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

Some studies concerning perception of isolated vowels

Carlson, R. and Granström, B. and Fant, G.

journal: STL-QPSR
volume: 11
number: 2-3
year: 1970
pages: 019-035

KTH Computer Science and Communication

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

II. SPEECH PERCEPTION

A. SOME STUDIES CONCERNING PERCEPTION OF ISOLATED VOWELS

R. Carlson, B. Granström, and G. Fant

Abstract

The purpose of these studies has been to examine the possibilities of describing the main phonetic quality of vowels in two dimensions, especially in terms of two frequencies. A mapping of the Swedish vowel space on this two-dimensional plane has been established. As means of investigation both a matching procedure and identification tests have been used.

The nature of the above projection has been studied, considering implications both to phonetic categorization and functional models. The feasibility of some technical solutions, including analysis by a cochlea model followed by zero-cross counting, will be discussed.

Introduction

A basic but yet not solved problem in acoustic phonetics is to express vowel quality in simple terms, preserving a unique relation to vowel identity. Normally, the vowel space is reduced to a plane showing the position of the first two formants, but this information is not sufficient to express vowel identity. The higher formants have considerable influence on the identity, especially for the close front vowels.

In the early fifties experiments performed at Bell Telephone Laboratories[1] and Haskins Laboratory[2] showed that a vowel can be synthesized using just two formants. The second formant, here called F2', should in that case substitute both the second formant and higher formants of a natural vowel, but is it possible to formalize this relation?

One specific formula, taking into account the first three formant frequencies, has been proposed by Fant[3]:

$$F_2' = F_2 + \frac{(F_3 - F_2)}{2} \cdot \frac{(F_2 - F_1)}{(F_3 - F_1)}$$

This formula, which was constructed to take into account F_3 when F_2 is far apart from F_1, was found to be useful for a two-dimensional mapping of Swedish vowels, improving the separation between unrounded and rounded

front vowels. It is, however, an ad hoc construction and is not claimed to represent a theory of perception. Fant and Risberg[4] performed a two-formant matching experiment involving the subject' s own prerecorded vowel as reference. From this study and the earlier American work it can be expected that a two-formant approximation will be more effective for back vowels than for front vowels and that F_2' matches F_2 of back vowels and is to be found in the region of F_3 of high unrounded front vowels.

One may argue whether formant frequencies is the best acoustic correlate to perceived vowel timbre and identity. It is well known that formant bandwidths and thus the exact formant amplitudes are not critical parameters whereas the overall spectrum shape as conditioned by the pattern of formant frequencies is an important stimulus aspect, Fant[5][6]. Accordingly, the location of F_1, F_2, and F_3 can be translated to the equivalent spectrum shape parameters. (These can be qualitatively studied in Fig. II-A-1.) Thus, F_1 affects the overall spectrum level, the main spectral balance is determined by F_2 which boosts the F_1-region and accounts for low level of F_3 and higher formants when close to F_1. This is the grave/acute dimension. When F_2 is close to F_3 there is a dominance of the upper part of the spectrum and we can establish a spectrum shape factor within the $F_2F_3F_4$-region by noting the location of F_3 closer to F2 (plain) or to F_4 (sharp). A movement of F_3 alone has several correlates, one is a shift of the center of gravity within the upper part of the spectrum, another is a change in spectral shape. The latter aspect would be perceived as a finer gradation than a two-formant perception model accounts for whereas the center of gravity would be a possible direct correlate of F_2'. The importance of the spectral shape aspect has been stressed by Fujimura[7] who rejected an equivalent two-formant interpretation of his three-formant vowel identification test. Experimental techniques for assessing such measures by means of a cochlea model and waveform zero-counting techniques have been attempted in our work.

The possible significance of various patterns within an F2F3F4F5-formant group has recently been discussed by Fant, Henningson, and Stålhammar[8].

The main object of our study is to learn more about vowel perception and the experiments with two-formant synthetic stimuli is merely a method to approach the more general problem. The following questions could be proposed.

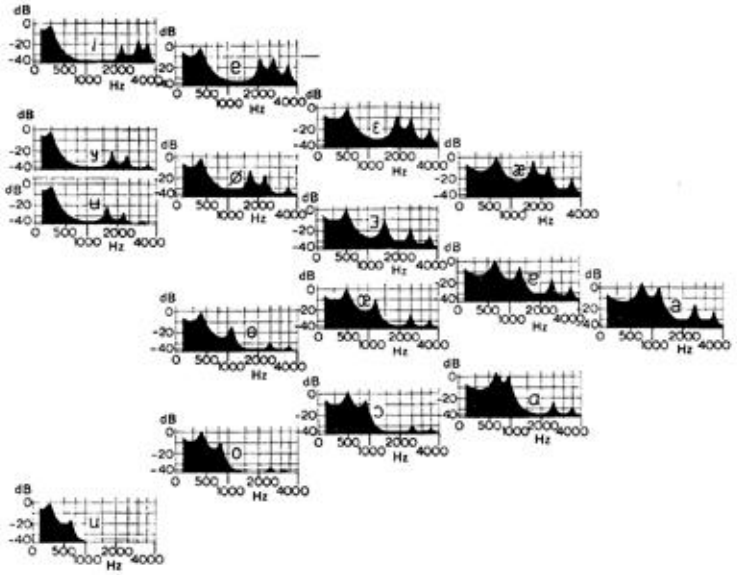

Fig. II-A-1. Spectra on an approximate mel scale of synthetic vowels ordered according to the particular F_1 and F_2. The changes in spectrum shape and in formant levels following a shift in one or more of the formant frequencies should be observed. (Fig. 8 in G. Fant: 'The acoustics of speech', Proc. 3rd ICA Stuttgart 1959, Vol. I.)

(1) How successful is a two-formant approximation of Swedish vowels?

(2) Can a subject match a natural or a high quality synthetic reference vowel against a synthetic vowel of a simplified structure without being influenced by an association to his own internalized inventory of vowel prototypes?

(3) To what extent can F_2' of the two-formant stimulus be predicted from the composition of the complete vowel?

(4) What is the neurophysiological and psychological significance of the two-formant model? Does it reflect basic constraints in the decoding process?

These questions indicate that the two-formant model cannot be studied isolated from the vowel perception process including e.g. parameter reduction, consistency and knowledge of language. A question which must be considered in this connection is: Does vowel perception behave in a continuous or categorical way in any sense? The literature gives several examples of contradictious answers. One alternative is proposed by Liberman et al[9] who claim that perception of synthetic vowels is continuous on the contrary to the perception of stops. On the other hand, Chistovich et al[10][11] have shown that vowel perception is discontinuous. A mixture of these opinions gives a third alternative, viz. the continuity is related to the time length of the stimulus, Fujisaki and Kawashima[12]. That means that if the perceptual identification process will force contradictory phonetical categories on the same vowel sound the corresponding F2' values will either be the same, in case of continuous perception, or different if vowel perception has a categorical component.

Pilot study

Here follows a brief report on a pilot study, the technique of which was later abandoned. The method may, however, have some interest per se. Simplified vowel stimuli were produced by substituting the upper formants (F2 and higher) of a natural vowel with the output of a narrow BP-filter (chosen from a filter bank) excited in synchrony with the fundamental frequency. The magnitude of the excitation was controlled by the energy of the upper formant area. These stimuli were presented alternatingly with the natural vowel and phonetically trained subjects were asked to give their opinion on the position of the "upper formant" of the semisynthetic stimulus being "too low, slightly lower, appropriate, slightly higher, too high".

The appropriate position of the "second formant" was found to be situated between the second and the third formant of the reference vowel for /ø/ and somewhat higher than the third formant for /y/, /e/, and /i/. The "second formant" value of /i/ was rather extreme (female 4.4 kHz and male 3.5 kHz). It should, however, be noted that the variance was quite low indicating that the subjects easily could make a decision. The stimulus quality was deteriorated by technical insufficiencies causing a decrease in similarity to the natural reference. Partially due to this the judgments tended to be made in a non speech mode. A need was thus felt to create a test situation more favorable to speech mode in which the test variables were more easily and accurately controlled.

Synthesis program for stimulus production

A computer program for vowel synthesis has been developed. To minimize the number of irrelevant variables both the "natural" reference vowel and the simplified vowel were generated by the computer, see Fig. II-A-2. This was done with the formant circuits arranged in parallel to facilitate the control of formant level parameters.* Two parameters of interest chosen among the formant frequencies, levels, and bandwidths, could be controlled with knobs connected to the computer.

To assure a reasonable degree of naturalness some pains were taken with generating a good intonation contour by varying the intensity, AO, and the fundamental frequency, FO, in an appropriate way.

Before the regular matching experiments could take place, we had to decide what kind of simplified vowel we should use, which parameters should be varied, and at which values the constant parameters should be fixed.

In order to settle this we carried out some preliminary investigations, the result of which might be of general interest.

As the second spectral peak we tried four alternatives, viz. a single resonance circuit (conjugate poles) with 50 Hz bandwidth, two cascaded resonance circuits the bandwidths of which were 50 Hz and 375 Hz, respectively, and two band-pass filters; one narrow (50 Hz) and the other broad (250 Hz).

