel quality diagrams. More recently, perceptual studies have supported the view that a two-formant model of a vowel is also a valid representation at some level (not necessarily peripheral) of auditory processing. In this connection, for example, the work of Plomp (1975) and his associates demonstrated that Dutch vowel identification scores based on a principal-components analysis of vowel spectra represented in multidimensional space (each dimension corresponding to the sound-pressure level in a one-third octave filter) displayed excellent agreement with vowel identification scores based on a two-formant (log F_1 versus log F_2) plane. Further evidence was provided by Carlson, Fant, and Granström (1975) and Carlson, Granström, and Fant (1970) in achieving successful matchings of Swedish vowels by a two-formant approximation F_1 versus F_2', where F_2' was derived from F_2 by a variable degree of upward weighting to allow for the contribution of upper formants.

As a means of predicting the results of such matching experiments, Carlson et al (1970) proposed a formula designed to produce a continuous shift of F_2' between the extreme values of F_2 and $(F_3F_4)^{1/2}$,

---

* Guest researcher at KTH 1 April to 30 June 1977. University College of North Wales, Bangor, Great Britain.

taking into account relations between formant frequencies and spectrum shape. That formula (henceforth the 1975 formula) was as follows:

$$F_2' = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c}$$

$$c = \left(\frac{F_1}{500}\right)^2 \left(\frac{F_2 - F_1}{F_4 - F_3}\right)^4 \left(\frac{F_3 - F_2}{F_3 - F_1}\right)^2 \qquad (1)$$

The 1975 formula appeared successful, since it predicted F_2' for nine Swedish vowels to within 120 Hz, with an average error of only 63 Hz.

We decided to test the 1975 formula for calculating F_2' against a full set of 18 IPA-cardinal vowels. These were spoken by one of the authors and analyzed spectrographically. A set of matching experiments was then carried out, in the following way. 18 four-formant vowels were synthesized, each with the formant frequencies observed in the analysis of a cardinal vowel. Four listeners were asked to match, as nearly as possible to a four-formant stimulus, a synthetic two-formant vowel whose F_1 was unchanged but whose F_2 was variable.

Application of the 1975 formula to the cardinal vowel data gives results which are most unsatisfactory. Of the 18 cardinal vowels, ten show an error in predicted F_2' of over 120 Hz, and the mean error is 232 Hz! These discrepancies are reflected in Fig. I-A-1, a mel scale plot of F_1 against F_2' (matched and calculated). We have adopted the "technical mel scale" of Fant (1959):

$$m = \frac{1000 \log (1 + f/1000)}{\log 2} \qquad (2)$$

It can be seen that the calculated values fail to bring out the perceptual distinctiveness of back unrounded vowels (which do not occur in Swedish and thus had not been part of the 1975 test material), and of [ɒ] and [ʉ]. In all these cases, calculated F_2' is much too high. It means, in particular, that the contrast between the rounded and unrounded counterparts [o] and [ɤ] is completely obscured, the relationship of [ɨ] to [ʉ] is displaced (inverted in the F_2' dimension) and the open vowel area is improbably overcrowded.

An improved version of a formula for predicting F_2' has therefore been developed. In this paper we report on some initial trials

