Two-formant Models, Pitch and Vowel Perception

Rolf Carlson, Gunnar Fant and Björn Granström

Department of Speech Communication,
Royal Institute of Technology (KTH),
S-100 44 Stockholm 70, Sweden

Introduction

In 1970 we reported on a set of experiments on vowel perception based on two-formant approximations to four-formant synthetic vowels (Carlson *et al.*, 1970). It was concluded that all Swedish vowels could be matched by two-formant approximations, and that the effective formant 2, F'_2 was placed close to F_2 in back and midvowels, inbetween F_2 and F_3 in non-high or rounded front vowels, and in the region of F_3 or higher for a typical [i:] vowel, see Fig.1. The results of this match could be rather closely predicted from cochlea analog filtering by a measure of the density of channels carrying the same output zero-crossing frequency within a given quantal interval. It was found that vowel identity was retained when presenting one or more formants to one ear and the remaining formants of the sound in the other ear -

indicating an integration of timbre at a non-peripheral level of the auditory system. Experiments on identification of Swedish two-formant synthetic vowels revealed a dependency of F_0 of a magnitude expected from earlier studies, Miller (1953) and Fujisaki and Kawashima (1968).

It is the purpose of the present article to review our earlier studies and to add further experimental data.

The specific problems we have had in mind are the following:

(1) The phonetic validity of F'_2.
(2) Can F'_2 be predicted from a knowledge of F_1, F_2, F_3, and F_4? If so, how?
(3) Further evidence on integration of vowel timbre in dichotic listening.
(4) How is F_1 perceived in specific at high F_0? By the most prominent harmonic alone or by a weighting of several spectral components?

The matching experiment

The vowel stimuli were produced by a computer simulation of parallel formant synthesis so as to retain control over formant amplitudes. The four-formant vowels were given formant amplitudes as computed from a serial analog model and the amplitude of the upper formant of the two-formant vowel was preset inversely proportional to its frequency, F'_2. The first formant of the two-formant vowels was chosen the same as in the four-formant reference and the same rise-fall intonation (with a mean F_0 of 120 Hz) contour was used. Three phonetically trained subjects

were used. No real difficulty was encountered in the matching although naturalness was typically better for back vowels than for high front vowels. The maximum deviation of any subject's preferred setting of F'_2 was of the order of ±10%. In occasional series of matching the spread was of the order of a difference limen in F_2. On some occasions the mean value of a series could vary from one day to the next. The consistency was less for the vowel [i] than for other vowels.

Prediction of F'_2 from formant data

Some indication of the relative importance of various peaks in the spectra of typical Swedish vowels is qualitatively seen in the mel scale diagram of synthetic vowels, see Fig.2. These calculated spectra are based on true spectrum envelopes.

When the 1970-article was written we felt that it would be difficult to design a weighting technique to predict the F'_2 from the set of formant frequencies and amplitudes. One reason for this was the highly non-linear dependency of F'_2 on F_3 in the boundary region between the [i] and the [y] vowels. The [i] has 100 Hz higher F_2 and F_4 than [y] and 500 Hz higher F_3. However, the F'_2 of [i] was found to be as much as 1200 Hz higher than in [y] which is larger than the shifts of all formants. In case of the [i] the average match was F'_2=3210 Hz which is 300 Hz above F_3, whereas [y] was matched at F'_2=2100 Hz or 80 Hz above F_2. The effect of a shift of F_3 alone in the boundary region [y]-[i] is demonstrated in Fig.3.

The subject's matching may be influenced by two main factors. One is that of his auditory impression of the test stimuli. The other is the mediation through his perceptual norm of standard phonemes. A part of the observed non-linearity could be related to the latter effect. There could also be some tendency to match F_2 of [y] and F_4 of [i], i.e. to match on a specific formant instead of a weighted mean, but this was observed occasionally only.

A first attempt to calculate and F'_2 as a linearly weighted mean frequency of F_2, F_3, and F_4 with associated amplitudes L_2, L_3, L_4 failed. Much better results were achieved by a direct search for regions of spectral prominence. Information on formant amplitudes was discarded since the main spectral shape features are derivable from the set of formant frequencies, Fant (1960). Also, it is known from our earlier study (1970) and previous work (Lindqvist & Pauli, 1968) that phonemic identity of vowels is preserved within a large range of variations of formant amplitudes. The intuitive approach followed was accordingly to design a formula which would place F'_2 somewhere between F_2 and $(F_3F_4)^{1/2}$. The lower limit $F'_2 = F_2$ should apply when F_2 is close to F_1 as in back vowels. The upper limit $F'_2=(F_3F_4)^{1/2}$ should apply when F_2-F_1 is large and F_3 is much closer to F_4 than to F_2 as in [i] type vowels. On the other hand, when F_3-F_2 is very small, F'_2 should be given a location just above F_2. Intermediate patterns should be taken care of by an appropriate weighting. These considerations eventually resulted in the formula:

(1) $$F'_2 = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c}$$

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

The factor $(F_1/500)^2$ in c was added for best overall match with the measured data.

Table I

Vowel IPA	F_1	F_2	F_3	F_4	F'_2 matched	F'_2 cochlea	F'_2 formula	F'_2 zerocross
i	255	2065	2960	3400	3210	3100	3100	2900
ɪ	375	2060	2560	3400	2370	2300	2340	2400
y	255	1930	2420	3300	2010	2100	2130	2400
æ	605	1550	2450	3400	1960	1900	1880	2000
ø	360	1690	2200	3390	1720	1700	1760	1900
ʉ	280	1630	2140	3310	1730	1600	1670	1700
a	580	940	2480	3290	960	900	1060	
o	400	710	2460	3150	720	700	735	
u	310	730	2250	3300	730	700	745	

The tabulation contains the formant data F_1, F_2, F_3, and F_4 of the reference vowel together with the matched F'_2, estimated F'_2 from the cochlea model described in the next section, and F'_2 given by the empirical formula. These three measures of F'_2 agree within 160 Hz and the average differences are of the order of 75 Hz. The last column in Table I shows the result as a simple zero-crossing frequency count in a 1 kHz-5 kHz band. It is less consistent with the matched F'_2 than that of the cochlea model and the [i]-[y] contrast is much reduced. Accordingly it is a less representative measure of the mean frequency of F_2 and higher formants.