* The initial level of each formant could be computed automatically by using distances in the complex plane according to a serial synthesis model. The contribution from higher formants was kept constant however, implying a constant vocal tract length (cf. in real speech there is a small difference in vocal tract length of rounded and unrounded articulation).

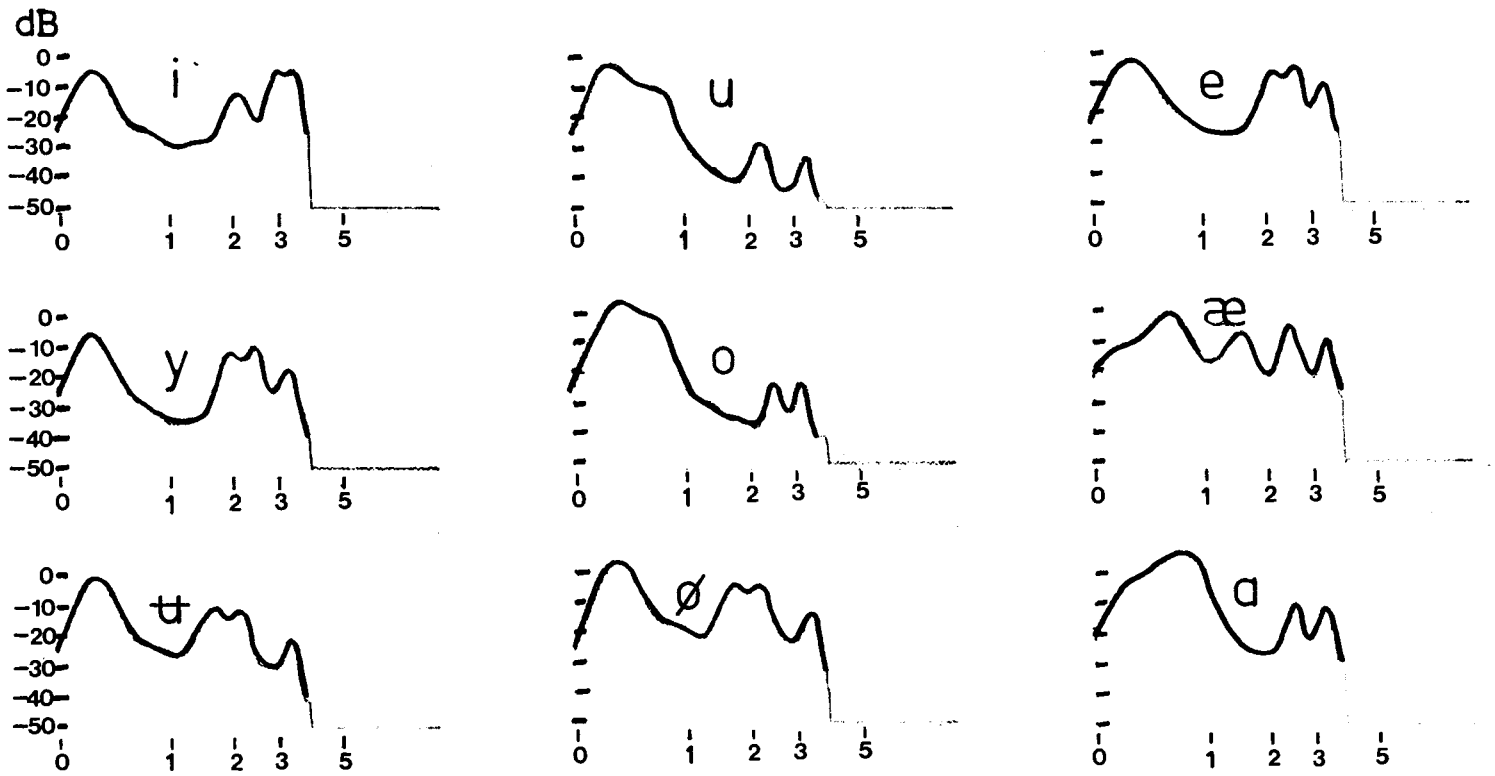

Fig. II-A-2. Spectra of synthetic reference vowels. Pre-emphasis +6 dB/oct.

Matchings were carried out using all these types. However, we could not find any significant difference in the results, as to mean matching frequency. There still seemed to be a difference in the spread of the results favoring the single resonance circuit as being the one with the least spread. We decided to use this circuit in the main experiments, especially as it is of the same kind as the ones generating the other formants.

The bandwidths appeared to be of minor importance. Hoping to add to the naturalness, we assigned frequency-dependent bandwidth values according to measured data, Fujimura and Lindqvist[13], taking into account that these measurements were made with closed glottis, giving too low values, especially for the low frequency region.

The formant level has often been claimed to be of importance in such a way that an increase in formant level has the same effect as an increase in formant frequency, e.g. Lindqvist and Pauli[14]. To get an estimate of the relative importance of this effect we set up a matching experiment with the frequency and the amplitude of the second peak, F2, as independent variables. The result is displayed in Fig. II-A-3. Every subject carried out two matchings for each vowel without time limitation. These pairs are connected in the graph. Out of these 24 lines (six subjects and four vowels) only 13 have a slope supporting the above-mentioned theory, i.e. a negative derivative. It is remarkable that all the /y/-pairs show a contradictory tendency. On the whole, the effect must be very weak and possible vowel dependent (cf. the relation between subjective pitch and loudness). This point, however, needs further investigation.

Thus, in our results we find justification for a rather crude rule of assigning formant levels to the two-formant vowel: in the "vocal tract" transmission function both peaks are equal in height to the first formant peak of the reference vowel, i.e. there will be a -6 dB/octave slope in the sound wave spectrum.

The main matching experiment

After having decided on the frequency position of the second peak in the two-formant synthesis as the sole manually controlled variable, we set up a matching experiment including all nine Swedish long vowels. The formant frequencies were taken from Fant's older measured data, Fant [3], as shown

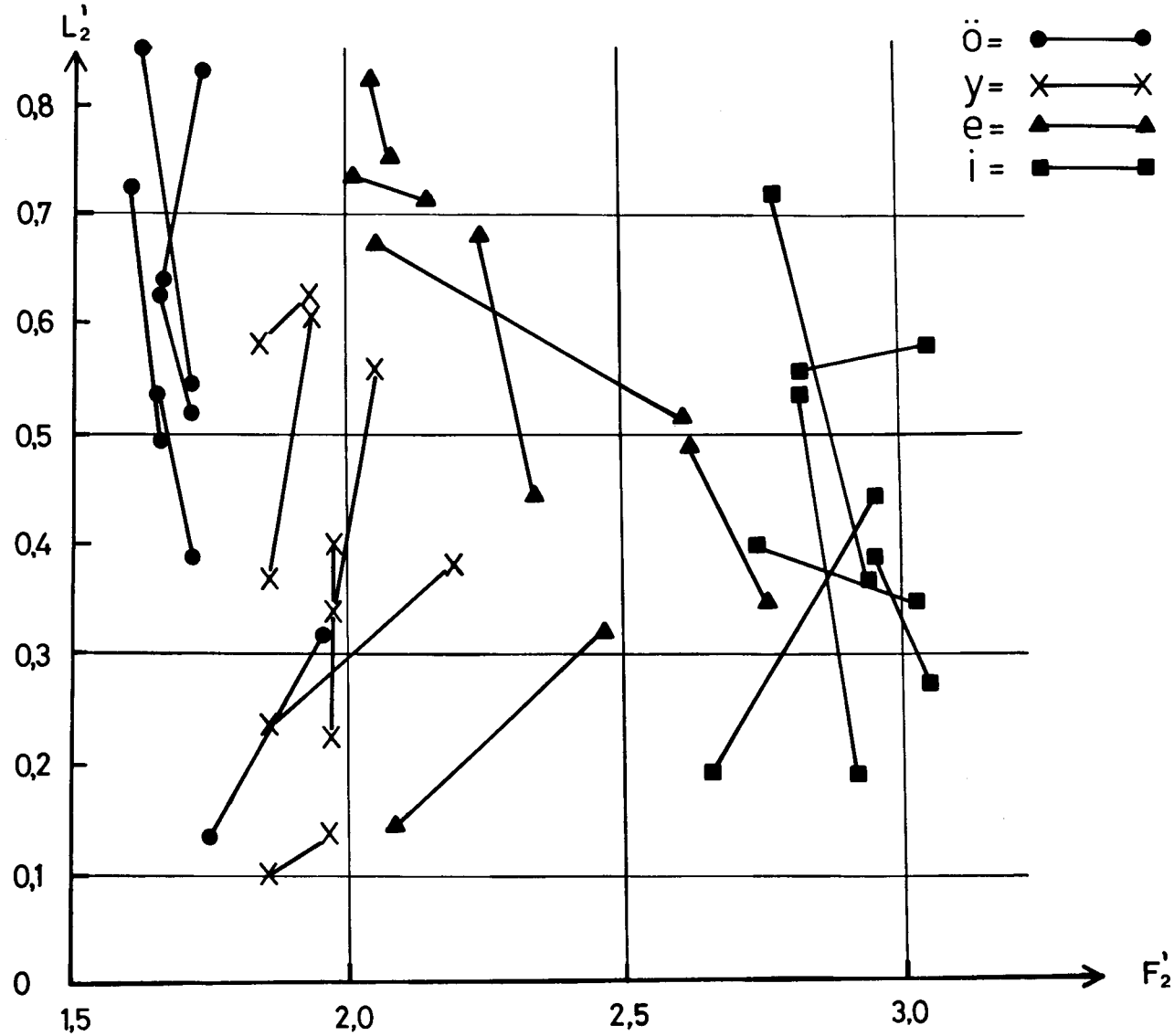

Fig. II-A-3. Result of a matching experiment. Reference vowels are presented in Table II-A-1 and Fig. II-A-2. The frequency and amplitude of F2 were chosen as independent variables.

in Table II-A-1. As representative of the "ä"-vowel we have chosen to use the pre-r allophone [æ:] being the one used when pronouncing the letter "ä" in isolation. The vowel sequence was randomized and controlled by the experimenter. The subjects were asked to make the two sounds as similar as possible, using the vowel quality as matching criteria. No time limit was put on the matchings. However, even though all the subjects approved of the quality of the synthesis almost everyone reported of a fatigue effect, i.e. they had difficulties in perceiving the sounds as vowels after very long matching attempts at one single vowel.