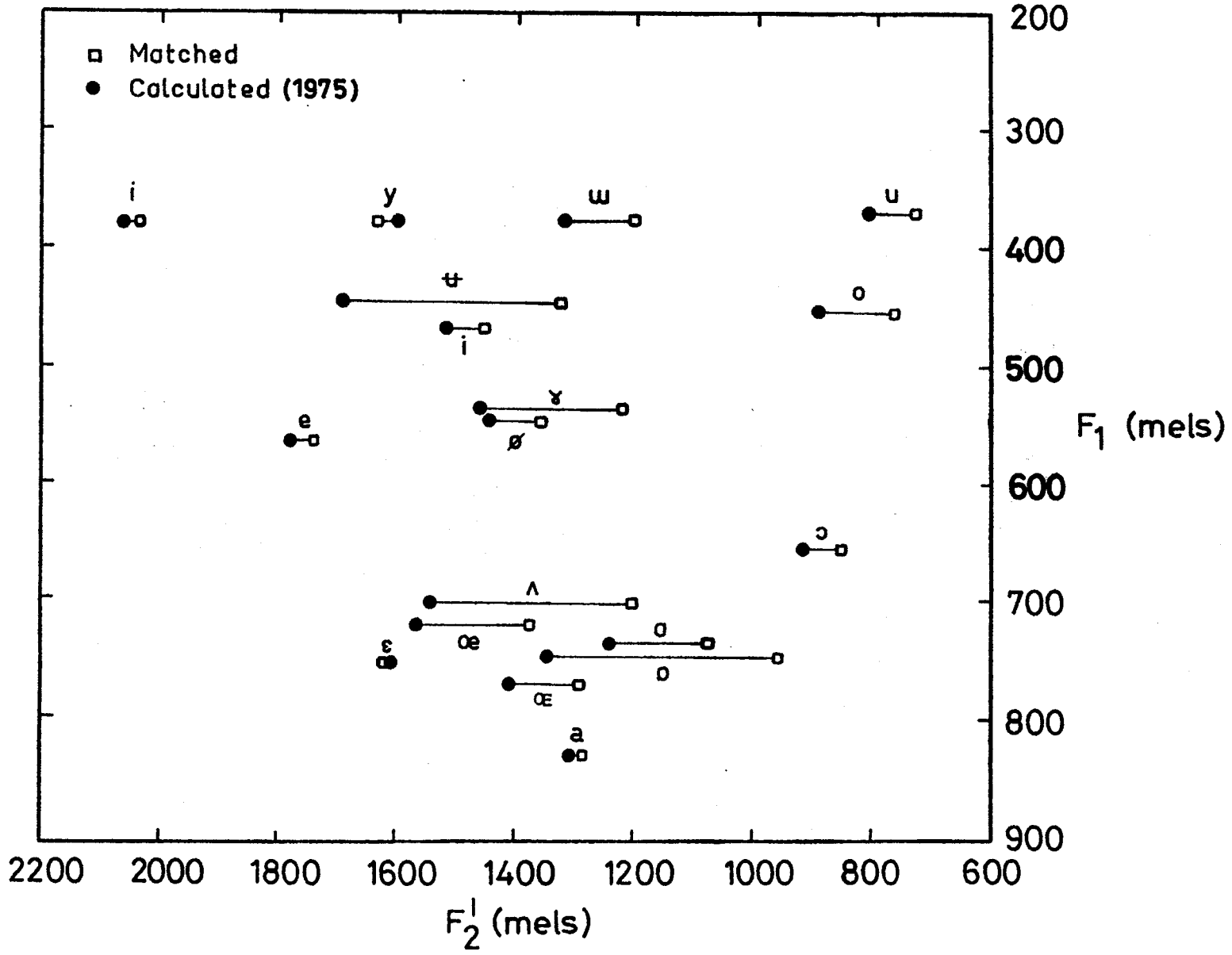

Fig. I-A-1. Matched and calculated F_2' of 18 cardinal vowels illustrating the shortcomings of the 1975 F_2' formula.

of the formula (henceforth the 1977 formula), on some possible refinements to it, and, superficially, on the extent to which two-formant matchings themselves were successfully accomplished.

The 1977 formula, like its predecessor, produces a continuous shift of F_2' between the extreme frequencies of F_2 and $(F_3F_4)^{1/4}$: however, its basis in acoustic theory is more firm than the rather intuitive 1975 version. The new formula is developed according to a spectrum prominence model which postulates interdependencies between formant frequencies and spectrum levels (Fant, 1960), and is more systematically designed from s-plane residue calculations:

$$F_2' = \frac{F_2 + c^2 (F_3F_4)^{1/2}}{1 + c^2} \tag{3}$$

$$c = K(f) \cdot \frac{A_{34}}{A_2}$$

Where A_{34} is the vocal tract transfer function in the valley between F3 and F4 at the frequency $F_{34} = (F_3F_4)^{1/2}$ and A_2 is the transfer function at the second formant peak F2.