A functional model for deriving F'_2

The model consists of a computer simulated bank of 120 rather broad filters, spaced 38 Hz apart, the output distribution of which can be seen in Fig.4 with the vowel [i] as input. In each channel this filtering is followed by zerocross counting averaged over 100 ms and converted to frequency. A density-measure is established by counting the number of channels in which the same frequency is measured within a quantum range of 75 Hz. A histogram of this measure, see Fig.4, brings out characteristic frequency regions. The two most prominent peaks were found to correspond closely to F_1 and the F'_2 of the two-formant matching, the difference being of the order of 75 Hz. It is also remarkable that the empirical formula generated F'_2 values coincide with those of the matching experiment and the cochlea model with the same degree of accuracy. These three F'_2 measures agree within a maximum deviation of 160 Hz in any pair.

The filtering in the cochlea model was designed to conform with Flanagan's (1965) model of frequency-place analysis along the basilar membrane matched to the Békésy-data. The zero-crossing information would accordingly reflect a temporal fine structure at the input to the primary neurons. More recent experimental measurements of basilar membrane motion show a much steeper response and higher selectivity (Rhode, 1971) and tuning curves from primary neurons also point towards a more selective filtering (Kiang *et al.*, 1965). However, there seems to be a general broadening of the response function higher up in the nervous system, Møller (1972).

The close agreement between the cochlea-based derivations of F'_2 and the empirical derivations from formant frequencies are in part only ascribable to the particular modelling of Eq.(1) which contains one numerical constant only. The common basis of the agreement is that of single component prominence.

A single sine wave or a dominating formant appear both as a spectral peak and as the same zero-crossing frequency in several adjacent taps of the cochlea output. To the extent available, zero-crossing frequency information may sharpen the spectral resolution as well as suppress weaker components thus bringing out elements of auditory prominence. However, the same process could be administered entirely in the spectral place-magnitude domain and we do not claim that our experiments would back up one or the other of the two models of parametric representation as being more valid in a physiological sense.

Split vowel experiment

It appears reasonable to assume that the perception of vowel timbre engages peripheral as well as more central auditory functions. To obtain some insight in the merge of sensation we designed a forced choice identification test with four-formant stimuli distributed in the [i]-[y] domain with F_3 as the only variable frequency. Test conditions included:

(1) (a) F_1 and F_2 were presented to the left ear and F_3 and F_4 to the right ear.
(b) Vice versa.

(2) (a) F_1, F_2, and F_3 to the left ear and F_4 to the right ear.
(b) Vice versa.

(3) Binaural presentation of
(a) F_1+F_2
(b) F_3+F_4
(c) $F_1+F_2+F_3$
(d) F_4
(e) $F_1+F_2+F_3+F_4$

The stimuli were presented over headphones in random order. In all 20 normal hearing subjects participated. Test conditions (a) and (b), interchange of ears, did not show significant differences and the results were therefore pooled. Test conditions (3) were set in as a control of the extent to which the split stimuli results in (1) and (2) could be predicted from the particular response to the stimulus in either ear. The result was negative.

The results from these tests are shown in Fig.5 where each point represents the average of 80 responses for [i] and [y] identity. The split stimulus presentation (1) and (2) evidently gives almost the same identification curve as the normal presentation (3.e), the slope and the 50% identification being nearly the same. One difference is that absolutely unanimous [i] responses were never obtained for the split vowels.

Interaction between F_0 and F_1

We have so far discussed formant patterns as frequency domain envelopes and spectral shapes without considering the harmonic fine structure. The

selectivity of the ear is sufficient for resolving low-frequency harmonics providing F_0 is greater than the critical bandwidth, or F_0>100 Hz (Plomp, 1964).

Two different hypotheses about the perception of the first formant could be proposed.

(a) The listener can reconstruct the peak of the envelope from the perceived harmonics irrespective of whether there is energy at the peak or not.

(b) The listener selects the largest peak of the auditory pattern and ignores other partials. At high F_0 a single harmonic is picked out.

According to Chistovich the second hypothesis is the most probable one, see Chistovich (1971) and Mushinikov and Chistovich (1972). In a matching experiment, where a two-formant synthesis was used as reference and the test stimulus consisted of the same second formant and one variable sinusoid positioned in the low-frequency domain, they got a result pointing towards hypothesis (b). The subject positioned the sinusoid close to a partial in the reference, especially when a high F_0 was used. Best phonetic equality was aimed at as matching criteria.

As a contribution to the discussion we would like to offer the following more specific interpretation. The auditory impression of vowel timbre includes the perceived fundamental pitch which attracts an increasing proportion of the listener's attention at high F_0-values and especially at a monotone pitch. The pitch may be remembered from the reference vowel, or perceived as a residue from the upper formant of the test vowel. The matching might accordingly be thought of as engaging the subject in two related but different tasks.

(a) Positioning of the sinusoid in the neighborhood of F_1.

(b) The sinusoid is tuned close to the most 'prominent' harmonic in the F_1 region to satisfy the additional demand for harmonic congruence.

Two separate hypotheses have been suggested by Mushnikov and Chistovich (1972) for the peak selection. One corresponds to the criterion of maximum loudness, the other to the criterion of maximum (phonetic) significance. These are illuminated in an experiment where the subject had to manipulate the level of one of two harmonically related sinewaves in the first formant region for equal loudness. After that a synthetic F_2 was introduced and the level of the sinewave was adjusted by the subject so as to place the stimulus in the phonetic boundary between [i] and [e]. This procedure defines the so-called 'equal significance' relation between the two sinusoids, and an equal significance space could be constructed. The result could be summarized as follows:

(a) The equal loudness curve has an approximate slope of -6 dB/oct below 600 Hz.

(b) The equal significance curve has a slope of 6 dB/oct below 600 Hz.

That means that the lower the frequency of a partial, the lower level is needed to enhance an [i] response whilst a higher level is needed to provide the same loudness as the sinewave of higher frequency.