TABLE II-A-1

vowel	reference vowel				matched mean	$\frac{F2'_{max}-F2'_{min}}{F2'}$
IPA	F1	F2	F3	F4	F2'	
i	255	2065	2960	3400	3210	0.31
e	375	2060	2560	3400	2370	0.21
y	255	1930	2420	3300	2010	0.12
æ	605	1550	2450	3400	1960	0.23
ø	360	1690	2200	3390	1720	0.14
ʉ	280	1630	2140	3310	1730	0.14
ɑ	580	940	2480	3290	960	0.17
o	400	710	2460	3150	720	0.14
u	310	730	2250	3300	730	0.14

One typical test series included two subjects making five thorough matchings on each vowel. These matching attempts were spread during several days. The results from this test are shown in Table II-A-1 and Fig. II-A-4. The implication of these results will be discussed later. However, the spread of the results seem to be quite great, especially for the vowels having no neighboring phoneme with equal F1 but higher F2, i.e. [i], [e], and [æ]. It should, however, be noted that some subjects performed considerably more consistently, especially if the matchings were carried out on one single occasion involving only one vowel. In these favorable cases variation widths of about 3 % were observed, i.e. approximately equal to the DL for formant frequency.

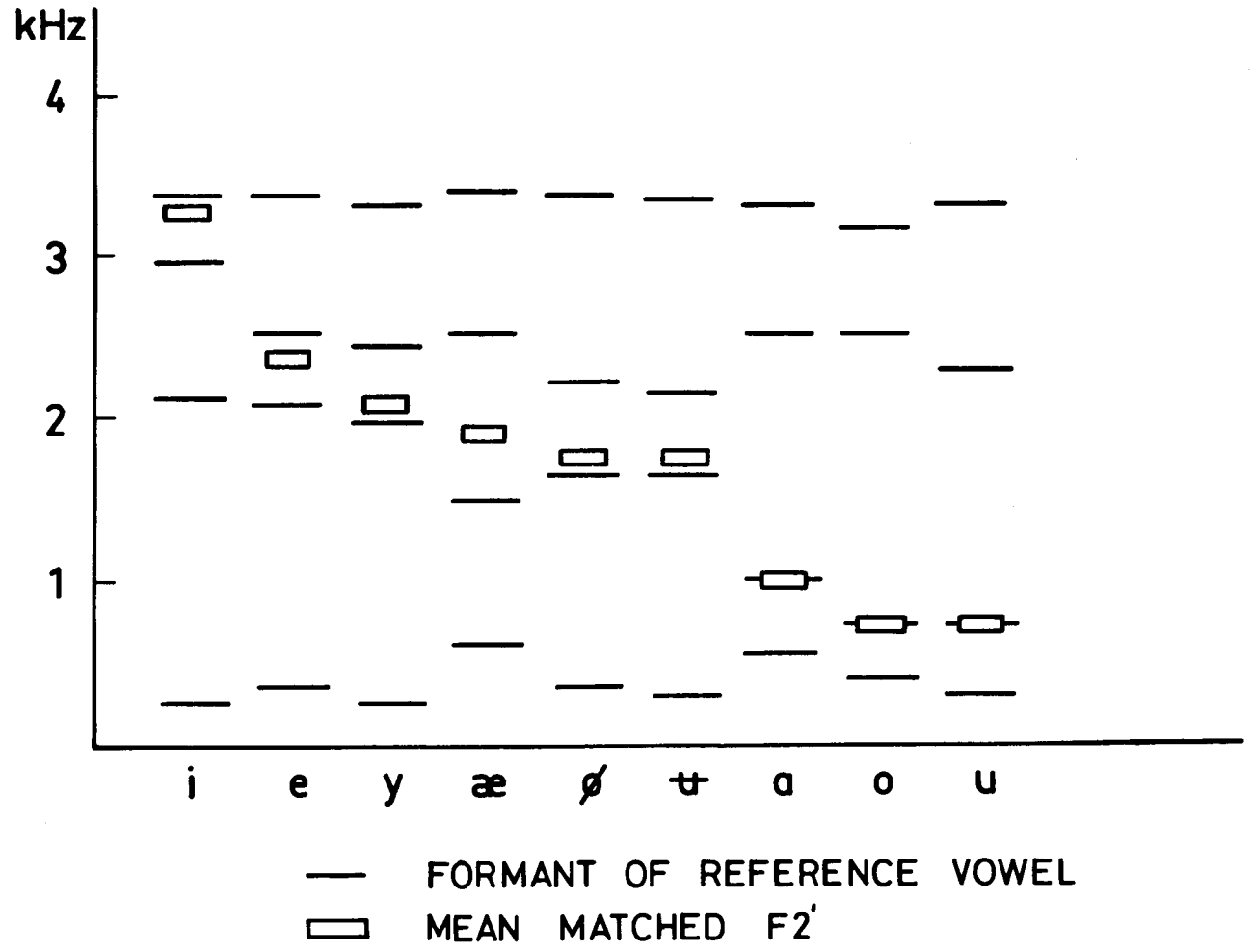

Fig. II-A-4. Result of a matching test.

In the matching situation the temporal and interindividual variations are quite great probably most due to varying matching criteria and unstable internalized references. These effects are interesting per se but it is not in the scope of this paper to dig any further into these problems.

Identification test. Background and realization

An important question always arises when dealing with a matching experiment. Is the result of the experiment representative of speech perception or is it just due to acoustic phenomena in a broader sense? In order to get some idea about that, it is logical to follow up with an identification test using the same type of stimuli as in the matching experiment.

Some preliminary identification tests were carried out and the experience gained was then used for the main test which was built up as follows. Two-formant stimuli were produced by the computer and recorded in four sequences. As we did not know where the true phoneme boundaries between the vowels were situated, we wanted the stimuli to be spread out in the plane so that every stimulus had the same perceptual distance to all neighbors. There is no doubt that the mel scale would serve this purpose better than the linear scale. We choose to rotate the F1-F2'-plane 45° in order to relate the test to the "spread" versus "flat" parameters adopted by Fant[15]. After this operation was done we obtained two new mel scale parameters, here called m and n, substituting the formant locations. These can be derived as follows:

$$M2 - M1 = m \cdot 135$$
$$M2 + M1 = (n + 6) \cdot 125$$

M1 and M2 are F1 and F2 expressed in the technical mel scale, e.g.,

$$M1 = ({}^{2}\log(1 + F1/1000))\ 1000$$

Each stimulus in the test could be identified by m and n, here defined as integers. A series consisted of 96 stimuli in six groups arranged in a special sequence, which will be discussed later. Two series had the same sequence but different pitch and running the sequence backwards gave the two others. The series using the same pitch contour were played one after another.

The test group consisted of 27 electrical engineering students from different parts of Sweden speaking various dialects.

The test was run in an auditorium where the subjects were placed in different parts of the room. Before the first and the third series started the test group listened to a random sequence of two-formant stimuli in order to get used to the specific quality.

The subjects were instructed to respond in the normal orthography of the nine long Swedish vowels. It should be noted that the vowel signs of the Swedish orthography have fairly stable pronunciations, though sometimes distinct from the pronunciation of the similar IPA-characters. To avoid confusion the response alternatives will henceforth be denoted by majuscules. In Table II-A-2 these are shown along with the approximate IPA-transcript of the letters pronounced in isolation. It is possible to introduce a greater number of natural response categories. (This will be touched upon in the discussion.) However, the gain was judged to be less than the loss, since the subjects were unfamiliar with this kind of test and the confusion would probably have increased considerably.

TABLE II-A-2

TEST CATEGORY	I	Y	U	O	Å	Ö	E	Ä	A
PHONETIC VALUE IN ISOLATION IPA	i	y	ʉ	u	o	ø	e	æ	ɑ

There are some observations on the result, discussed below, that might throw some light on the subjects' behavior and, consequently, also on the validity of the conclusions drawn. First of all no correlation could be found between a subject's location in the room and his answers. Only 0.15 % of the responses were labeled "no opinion", which indicates that the subjects had no difficulties in following the test.