The ratio between A_{34} and A_2 may be derived from S-plane vectorial products as

(4)

$$\frac{A_{34}}{A_2} = \frac{F_2B_2(F_2-F_1)(F_2+F_1)(F_3-F_2)(F_3+F_2)(F_4-F_2)(F_4+F_2)}{(F_{34}-F_3)(F_4-F_{34})(F_{34}-F_2)(F_{34}+F_2)(F_{34}-F_1)(F_{34}+F_1)(F_{34}+F_3)(F_{34}+F_4}$$

With the approximation $(F_3F_4)^{1/2} = \frac{1}{2}(F_3+F_4)$ this expression reduces to

$$\frac{A_{34}}{A_2} = \frac{B_2 \cdot F_2 (1-F_1^2/F_2^2)(1-F_2^2/F_3^2)(1-F_2^2/F_4^2)}{(F_4 - F_3)^2 (\frac{F_3F_4}{F_2^2} - 1)} \tag{5}$$

The factor K(f) in the weighting function c is intended to include additional preemphasis originating from source, radiation, and higher pole corrections and in addition a correction for differences in equal loudness levels. In the first approximation we shall disregard their combined frequency dependency and start out the calculations with $K(f) = 12$ and $B_2 = 100$ Hz.

Table I-A-I. Comparison of matched values of F_2' with calculated values of F_2'. Each calculated value is followed by its difference Δ from the matching (= prediction error).

	Measured formant frequencies				F_2' matched		F_2' calculations						
							1975		1977				
									$B_2=100$ *)		$B_2=67$ **)		
	F_1 Hz	H_2 Hz	F_3 Hz	F_4 Hz	Hz	mels	mels	Δ	mels	Δ	Hz	mels	Δ
i	300	2300	3070	3590	3095	2034	2062	26	2059	25	3190	2067	33
e	470	2180	2720	3790	2361	1749	1780	31	1744	5	2361	1749	0
ɛ	680	1890	2580	3940	2076	1621	1614	7	1557	64	1932	1552	69
a	770	1400	2460	3710	1452	1294	1316	22	1276	18	1410	1269	25
ɑ	660	1170	2770	3650	1103	1073	1248	175	1144	71	1182	1126	53
ɔ	570	840	2640	3310	806	853	922	69	892	39	842	881	28
o	370	730	2670	3240	700	766	896	130	814	48	733	793	27
u	290	700	2550	3280	669	739	807	68	775	36	700	766	27
y	300	1890	2250	3000	2101	1633	1600	33	1659	26	2125	1644	11
ø	460	1520	2290	3290	1570	1362	1442	80	1398	36	1583	1369	7
œ	640	1450	2330	3030	1637	1399	1578	179	1458	59	1612	1385	14
Œ	700	1430	2390	3350	1458	1294	1420	126	1330	36	1471	1305	11
ɒ	670	1050	2900	3490	947	961	1346	385	1096	135	1072	1051	90
ʌ	620	1260	2390	3610	1284	1192	1542	350	1189	3	1266	1180	12
ɤ	450	1300	2640	3470	1326	1217	1460	243	1285	68	1354	1235	18
ɯ	300	1320	2480	3440	1300	1202	1325	123	1273	71	1359	1238	36
ɨ	380	1690	2460	3570	1754	1462	1517	55	1485	23	1763	1466	4
ʉ	360	1550	2430	3030	1503	1324	1696	372	1656	332	1479	1575	251
Total prediction error (mels)								2474		1095			716

*) $K(f) = 12$

**) $K(f) = 12 \cdot \frac{F_2}{1400}$

As seen in Table I-A-I (p. 4) there is a substantial improvement. The total prediction error in mels summed over the entire ensemble of vowels drops from 2474 in the 1975 formula to 1095. The residual error is further decreased to 676 mels by the choice of B_2 = 50 Hz and

$$K(f) = 12\ (\frac{F_2}{1400}) \qquad (6)$$

This empirical correction takes into account a falling -6 dB/octave slope for the sum of source and radiation in the lower F_2 domain.

The optimal B_2 = 50 is acoustically representative but has no independent significance here since it combines with the factor 12 F_2/1400. At F_2 = 1400 K(f) = 12 or 21.5 dB which is a representative difference between the spectrum level of F2 and the valley between F3 and F4 for a neutral vowel.

Table I-A-II.