Fig.6 shows an F_1 envelope contour with three different pre-emphases, 0, +6 dB/oct, and -6 dB/oct. These are SPL, loudness (L) and significance (S), respectively. There is a significant difference between L and S. L sharpens the formant peak and the pattern is similar to the vocal tract transfer

function. S, on the other hand, introduces a low-pass shape of the pattern with a maximum below F_1 and the perceptual decision procedure has to be of a different kind than a peak-picking in the S-domain.

In order to study the interaction between partials, formant frequency, and F_0 we made an identification test using synthetic steady-state vowels. Formants F_2, F_3, and F_4 were the same in all stimuli and the position of F_1 was varied in a range covering the boundary between the Swedish [i] and [e]. F_0 was held constant or followed a contour with a maximal deviation of 4% from the mean. The result of the identification test is shown in Fig.7a. With increasing F_0 the boundary is shifted towards higher F_1 values as could be expected from the data of Miller (1953), Carlson, Granström and Fant (1970), and Fujisaki and Kawashima (1968).

A slight perturbation of F_0 should provide the listener with more detailed information about the spectral envelope. However, such an improvement is not detectable in our data. On the contrary a slight uncertainty in the decision appears to be added.

To compare our results with the equal significance concept the most significant harmonic was estimated for each stimulus and plotted in Fig.7b. The ordinate is now the frequency position of this harmonic but the identification curve (constant F_0) is included for comparison of e.g. phoneme boundary position and shift of most significant harmonic. Obviously the correlation is very low and sometimes negative.

In Fig.7c the identity curves are shown together with frequency position of the highest partial in the loudness dimension.

No simple method was found to predict the identity scores from these measures. However, the second loudest partial indicated by the direction of arrows in Fig.7c appears to add a systematic trend.

Let us accordingly hypothesize four different ways the listener might extract a parameter representing the most important frequency (MIF) in the low frequency region:

(a) estimate the most prominent partial in the 'equal significance' space.

(b) estimate the most prominent partial in the loudness space.

(c) compute the weighted means of the two most prominent partials /m/ and /n/ in the loudness (sone) space.

$$MIF = \frac{f_m S_m + f_n S_n}{S_m + S_n}$$

(d) compute the weighted mean of the *three* most prominent partials in the loudness space.

The hypotheses (a) and (b) have been rejected in the discussion above.

In Fig.8 hypotheses (b), (c), and (d) are represented by computed MIF at the observed phoneme-boundaries from the identification test. Since phoneme-boundaries in the vowel space are monotonous functions of F_0 the MIF has to be a monotonous function of F_0. Hypothesis (c) is the only one that provides a monotonous MIF-F_1 relation and it also shows the best fit of MIF to the physical F_1.

The result suggests that the listener could use some interpolation mechanism to estimate the formant frequency as opposed to a selection of the loudest or most 'significant' harmonic.

General discussion

Phonetic significance of the F'_2 parameter

The unavoidable compromise in speech research is that between the simplicity of models and their accuracy. Two-formant approximations hold well for back vowels where F_3 and F_4 are relatively weak. Even single-formant approximations preserve essentials of the phonetic value of back vowels. The two-formant approximation holds least well for high front vowels. However, all vowels of the rich Swedish vowel system could be satisfactorily matched and identified by two-formant approximations.

A more detailed vowel stimulus model would include an extra upper formant or a measure of spectra spread in the F'_2 domain. Such an extension would improve [i] vowels but is not necessary for the distinction between Swedish [y] and [ʉ]. The tendency of increasing [ʉ] response and decreasing [y] response when increasing the distance between F_2 and F_3 whilst maintaining their geometrical mean, Fujimura (1967), has two possible explanations. One brought forward by Chistovich and Kozhevnikov (1970) is that the spectral spread in the F_2-F_3 region is a secondary perceptual parameter in Swedish, the F_2 and F_3 of [ʉ] being further apart than a critical bandwidth. We would rather suggest that the effect of separating F_2 and F_3 of vowels located in the [y]-[ʉ] boundary

region is to shift the spectral balance towards a lower F'_2, which can be verified from the empirical formula.

Thus, a vowel halfway between [y] and [ʉ] of F_1=270 Hz, F_2=1780 Hz and F_3=2280 Hz, F_4=3350 Hz has an F'_2 of 1850 Hz whereas the extreme condition of $F_2=F_3=\sqrt{1780 \cdot 2280}=2015$ Hz corresponds to F'_2=2015 Hz. In connected speech the distance between F_2 and F_3 is about the same for the long and stressed [y:] and [ʉ:]. In most dialects [y:] displays a rising F_3 diphthongal glide towards a rounded [j] and the [ʉ] vowel is diphthongized towards a bilabial closure which lowers both F_2 and F_3. These diphthong elements can also be approximated in two-formant synthesis which improves the naturalness of the [i], [y], [ʉ], and [u] vowels.

On the other hand, the interpretation of Chistovich and Kozhevnikov (1970) that Swedish [y] is characterized by F_3 very close to F_2 has a support in the Fujimura-Lindqvist (1971) sweep-frequency analysis of sustained silent articulations. In their data F_2F_3 form a single peak. This represents a very extreme articulation.

There is not much gain in restricting speech synthesis to two-formant representations except for special research in perception. The early Haskins Laboratories' synthetic speech employed two-formant vowels with typical data close to those we have found in the matching experiment. The higher F'_2 of Swedish [i] vowels, compared to 2700-2900 Hz of Delattre, Liberman, Cooper (1951) may in part reflect a phonetic difference, in part be the consequence of the English vowel system lacking a vowel [y] to be contrasted

with [i]. It should be observed that the upper limit of F'_2 values available for the subject in an identification test and the relative crowdiness of the vowel system will bias the F'_2 measures, see Ainsworth (1971) who reports exceptionally low F'_2 values for the vowel [i].

The very close agreement between the empirical formula Eq.(1) relating F'_2 to F_1, F_2, F_3, and F_4 and the results from the matching experiments and the 'cochlea' functional analog suggest that the formula might be useful for descriptive phonetic work replacing the simple F'_2 formula developed by Fant (1959). Alternatively, instead of measuring formants, which is notoriously difficult, one might employ a cochlea analog for taking speech spectrograms. A simpler method is to take the mean zero-crossing frequency from a high-pass filter removing F_1. This was tried but did not provide the same accuracy as the cochlea analog, see Table I. However, it is interesting to note the close similarity between our F'_2 and measures of Scully (1968) who performed a simple time-domain 'ripple' analysis of front vowel formants above 1000 Hz.