The stimulus sequence was built up by triads, consisting of one long jump in the plane (about 600 mel) and one short (about 200 mel) in the opposite direction, i.e. when running the sequence in one direction the contrast between the second stimulus in the triad and the first is great, and when running the test sequence backwards the first step is small. The triads were overlapping so that the last stimulus in one triad was the first in the following. By this method the sequential effect on the result could be studied. However, no significant influence was found. The relatively long interval between the stimuli (2.5 sec) might account for this result.

An alternative explanation would be that both the long and the short jump tend to move the response in the same direction. In the case of a short jump the two stimuli could be perceived as different sounds due to the proximity in time, impelling the subject to label them differently even if they in another context should be given the same identity. The long jump might, on the other hand, be perceived as intended longer in the same way as in natural speech, where an extreme change in articulation seldom results in the intended target values.

A second aspect to be considered is the relation between the subject's classificatory reference and his production.

In order to study this relation, formant frequencies of the listeners' spoken vowels were extracted and the mean frequencies for different vowels were calculated. It was now possible to examine if a subject had a lower or a higher position of a formant than the mean. The test result was examined in the same way and no correlation between this result and the subject's own production could be found, discouraging any belief in extreme motor theory. It should be noted that the average vowel position in the test was calculated as means of the positions of the stimuli, which had been given the same judgment by the subject. This point has significance only in relation to the total mean, and is not to be regarded as an ideal vowel position.

Discussion of the identification test

Comparing the results of the matching experiment, Fig. II-A-4, and the identification test with low F_0, Fig. II-A-5, it can be seen that the matched F2 falls within the right phoneme area, or at least on the border, obtained in the identification test. This conforms well with the phonetic categorization as an important part of the matching procedure which will be discussed later. The reference formant data have been revised recently, Fant et al[8]. The border-line cases had possibly disappeared if these new data had been used. It is obvious, however, that the old data evoked the appropriate response as to vowel identity.

The response areas in Figs. II-A-5 and II-A-6 for the maximally closed vowels [i], [y], [ʉ], and [u], especially [u], are rather small, considering the spread of formant data in natural speech. This is certainly due to the commonly diphthongized realization of these phonemes in Swedish. Thus, [i] and [y] are often palatalized and [ʉ] and [u] labialized towards the end,

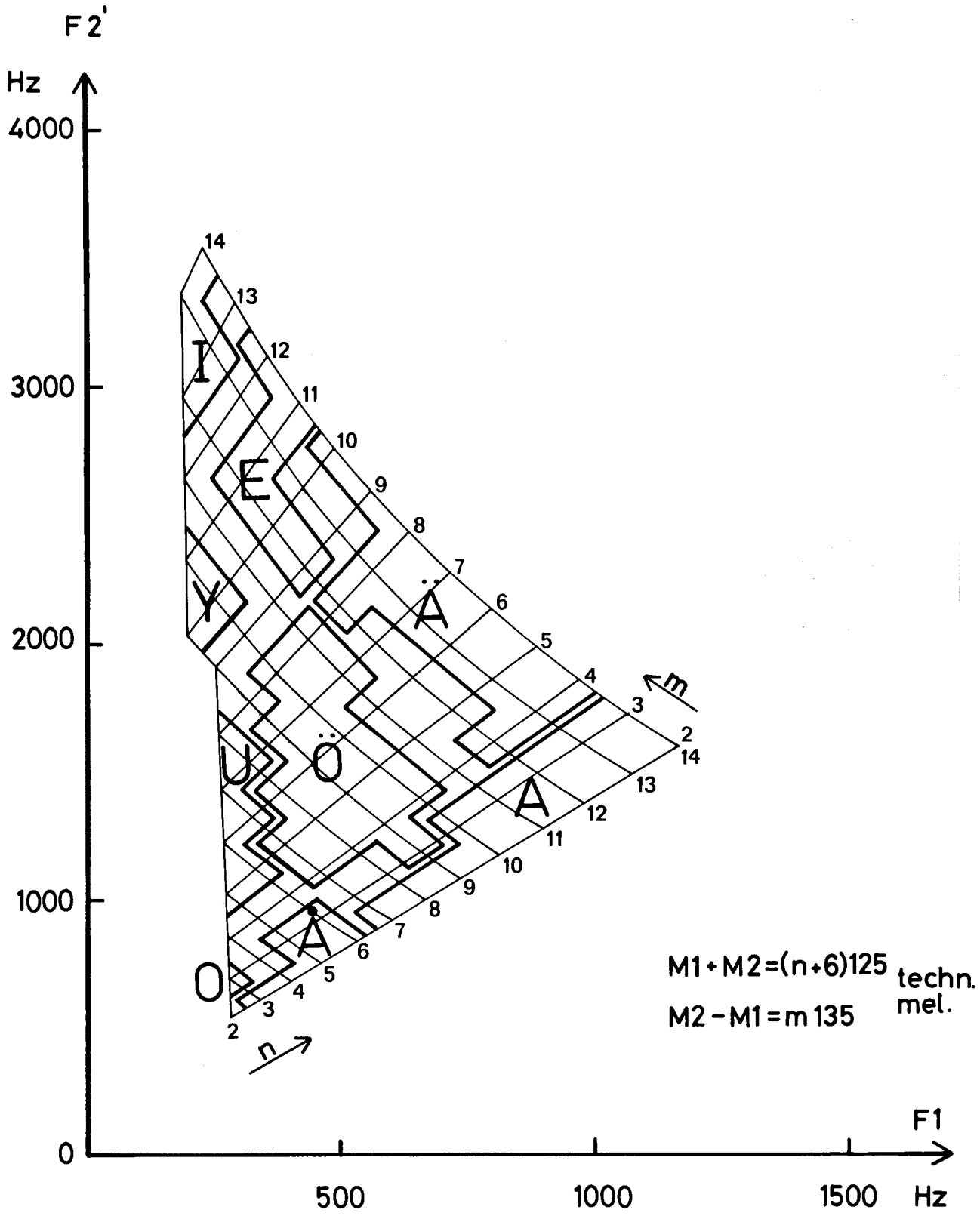

Fig. II-A-5. Response areas including points where 75 % of the responses in the identification test are equal. Low pitch (100-120 Hz).

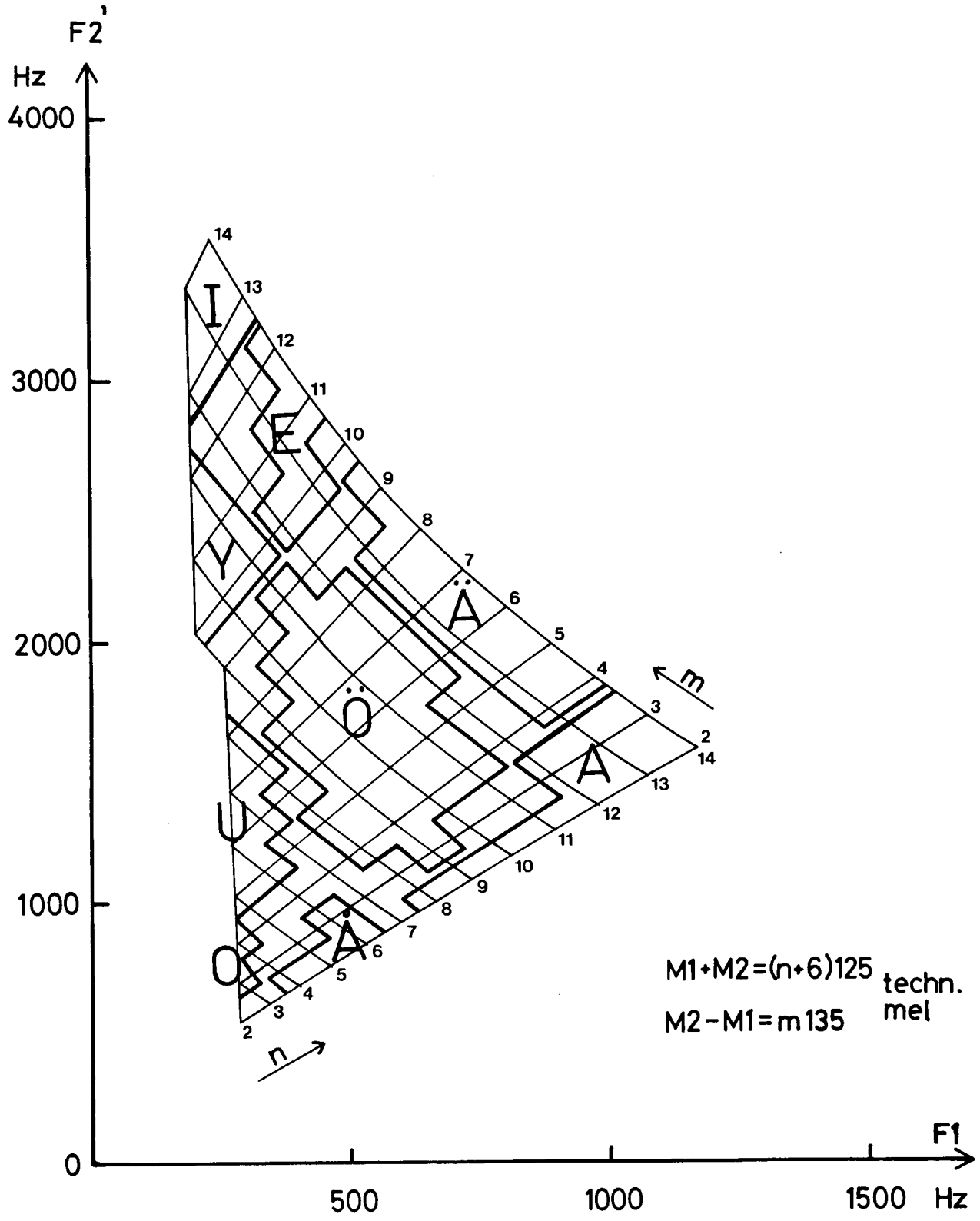

Fig. II-A-6. Response areas including points where 75 % of the responses in the identification test are equal. High pitch (200-240 Hz).

approximately corresponding to a move in the "+m" and "-n" direction, respectively (see Fig. II-A-5). This diphthongization was not present in the test stimuli, but an informal listening test carried out with accordingly diphthongized stimuli shows an expected broadening of the response areas in addition to increased naturalness hinting that these specific diphthongizations are important secondary cues for the identification of [i], [y], [ʉ], and [u] in Swedish.