B_2 =	Residual prediction error 150	100	83	67	50	33 Hz
K(f) = 12	1828	1095	868	787	817	899
$K(f) = \frac{12F_2}{1400}$	1901	1202	952	716	676	771

Discussion

The outcome of the F_2' calculations in relation to the matched F_2' values is shown in Fig. I-A-2. We have chosen the near optimal B_2 = 67 Hz and K(f) = 12 F_2/1400. The relations within the vowel system are now restored. With the exception of the vowel [ʉ] the prediction error is of the order of 20-30 mels or 40-100 Hz, which is of the order of a difference limen in formant frequency discrimination. A further insight in the outcome of the matchings and the calculations is obtained from Fig. I-A-3 where these data are compared to the set of measured F_1, F_2, F_3, and F_4 of the natural vowels.

The calculated F_2' values of [u] [o] [ɔ][a][ɒ][ʌ] follow F_2 within an average of 7 Hz, whilst the matched F_2' of these vowels differ

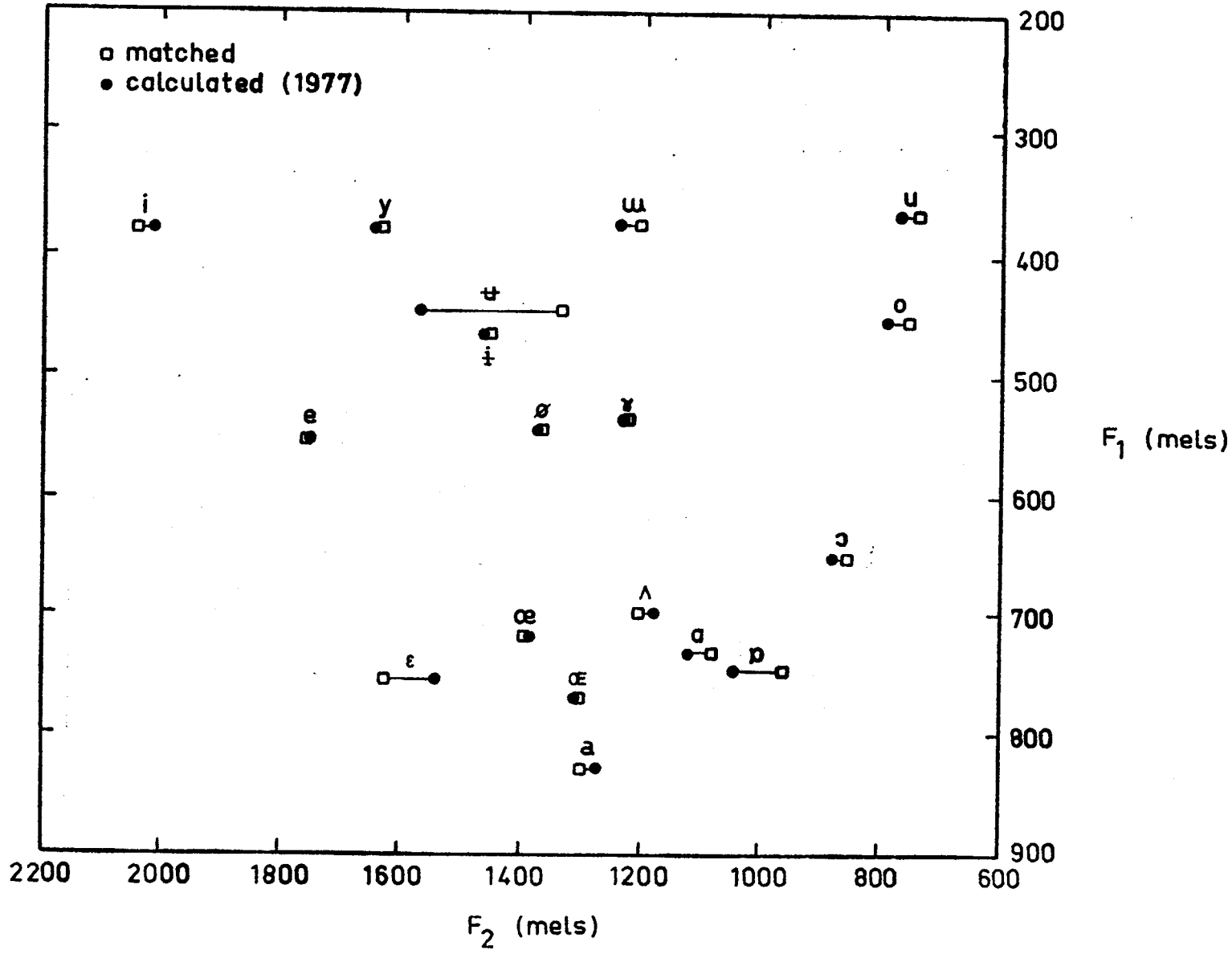

Fig. I-A-2. Matched and calculated F_2 of 18 cardinal vowels. 1977 formula with B=67 Hz and $K(f)=12F_2'/1400$.

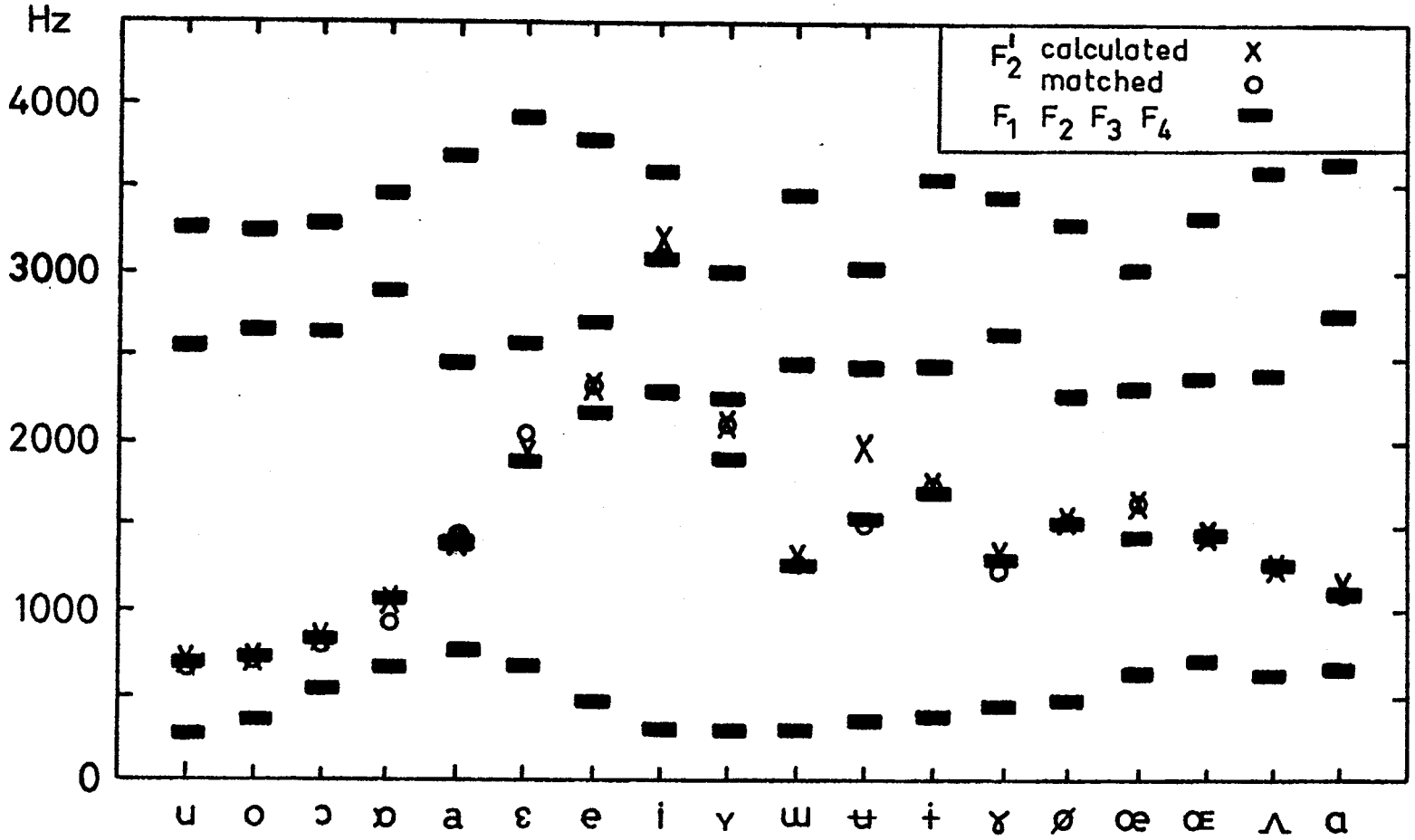

Fig. I-A-3. Sequential presentation of the vowels in terms of measured formant data F_1, F_2, F_3, and F_4 together with matched and calculated F_2'.