We feel that more experience is needed to assess the practical value of F'_2 specifications of vowels. It appears easier to detect reliable F'_2 values for synthetic speech than from natural speech. It would also be advisable to include the relative amplitude of F'_2 as an additional parameter for discriminating between vowels and voiced consonants. The [r] colored vowels characterized by F_3 very close to F_2 would need higher amplitudes of F'_2 than other vowels, Miller (1953), and voiced consonants would need lower amplitudes of F'_2 than vowels.

One interesting property of our models is the nonlinear relation between formant movements in the incoming sound and the associated shifts in F_2. Such relations should be held in mind when discussing the relative sharpening of the discrimination at phonetic boundaries, Stevens *et al.* (1969).

Mechanisms of data reduction in vowel perception

It has not been our intent to develop a complete theory of vowel perception. A more systematic development of an auditory model of the perception of steady-state vowels on the basis of psychoacoustic data is being presented in the paper of Karnickaya *et al.* (1973) to this symposium. Their approach and ours provide similar results with respect to the gross type of data reduction in the auditory system before the stage of phonetic identification. This similarity involves the calculation of the frequency locations of two major peaks whilst the configuration of the spectrum between these peaks is ignored. The suppression of secondary peaks is the result of the finite selectivitiy of the filter bank model, i.e. the dominance of a signal of one frequency over signals of other frequencies at a specific spatial coordinate (interband masking). An additional stage of 'lateral inhibition' provides additional sharpening of the main peaks in the model of Karnickaya *et al.* (1973).

In our model the broad filters account for extremely unselective amplitude-coordinate excitation patterns, whereas the particular parameter we have extracted, i.e. the density of taps carrying the same response frequency brings out the major peaks

and eliminates insignificant amplitude information. The phonetic identity of a vowel is independent of the overall level of presentation and within wide limits also independent of the relative amplitudes of the two major peaks. On the other hand, the relative amplitudes of formants within a main spectral peak are of importance to the extent only that they influence the spectral balance within the group of formants and thus its center of prominence.

Although this F_1F_2 extraction has been found to be effective for data reduction of spoken and synthetic four-formant vowels, it does not hold equally well for all voices and less well for back vowels. We need more experience of this technique and how it performs when assuming a more selective cochlea analog.

There still remains an argument as how to describe the peak-picking mechanism in the first formant range, whether single harmonics or a weighted mean of adjacent harmonics represent the perceived formant frequency. We find evidence in favor of the latter view from the essentially monotonous [i]-[e] boundary shift in F_1 when F_0 is successively increased. If single harmonics were picked out to represent F_1 we would expect the phonetic quality of a vowel produced with constant F_1 and varying F_0 to display discontinuities whenever the envelope peak of F_1 falls halfway between two harmonics. This does not appear to be the case. We accordingly hypothesize a mechanism of spatial integration and weighting of adjacent auditory components at a stage above that of spectral sharpening and

secondary peak suppression.*

The fact that a vowel retains its phonetic identity when some of the formants are presented to the right ear, the other formants to the left ear indicates a central or at least non-peripheral summation of auditory components. It would be interesting to study how the phonetic identity of a vowel is dependent on the relative intensities of the components presented to both ears. It is not claimed that the time-domain aspect of our functional model has a physiological significance. Whether it is the spatial distribution of intensities or of zero-crossing frequencies in adjacent filter bands, that are signalled to higher levels, may be of less importance in view of the dual nature of these parameters.

Average F_0-F_n trading. Mel scale spacing

We find a monotonous relation of higher F_1 values required for maintaining phonetic identity at increasing F_0. This result was confirmed by the identification tests in our previous work, Carlson, Granström, Fant (1970) and several earlier investigations. More generally we have observed a trading relation between F_0 and formants, a rise in F_0 from 120 to 240 Hz requiring a compensatory increase of M_1+M_2 by on the average 70 mel. This shift is 2-4 times smaller than the shift in formant scale factor on the mel

*As revealed by the discussion at the meeting, the Leningrad group now favors a statistical approach, based on the probabilities of vowel identity associated with each of the competing harmonics. Sharp discontinuities are thereby avoided. In our approach the probability function follows the mean frequency of the formant.

scale comparing males and females and might represent an out of context speaker-sex association.

We would like to offer the following alternative explanation. When a vowel is sustained at a fairly constant and high F_0 the auditory impression is very much colored by the fundamental. In the extreme high register of a soprano singing voice the vowel loses most of its phonetic identity. In less extreme instances with F_0 in the range of 200-300 Hz the individual harmonics, although separated by larger distances than the critical bandwidths, will combine to evoke a strong sensation of the fundamental. This tone will fuse with the timbre and shift the mean phonetic pitch of the sound from the auditory mean of F_1 and F_2 or alternatively F_1 alone to F_0, i.e. to a lower equivalent frequency. The seemingly paradoxal result is that an increase in the frequency of one component, F_0, lowers the mean pitch in the timbre domain. Thus, if the mean timbre pitch is denoted by $M_1+M'_2$ and the voice pitch by M_0 we might substitute

$$(2) \qquad M_g = \frac{(M_1+M'_2) + bM_0^q}{1 + bM_0^{q-1}}$$

for the $M_1+M'_2$. Here the exponent q represents the relative growth of the pitch interference in the perceived timbre and should be greater than 1, perhaps 2 or 3.

There is some tendency for distance between phonetic prototypes in the $M_1M'_2$ domain for Swedish vowels to conform to ordering within a set of M'_2-M_1 lines quantized approximately as

$$M_2-M_1 = n\ 250 \text{ mel}$$
$$M_2+M_1 = (n+3)\ 250 \text{ mel}$$

The particular mel scale adopted here is the analytical approximation of Fant (1959)

(3) $$M = 1000 \log_2 (1+F/1000)$$

where F is frequency in Hz and M is 'technical mel'. This choice of parameters has the benefit that all back vowels have approximately the same M_2-M_1 and all unrounded front vowels the same M_2+M_1, see Fant (1971).*

*A demonstration tape is available for those who have an interest in listening to and evaluating this constant mel spacing vowel ensemble.