All the vowels, except the back vowels [u], [o], and [ɑ], have an F2' more or less distinct from F2. This observation seems to be important for the discussion of invariance since the front vowels have a considerably greater male/female F2 difference than the group [u], [o], [ɑ], Fant[3][16]. In other words, when F2 dominates perceptually over higher formants there is a demand on speakers to maintain a similarity in F2, whilst other parameters mainly serve as cues for differences in vocal tract length and the sociologically established distinction between male and female voices.

In spite of the fact that the stimuli were deprived of sex qualities other than F_0, a comparison of the result with high and low F0 exposes a marked shift in phoneme location, as shown in Fig. II-A-7. This shift, 75 mel as an average, follows approximately the "+n", i.e. the "flatness" direction. Both facts are in accordance with earlier findings, Fant[3] and Miller[1]. In this case, however, the effect is induced by F0 alone. The quantitative data must be adopted with some precaution since the result may partially be due to the quantized stimulus locations and the specific stimulus ensemble used in the test.

How do our data conform with the theory of maximal contrast, i.e. equal spacing of categories in a perceptual plane, NB if such a place exists, Fant[15].

As mentioned earlier, it is possible to split some of the response categories that are phonetically relevant in Swedish. The criteria must be that the splittings rely on distinctions in timbre felt by normal language users. The long/short distinction relies primarily on the temporal pattern of speech, but in the case of /A/, /Å/, and /U/ a clear timbre difference exists. /Ä/ and /Ö/ contain the pre-r allophones [æ] and [œ], respectively. Thus, we find justifications in splitting /A/ into [ɑ] and [a], /Å/ into [o] and [ɔ], /U/ into [ʉ] and [ə], /Ä/ into [ɛ] and [æ], and finally, /Ö/ into [ø] and [œ]. An attempt has been made to map these additional categories

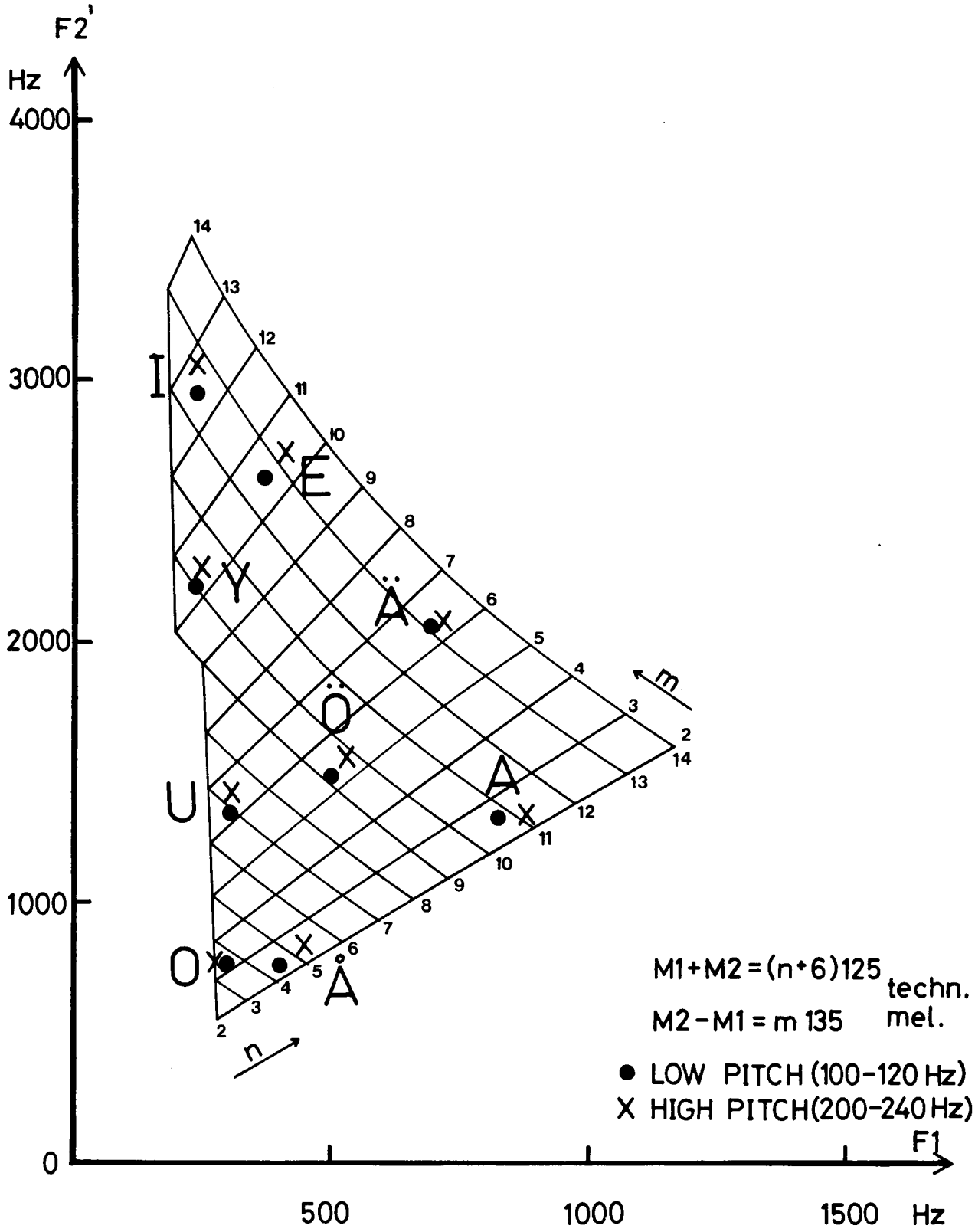

Fig. II-A-7. Mean of the responses in the identification test.

in Fig. II-A-8, where the nearest IPA-symbols have been used. In the (m, n) plot, the division appears to be more uniform than in the orthogonal (F1, F2) plot. The boundaries tend to follow the "m" or "n" directions, i.e. the spread and flat dimensions hint the perceptual significance of these parameters. The relations between vowels can be discussed in terms of quantal steps of the order of 250 mel, see Fant[15]. The two-formant stimulus corresponding to the /a/ and /ɑ/ areas are quite extreme and these locations should be moved in the "-n" direction. An identification test with an expanded number of response alternatives has been planned.

Further experiments on parameter reduction

To gain insight into the mechanisms of reduction of the full vowel to a two-formant sound, we carried out some additional experiments. We regarded the [y]-[i] region as suitable for investigation, since the F2' is clearly distinct from F2 and furthermore the [y]-[i] distinction is sometimes signaled by one single formant, namely F3. It has been proposed that the perceptually defined F2 could be related to the acoustic signal as some sort of weighted mean of the upper formants. With F3 varying between the typical [y] and [i] position and all other formants fixed a number of matchings were carried out, the result of which can be seen in Fig. II-A-9. Two representative spectra of the synthetic reference vowel has been added in Fig. II-A-9, however, they should not be regarded as ideal vowel synthesis of /y/ and /i/, cf. Fig. II-A-1 and Fig. II-A-2.

The distribution of points seems to be bimodal and the slope is remarkably large. A shift of F_3 from 2500 Hz to 2800 Hz requires a shift of F_2' from 2400 to 3300, i.e. $\Delta F_2'/\Delta F_3 = 3$. Can this shift be predicted from some weighted mean of F_2 F_3 and F_4? Any attempt to formulize a perceptual mean F_{2e} would have to rely on rather arbitrary assumptions since we lack knowledge of the relevant auditory weighting mechanisms at the cortical level. From an impressionistic point of view the "center of gravity" F_{2e} of the $F_2 = 2000$, $F_3 = 2500$, $F_4 = 3350$ region would not be too far from the matched $F_2' = 2400$ Hz whereas it seems less likely that F_{2e} of the $F_3 = 2800$ sample would fall as high as 3300 Hz, i.e. close to F_4 which is at the extreme high end of the vowel spectrum. The linear models we have tried provide a $\Delta F_{2e}/\Delta F_3$ of the order of 1 but could be expanded by squaring the weighting function. More sophisticated models retaining the general notion of F2 being enhanced at low F_3 and F4 being enhanced at high F_3 could of course be tried.