from the F_2 values by a mean of 44 Hz. There is a tendency of matched F_2' being on the low side of F_2, e.g. in [ɒ] where F_2 = 1050 Hz and F_2' is matched to 947 Hz whilst the calculated F_2' is 1072 Hz.

The calculated and matched F_2' of [i] came out close to F_3 and between F_2 and F_3 in [Y] [e] and [œ]. In the remaining vowels [ʏ] [ɨ] [ɛ] and [ŒE] the matched and calculated F_2' are again close to F_2.

The only case of a major discrepancy between calculated and measured F_2' appears to be the vowel [ʉ]. Spectrograms of the vowels [ʉ] and [ɨ] are shown in Fig. I-A-4. In spite of the relative intense F4 the matched F_2' of [ʉ] coincides with F_2 which is lower than the F_2 and matched F_2' of [ɨ]. On the other hand, in terms of calculated F_2' this expected phonetic order is not retained. The excessively high calculated F_2' of [ʉ] can be traced back to the high sensitivity of the formula to a small distance between F_4 and F_3. This property of the formula is needed to ensure the proper contrast between [i] and [y] but it appears to be less representative of spectral dominance in mid and back vowels. A gross error in the estimate of F_4 may accordingly have rather noticeable effects in the calculations.

We now turn to the more general question of the validity of a two-formant model in perception. In setting an experimental subject the task of matching a two-formant vowel to a four-formant one, and in using those matchings to evaluate the success of numerical methods, we are making the assumption that the auditory pattern above F_1 can be approximated by a single perceptual variable. How valid is that assumption? How difficult is the matching task? From informal observations of the subjects' reactions it seemed clear that the matching task is a realistic one to set, and that for the majority of cardinal vowels it is not at all difficult to do.

However, employing a readily available but gross and rather superficial measure, we can consider the number of listening trials that preceded the decision by a subject that a match had been obtained.