References

Ainsworth, W.A. (1971). Perception of synthesized vowels and h-d words as a function of fundamental frequency. *Journal of the Acoustical Society of America* 49, 1323-1324.

Carlson, R., Granström, B., & Fant, G. (1970). Some studies concerning perception of isolated vowels. *STL-QPSR* 2-3, 19-35.

Chistovich, L.A. (1971). Problems of speech perception. Pp.83-93 in *Form and Substance*, Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen, eds. L.L. Hammerich, R. Jakobson & E. Zwirner. Copenhagen: Akademisk Forlag.

Chistovich, L.A., Fyodorova, N.A., Lissenko, D.M., & Zhukova, M.G. (1973). Auditory segmentation of acoustic flow and its possible role in speech processing. Paper to be discussed in Session IV of Symposium on Auditory Analysis and Perception of Speech, Aug. 21-24, 1973, Leningrad - this volume.

Chistovich, L.A., & Kozhevnikov, V.A. (1970). *Theory and Methods of Research on Perception of Speech Signals*. JPRS-50423, Washington,DC, translated from the Russian.

Delattre, P.D., Liberman, A.M., & Cooper, F.S. (1951). Two-formant synthetic vowels and cardinal vowels. *Le Maître Phonétique*, July-December.

Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics* 1, 3-108.

Fant, G. (1960). *Acoustic Theory of Speech Production*. 's-Gravenhage: Mouton (2nd ed. 1970).

Fant, G. (1971). Distinctive features and phonetic dimensions. Pp.219-239 in *Applications of Linguistics. Selected Papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, eds. G.E. Perren & J.L.M. Trim. Cambridge University Press.

Fant, G. & Risberg, A. (1963). Auditory matching of vowels with two formant synthetic sounds. *STL-QPSR* 4, 7-11.

Flanagan, J.L. (1965). Computational models for ear operation. Pp.91-118 in *Speech Analysis Synthesis and Perception*. Berlin: Springer-Verlag.

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

Fujisaki, H. & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics* AU-16, No.1, 73-77.

Karnickaya, E.G., Mushnikov, V.N., Slepokurova, N.A. & Zhukov, S.Ja. (1973). Auditory processing of steady-state vowels. Paper to be discussed in Session III of Symposium on Auditory Analysis and Perception of Speech, Aug.21-24 1973, Leningrad.

Kiang, N., Y-S., Watanabe, T., Thomas, E.C. & Clark, L.F. (1965). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*. Research Monograph 35. Cambridge, Mass.: The MIT Press.

Mushnikov, V.N. & Chistovich, L.A. (1972). Method for the experimental investigation of the role of component loudnesses in the recognition of a vowel. *Soviet Physics-Acoustics* 17, 339-344.

Lindqvist, J. & Pauli, S. (1968). The role of relative spectrum levels in vowel perception. *STL-QPSR* 2-3, 12-15.

Miller, R.L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America* 25, 114-121.

Møller, A.R. (1972). Coding of sounds in lower levels of the auditory system. *Quarterly Review of Biophysics* 5:1, 59-155.

Plomp, R. (1964). The ear as a frequency analyzer. *Journal of the Acoustical Society of America* 36, 1628-1636.

Rhode, W.S. (1971). Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *Journal of the Acoustical Society of America* 49 1218-1231.

Stevens, K.N., Liberman, A.M., Studdert-Kennedy, M. & Ohman, S. (1969). Crosslanguage study of vowel perception. *Language and Speech* 12, 1-23.

Zwicker, E. & Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfänger. 2nd revised edition. Stuttgart: S. Hirzel Verlag.

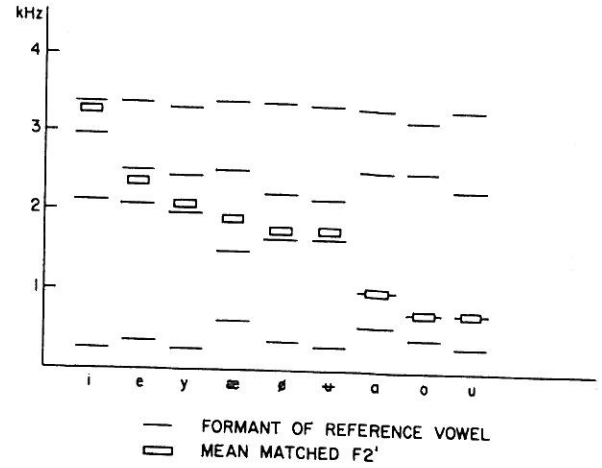

Figure 1 Result of a matching test.