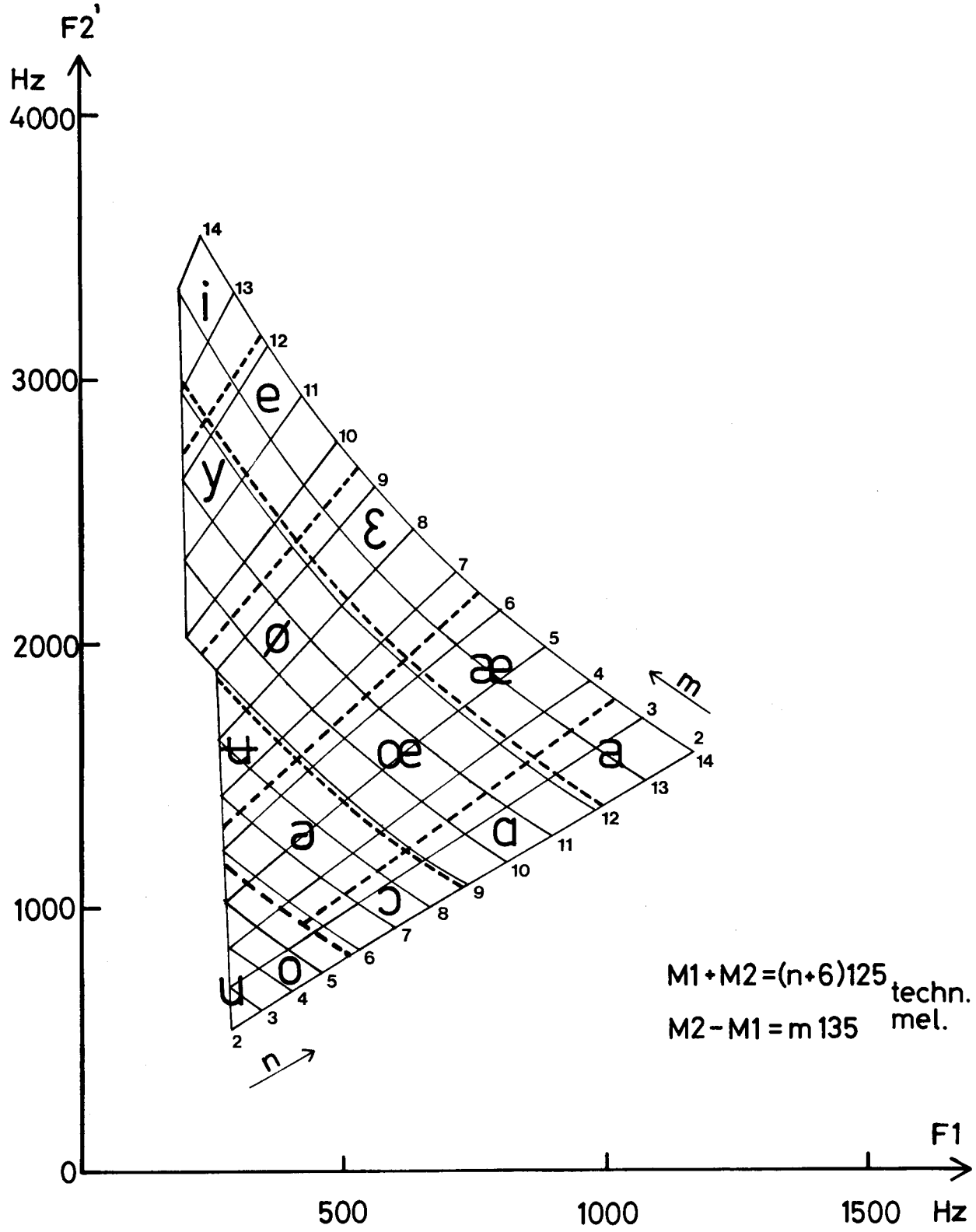

Fig. II-A-8. A hypothetical phonetic display explained in the text.

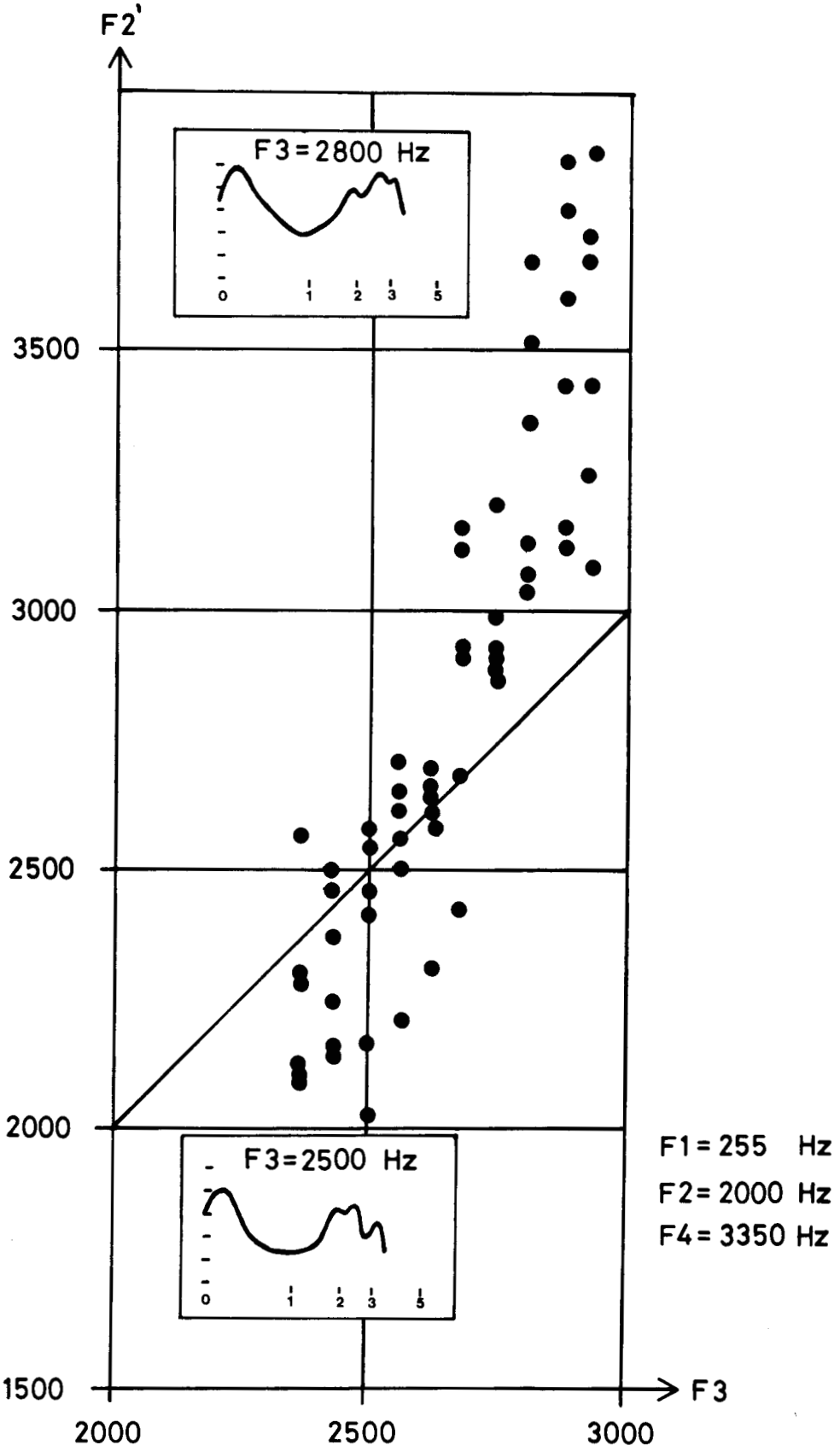

Fig. II-A-9. Result of a matching experiment. F3 of the reference vowel varying from an [y] to an [i] position.

It does not seem likely that F_2' of the two-formant stimulus is matched to one single formant of the complete vowel.

A more probable explanation appears to be that the subjects categorical identification of the F_3 = 2800 Hz reference as an [i]-vowel requires a match to an internalized vowel of higher F_{2e} than that of the F_3 = 2800 Hz sample. Also any F_2' above 3000 Hz would be identified as an [i]-vowel. We accordingly have to assume that the F_2' -matching process involves both continuous and discrete evaluations.

In order to follow up these results we arranged an identification test with the same type of four-formant stimuli, though with systematically varied formant levels, see Fig. II-A-10. The parallel synthesis is not the ideal technique if just a change of a formant level is wanted, since the level is changed in the domain between the formant peaks, too, and followingly the total envelope will be influenced.

The only conclusion that could be drawn concerning level dependency is that the uncertainty increases with extreme formant levels. However, the 50 % point fits rather well as delimiter between the two clusters in the matching result.

Considering the above-mentioned results it appears difficult to extract from the four-formant vowel a frequency equal to the matched F2', by any formalism, before the vowel has been identified. But is it still possible to compute a mean frequency F_{2e} close to F2' or at least within the right phoneme area defined by the identification test?