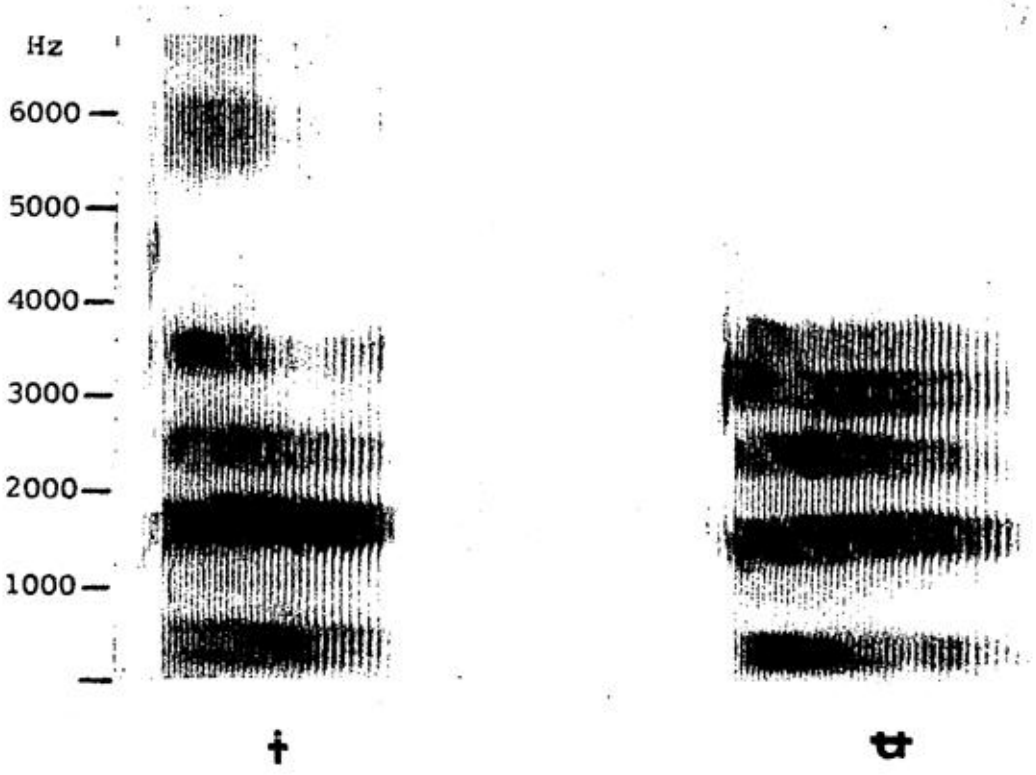

Fig. I-A-4. Spectrograms of the vowels [ɨ] and [ʉ].

On that basis, the spectrum of the vowels [i] [e] and [y] could be matched only with some difficulty. The number of trials needed is plotted on Fig. I-A-4 as isolated data points. Also shown in Fig. I-A-4 is some further evidence which could be interpreted in the same way, namely that the standard deviation in F_2' responses was especially high in the vowels [i] and [e]. This finding is consistent with the fact that the upper-frequency spectrum of [i e y] is more significantly shaped by the upper formants $F_3 \ldots F_n$ than is the case in other vowels.

Indeed a tendency was observed (which needs further corroboration) for listeners to fall into two groups according to whether they selected their match for say [i] at around 1700 mels ($\approx$ 2250 Hz) or at around 2200 mels ($\approx$ 3600 Hz). The first of these frequencies is closely coincident with F_2, the second with F_4, and it therefore appears that subjects tended to allow either F_2 or F_4 to dominate the percept. We can therefore conclude that for the vowels [i e y] it is, tentatively, less convincing than for the remaining 15 cardinal vowels to suppose that a perceptual correlate of the formants above F_1 exists in the form of a single parameter F_2.

Other spectrum attributes such as the relative spacing of formants within the F2 F3 F4 F5 probably enter as additional cues.

Although the F_2' formula has been substantially improved we do not consider it to have reached a state of perfection, where it can be recommended for routine data reduction of formant measurements. Further work should be directed to auditory projections of spectrum profiles to test the relevance of loudness density mappings and possible lateral suppression. A fundamental question is to what extent the percept of phonetic quality is discontinuous in situations where the stimulus is allowed to change from F2 to F3 or F4 prominence. Karnickaya et al (1975) report bimodal distribution of response just as we have noted in our study. This finding would support their view that the major correlate of vowel quality is the position in the auditory space of F1 and the next highest peak in the processed spectral projection. In the phonetic boundary of equal prominence of F2 and