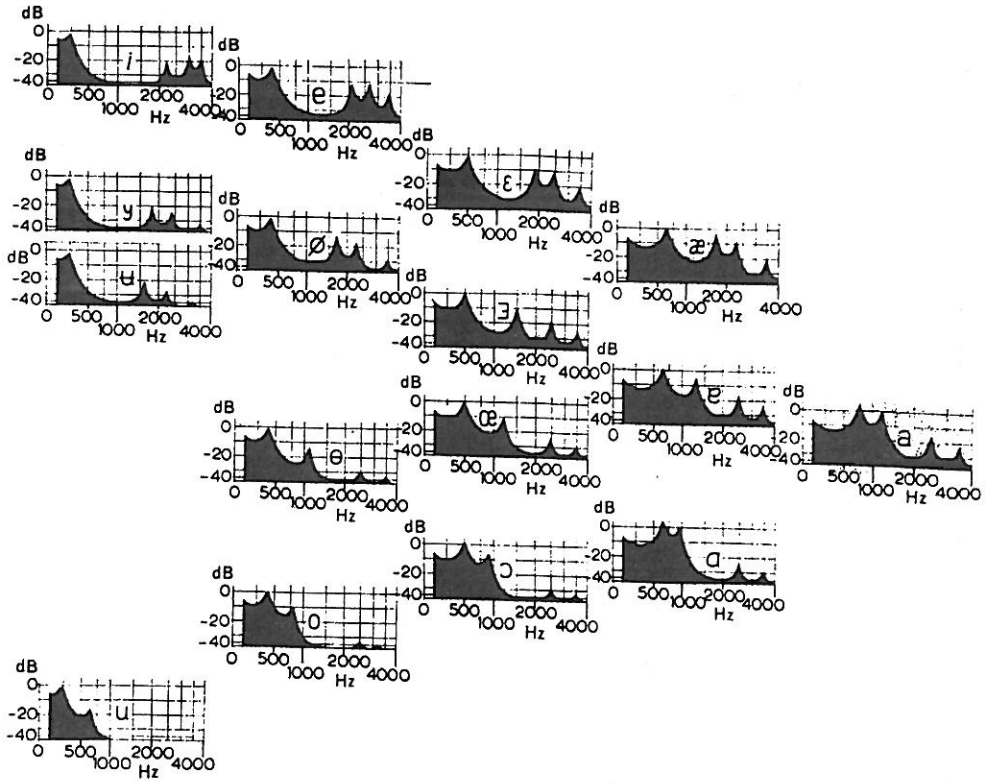

Figure 2 Spectra on an approximate mel scale of synthetic vowels ordered according to the particular F_1 and F_2. The changes in spectrum shape and in formant levels following a shift in one or more of the formant frequencies should be observed. (Fig.8 in G. Fant: The acoustics of speech, *Proc. 3rd ICA*, Stuttgart 1959, Vol.I).

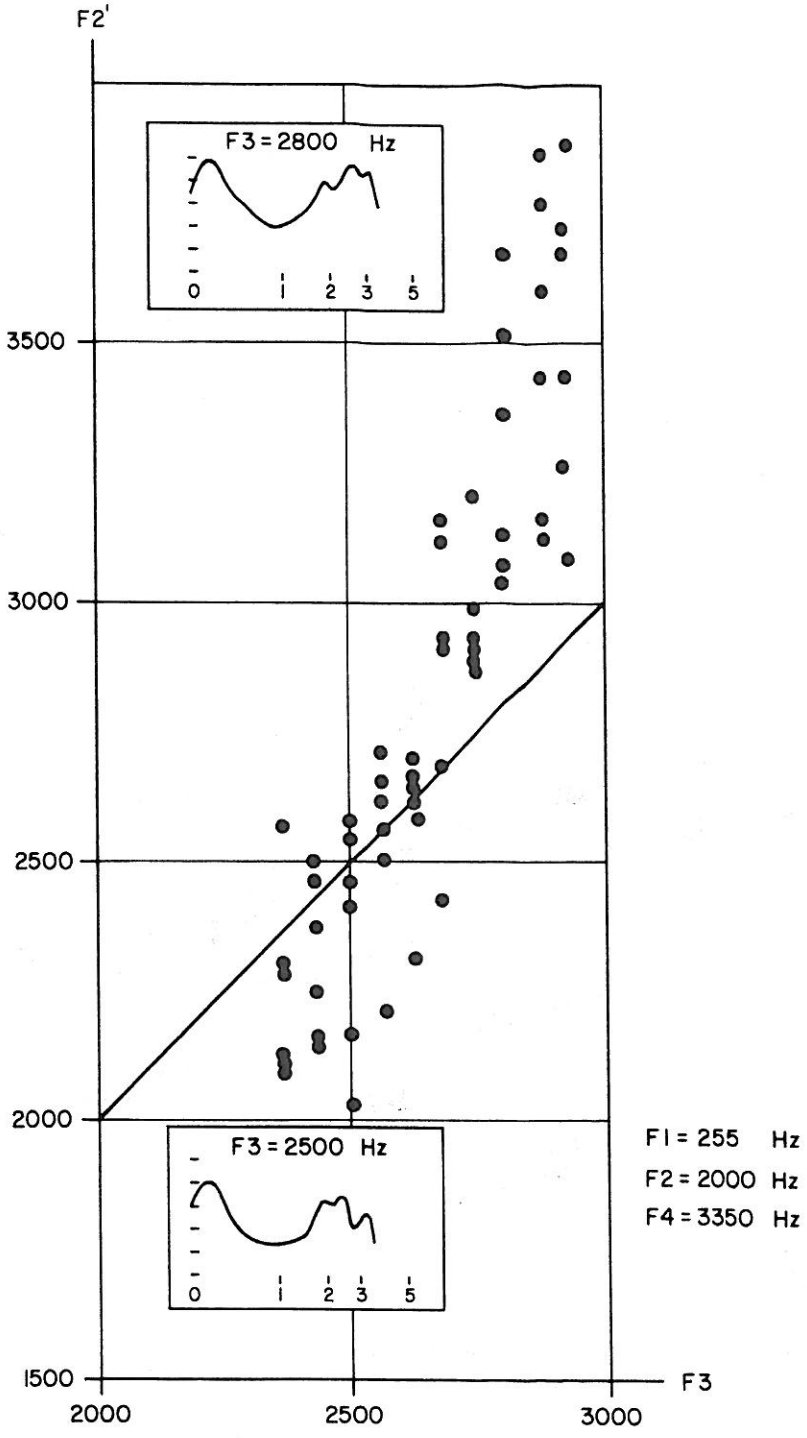

Figure 3 Result of a matching experiment. F_3 of the reference vowel varying from an [y] to an [i] position.