We have tried two experimental methods to attack this problem. One is to use some sort of time analysis, e.g. zero-cross counting, the frequency domain equivalence of which is a center of gravity (exact relation holds for band-pass filtered random noise only). Another method is to build up a filterbank and study the energy in different bands or some sort of envelope. We started with a rather simple one: high-passing (0.970 kHz 36 dB/oct) the signal and counting the number of zero-crossings during a certain time (100 msec). This number was used to derive a representative frequency, which was compared with the mean of the matchings, see Table II-A-3. The 100-msec integration time was selected in accordance with the work of Zwicker and Feldtkeller[17], who demonstrate changes in psychoacoustical behavior for stimuli shorter than 100 msec. The F1 is

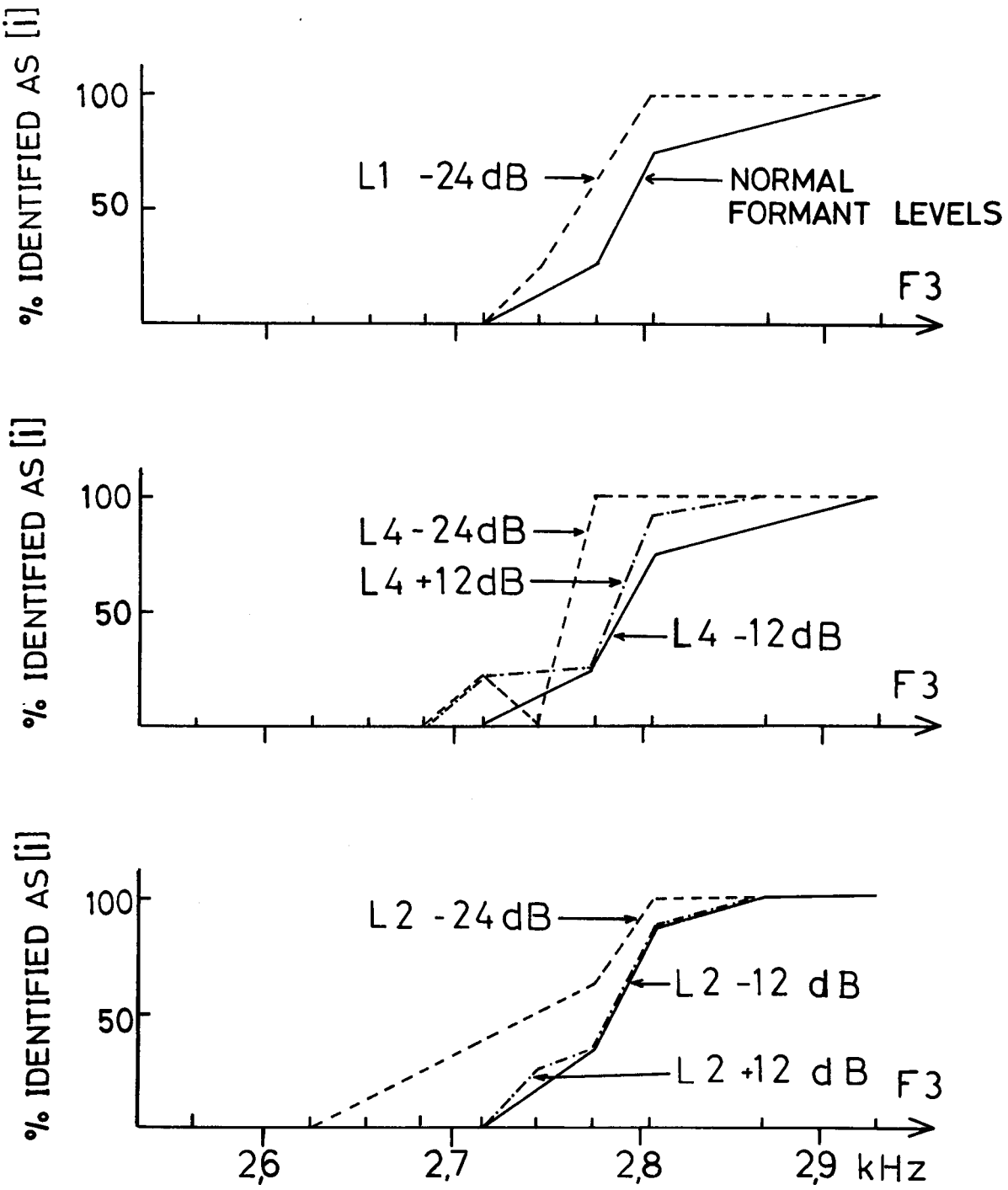

Fig. II-A-10. Results of identification tests with varying formant levels.

separated from higher formants by the filter. The sampling rate used corresponds to a low-pass limit at 5 kHz. The correspondence between the four-formant stimulus and the two-formant stimulus (the mean of the matchings) is fairly good, but the separation between the four-formant synthesis of [y] and [i] is not satisfactory.

TABLE II-A-3

vowel IPA	matched mean F2' kHz	Model for extraction of F2'						
		HP and zero-cross count. kHz	Model of the basilar membrane					
			Amplitude peaks kHz			Zero-cross count. Histogram peak kHz		
i	3.21	2.86	0.2	3.1		0.3	3.1	
I	2.37	2.38	0.4	2.5	3.3	0.3	2.3	
y	2.01	2.38	0.2	2.1	2.7	0.3	2.1	
æ	1.96	1.95	0.7	1.9	3.0	0.5	1.9	
ø	1.72	1.89	0.4	1.9	2.3	0.4	1.7	
ʉ	1.73	1.72	0.2	1.8		0.3	1.6	
ɑ	0.96		0.5	0.9		0.5	0.9	
o	0.72		0.4	0.6		0.4	0.7	
u	0.73		0.3	0.6		0.4	0.7	

For further analysis a model of the basilar membrane was constructed. The design of this model was based on the work of Flanagan[18]. The output gave the amplitude distribution along the membrane at 120 points. The maximum amplitude during every 100 msec and in every channel was stored and the envelope of the relative movements of the basilar membrane could be examined, see Fig. II-A-11 and Table II-A-3. In this table the localization of the first two or three peaks in the envelope of the four-formant synthesis is shown. The result seems promising in relation to the matching but there are at least five discouraging problems:

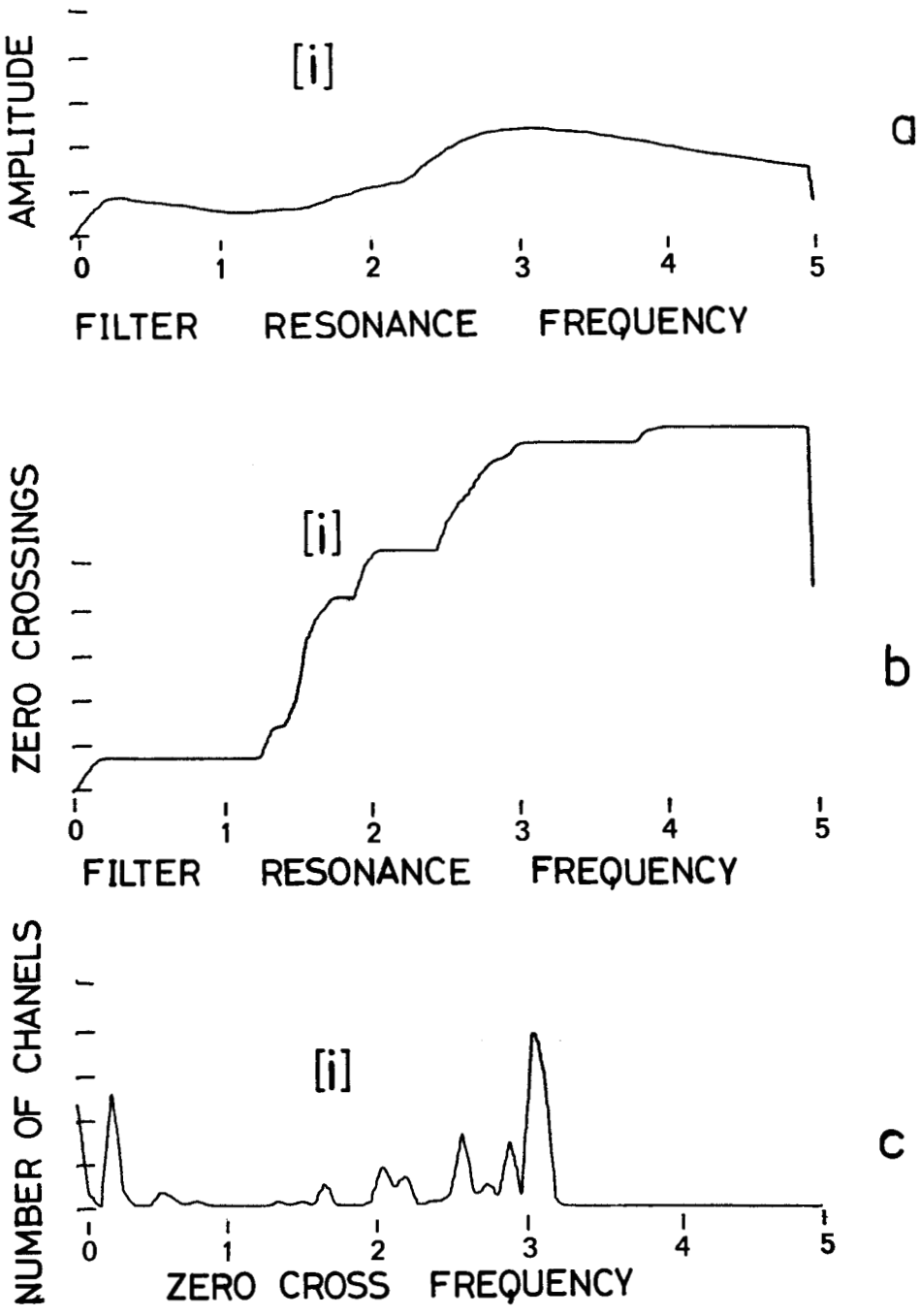

Fig. II-A-11. Output from the ear model described in the text.
a) Amplitude envelope on the basilar membrane.
b) Zero cross distribution along the basilar membrane.
c) Histogram. Zero cross frequencies are grouped in 75 Hz intervals.