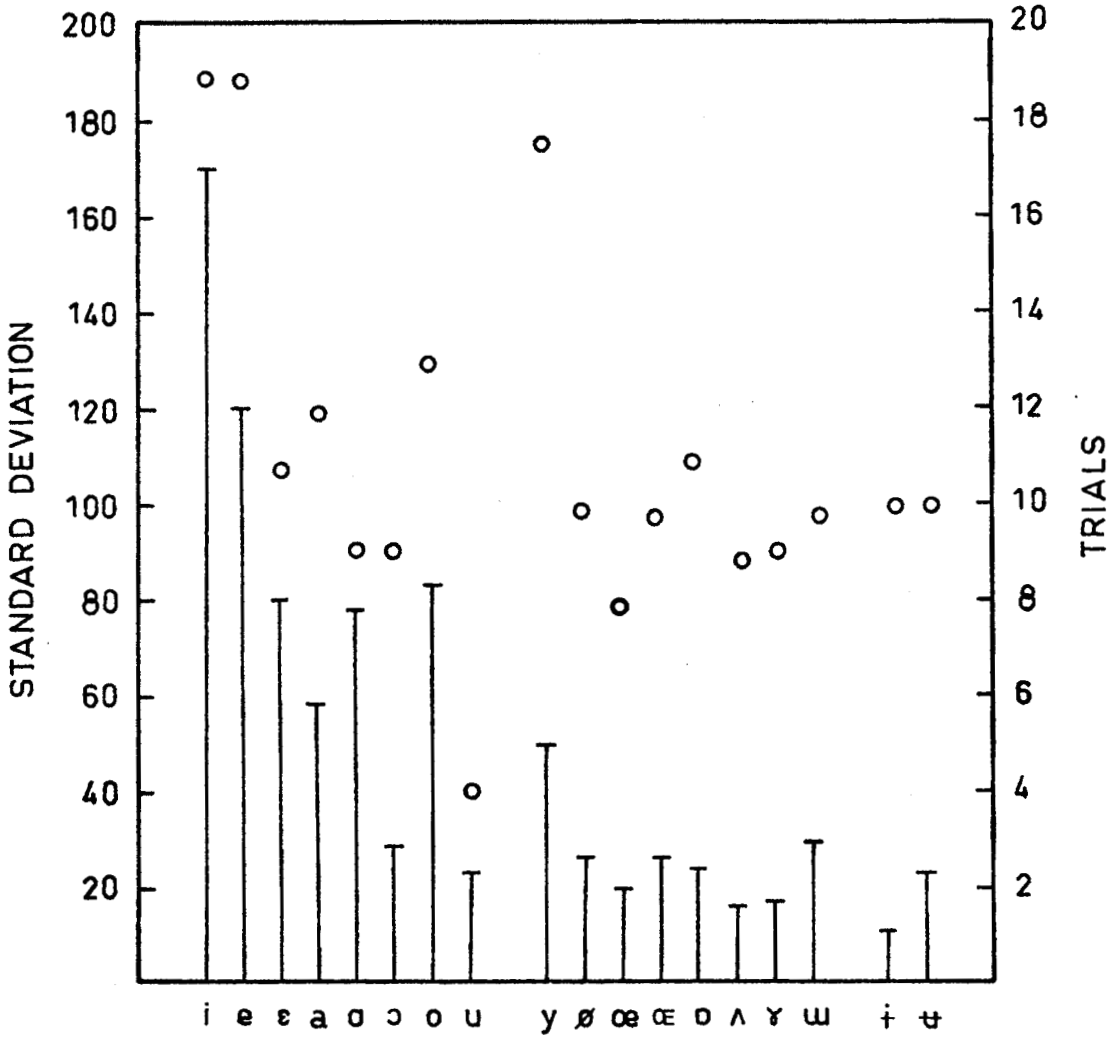

Fig. I-A-5. Standard deviation of F_2' matchings in mels (vertical lines) and the mean number of trials needed for a match (open circles).

a higher peak there is equal probability of either response. The F_2' formula is designed to retain the sensitivity in the change of percept when crossing an equal prominence boundary but it cannot be expected to display exactly the same location of the boundary as we encounter in any specific test. Moreover, the particular language and dialect of the subjects may bias the performance.

Acknowledgments

We are indebted to R. Carlson and B. Granström for discussions and for the use of their synthetic vowel matching program.

References

Carlson, R., Fant, G., and Granström, B. (1975): "Two-formant models, pitch and vowel perception", pp. 55-82 in Fant, G. and Tatham, M.A.A. (eds.), Auditory Analysis and Perception of Speech, Academic Press, London.

Carlson, R., Granström, B., and Fant, G. (1970): "Some studies concerning perception of isolated vowels", STL-QPSR 2-3/1970, pp. 19-35.

Fant, G. (1959): "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Technics, no. 1.

Fant, G. (1960): Acoustic Theory of Speech Production, Mouton, The Hague, p. 52.

Karnickaya, E.G., Mushnikov, V.N., Slepokurova, N.A., and Zhukov, S.Ja.(1975): "Auditory processing of steady-state vowels", pp. 37-53 in Fant, G. and Tatham, M.A.A. (eds.), Auditory Analysis and Perception of Speech, Academic Press, London.

Plomp, R. (1975): "Auditory analysis and timbre perception", pp. 7-22 in Fant, G. and Tatham, M.A.A. (eds.), Auditory Analysis and Perception of Speech, Academic Press, London.