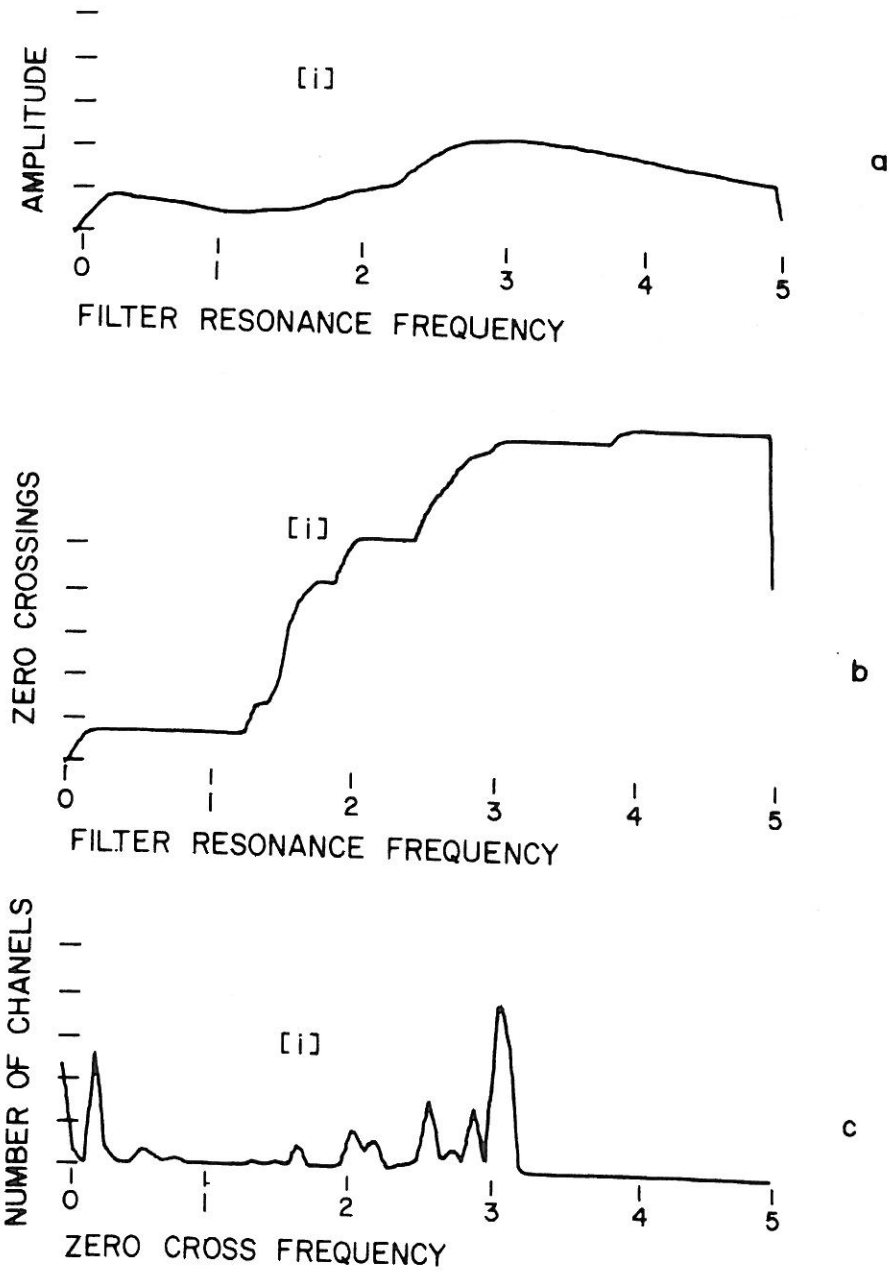

Figure 4 Output from the ear model described in the text.
(a) Amplitude envelope on the basilar membrane.
(b) Zero cross distribution along the basilar membrane.
(c) Histogram. Zero cross frequencies are grouped in 75 Hz intervals.

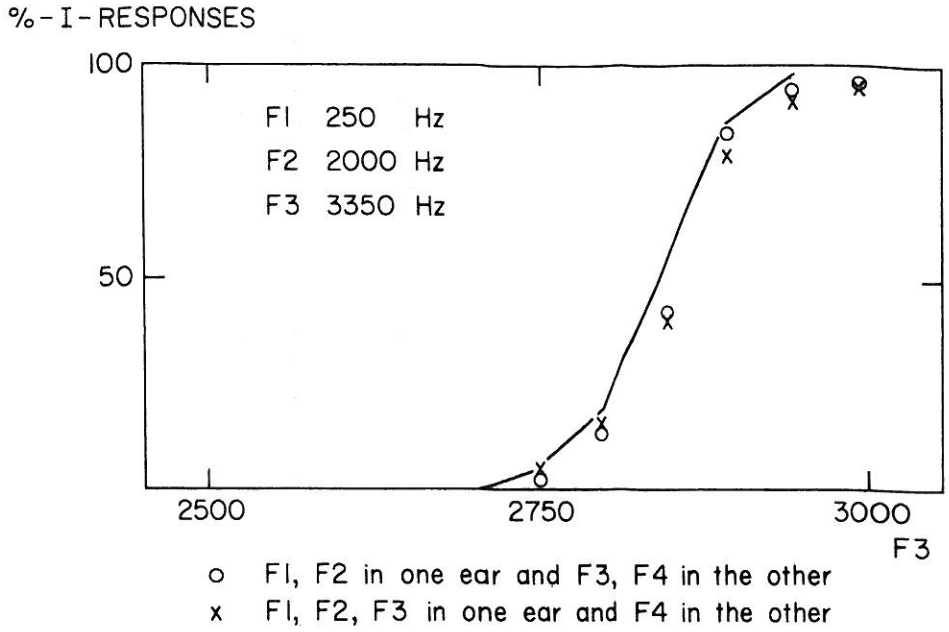

Figure 5 Solid line pertains to normal listening conditions.

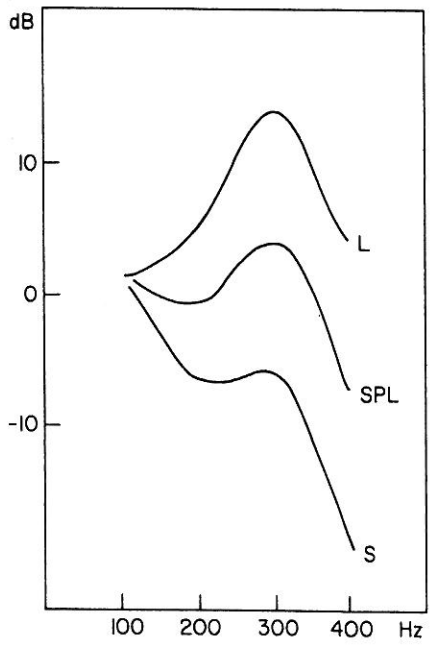

Figure 6 Envelope of first formant a fter different pre-emphasis.
L: +6 dB/oct approx. equal loudness,
SPL: unfiltered,
S: -6 dB/oct approx. equal significance (see text).