(1) In [y], [e], and [æ] the third peak sometimes has a higher level than the second.

(2) The peaks in [ʉ], [u], [o], and [ø] are quite small in comparison with the environment.

(3) The peak in [ø] seems to fluctuate but has the printed values as mean.

(4) The first peak in [ɑ] is seldom possible to discriminate.

(5) Tests with natural speech show as much speaker dependency as normal spectrum analysis.

Since the result of the last two experiments was rather promising, but not satisfactory, an attempt was made to combine the methods. The ear model was used as a filterbank with human relation and a zero-cross counter was connected to every output channel. This model should only be regarded as functional. The result was displayed as seen in Fig. II-A-11. For every vowel sound, both synthetic and natural, there existed frequency intervals where the output from different channels gave the same representative frequency. In the histogram, which shows how many channels gave the same value, two or three isolated peaks always existed, see Fig. II-A-11. The location of the highest two peaks is presented in Table II-A-3. The figures are representative for the stationary part of the sound. The fluctuation at the beginning and the end of the sound was great in spite of the fact that the synthetic sound was stationary. The noise has, of course, a great influence on the result in these parts since the SNR was low and the zero was defined as zero and not as an interval between two limits. The difference between [y] and [i] is now improved and the position of the second peak is related to the matching result. But does this relation have anything to do with perception and human behavior in a matching situation? Some small experiments with natural vowels gave discouraging results and it was not possible to discriminate different vowels by simple decisions.

A principal difference in vowel and consonant identification has often been proposed, some authors being of the opinion that the vowel analysis should take place at a low level in the auditory system. Some of our results may be interpreted in that way. In order to test such a hypothesis we made a modification of the synthesis program to be able to split the vowel sound between formants. This was done by connecting the outputs of the formant circuits to two different channels. Thus, the different channels could supply different ears or both ears could be exposed to the whole signal. The vowel identity proved to be invariant to such changes, leaving space impressions and minor timbre changes as distinguishing cues.

Comments on parameter reduction

Our experiments suggest that the frequency value of F2' in the matching experiment is related both to the perceived identity of the reference vowel and to its timbre, and furthermore the formant frequencies of a natural vowel have much greater influence than the formant levels on the identity. Thus, a prediction of F2' presumes identification of the reference vowel. Dialectal and language dependent variability may accordingly be expected.

Our experiment with the formants divided on the two ears also indicates that the timbre is established at an auditory level above that of central summation. From a speech perception point of view this level can still be subordinate to that of phonetic categorization. This means that the position of the second formant in the two-formant synthesis mentioned above cannot be extracted by looking for some sort of peak amplitude in the cochlea. Instead, information from both ears must be transmitted to a higher level, probably in many parallel lines, before the parameters used for the decision rules could be derived. In our last model the mechanical filtering in the cochlea is followed by a time domain analysis. Is this possible in the human nervous system at the frequencies we have used? The time domain analysis may provoke objections since the limit of synchronous excitation mostly is settled at 1.5 kHz. This limit, however, does not seem to be conclusively established. The many parallel lines, without essential information reduction, constitute no objection to adapt the model as a part of the speech mode analysis. In the auditory system this parallelism gives no additional cost contrary to the case in machine recognition.

Summary and conclusions

The objective of our work has been to study the feasibility of two-parameter models of vowel perception by experiments involving the matching of two-formant synthetic vowels with four-formant vowels, identification tests, and various means of signal transformation and reduction. In a broader sense these studies are intended to provide some general insight in the perception of steady-state vowels. Our findings can be summarized as follows:

(1) All Swedish vowels may be synthesized at a first order of approximation by two formants only, as revealed by identification test. Identification results are not correlated with the subject's individual vowel production.

(2) A more detailed representation of a vowel spectrum than by two formants is probably of greater importance for perceived timbre than for the identity.

(3) The matching of a two-formant synthetic vowel against a four-formant reference appears to be mediated by the subject's internalized inventory of vowel prototypes, by external biasing and by other matching criteria which are not always stable.

(4) Since matching involves both continuous and categorical evaluations which are not fully known there can be only a limited success in predicting matching results from the measurable aspects of the complete vowel.

(5) Experiments with dichotic split of the vowel sound suggests that the continuous aspects of the vowel are carried to the cortical level above that of central summation.

(6) F_2', the second formant of the two-formant stimulus, matches F_2 for back vowels, a location intermediate between F_2 and F_3 for open front vowels, and a location close to F_3 or higher for high unrounded front vowels. These results conform with earlier investigations.

(7) The F_2' of front vowels conforms reasonably well with the rate of zero-crossings of the four-formant reference vowel high-pass filtered above F_1 or by measuring the second peak of the envelope in a cochlea model. A combination of band-pass filtering and zero-crossing counting appears to offer an interesting alternative with even better fit to F_2'. Although such methods may be useful in identification they cannot be accepted as methods of the perceptual process. The sensitivity of such measures to formant level changes does not conform with perception. Further research is needed to investigate alternative means of parameter reduction of normal vowels. The normal relation between formant frequencies and spectrum envelope shape factors offers a guiding principle for this search.

(8) The trading relation between F_0 and F_1 and F_2' requires on the average a shift of the -flatness parameter M_1+M_2' by 75 mel for a change in F_0 from 110 to 220 Hz. This is of the order of 1/4 of a minimal phonetic step.

(9) A mel scale representation M_1 versus M_2 of the identified vowels brings out certain regularities and equidistant spacings in the vowel system with minimum quantal steps of the order of 250 mel, see Fant[15].

References:

(1) Miller, R.L.: "Auditory tests with synthetic vowels", J.Acoust.Soc. Am. 25 (1953), pp. 114-121.

(2) Delattre, P.C., Liberman, A.M., and Cooper, F.S.: "Two-formant synthetic vowels and cardinal vowels", Le Maître Phonétique, July-Dec. (1951).

(3) Fant, G.: "Acoustic analysis and synthesis of speech with applications to Swedish", Ericsson Technics 15, No. 1 (1959).

(4) Fant, G. and Risberg, A.: "Auditory matching of vowels with two formant synthetic sounds", STL-QPSR 4/1963, pp. 7-11.

(5) Fant, G.: "On the predictability of formant levels and spectrum envelopes from formant frequencies", in For Roman Jakobson, pp. 109-120 ('s-Gravenhage 1956).

(6) Fant, G.: Acoustic Theory of Speech Production ('s-Gravenhage 1960).

(7) Fujimura, O.: "On the second spectral peak of front vowels: A perceptual study of the role of the second and third formants", Language and Speech 10 (1967), pp. 181-193.

(8) Fant, G., Henningsson, G., and Stålhammar, U.: "Formant frequencies of Swedish vowels", STL-QPSR 4/1969, pp. 26-31.

(9) Liberman, A.M., Cooper, F.S., Harris, K.S., and MacNeilage, P.F.: "A motor theory of speech perception", paper D3 in Proc. of the Speech Communication Seminar, Stockholm 1962, Vol. II (Stockholm 1963).

(10) Chistovich, L., Fant, G., de Serpa-Leitão, A., and Tjernlund, P.: "mimicking of synthetic vowels", STL-QPSR 2/1966, pp. 1-18.

(11) Chistovich, L., Fant, G., and de Serpa-Leitão, A.: "Mimicking and perception of synthetic vowels", STL-QPSR 3/1966, pp. 1-3.

(12) Fujisaki, H. and Kawashima, T.: "On the modes and mechanisms of speech perception", Annual Report No. 1, July 1968-June 1969, Engineering Research Institute, University of Tokyo, pp. 67-73.

(13) Fujimura, O. and Lindqvist, J.: "On the sinewave response on the vocal tract", STL-QPSR 1/1964, pp. 5-10.

(14) Lindqvist, J. and Pauli, S.: "The role of relative spectrum levels in vowel perception", STL-QPSR 2-3/1968, pp. 12-15.

(15) Fant, G.: "Distinctive features and phonetic dimensions", STL-QPSR 2-3/1969, pp. 1-18.

(16) Fant, G.: "A note on vocal tract size factors and non-uniform F-pattern scalings", STL-QPSR 4/1966, pp. 22-30.

(17) Zwicker, E. and Feldtkeller, R.: Das Ohr als Nachrichtenempfänger, 2nd revised edition (Stuttgart 1967).

(18) Flanagan, J.L.: "Computational models for ear operation", in Speech Analysis Synthesis and Perception, pp. 91-118 (Berlin 1965).