$F_2 = 2065$
$F_3 = 2850$
$F_4 = 3400$

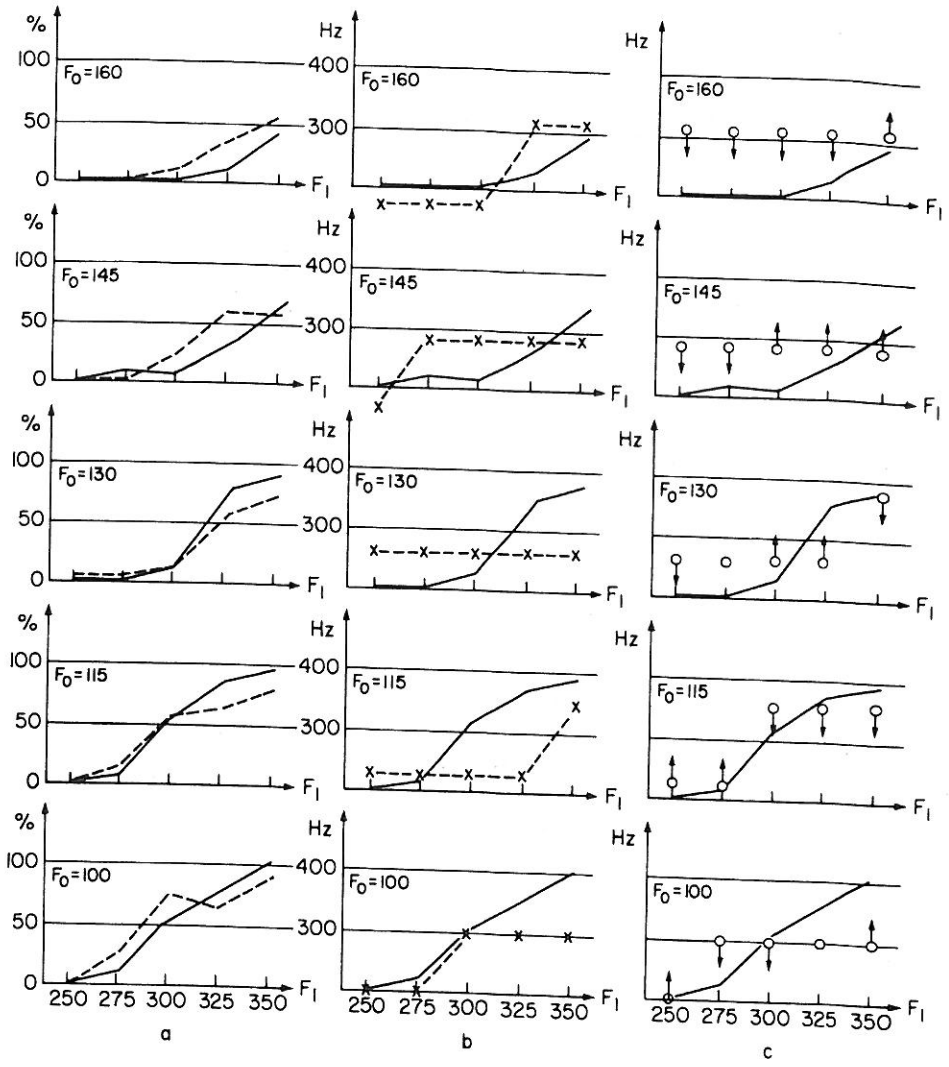

Figure 7 (a) Percent of [e] responses obtained from an identification test with varying F_0 and F_1 (—— monotone pitch, ---- varying pitch).

(b) x---x frequency value of the most significant harmonic (see text).
—— the same as in (a).

(c) ↑ ↓ frequency value of the loudest harmonic. Arrow indicate the direction to the second loudest harmonic (see text).
—— the same as in (a).

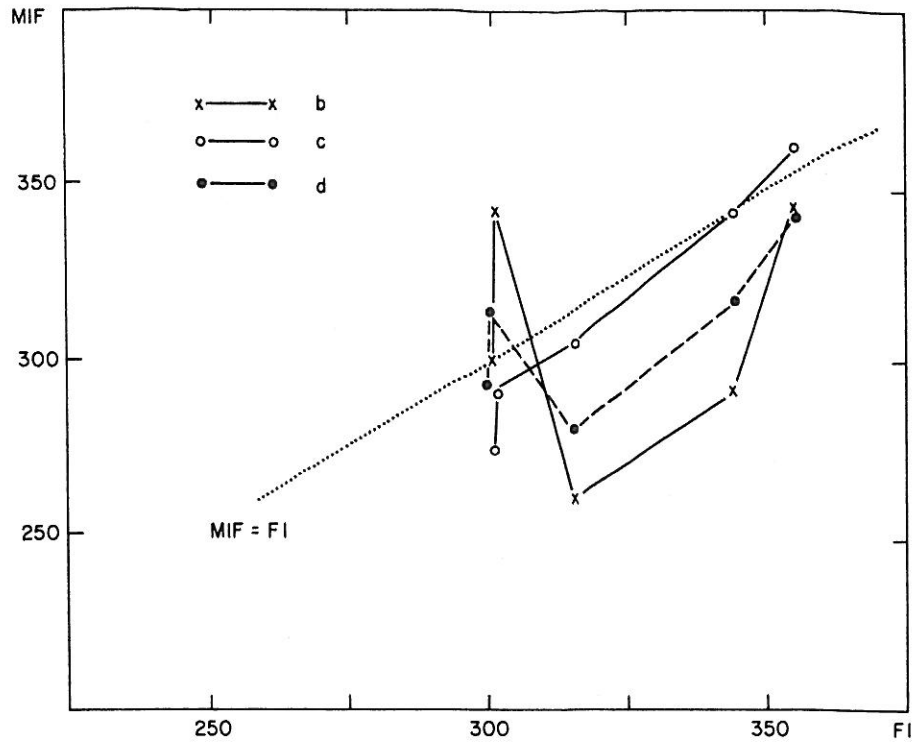

Figure 8 Estimated most important frequency, MIF, of the first formant as a function of the F_0 dependent F_1 of the [i]–[e] boundary for the different hypotheses b, c, and d (see text).