ACCURACY AND LIMITATIONS OF SONA-GRAPH MEASUREMENTS

BJÖRN LINDBLOM

SUMMARY

In the following we shall briefly review the basic terms involved in the description of speech spectra. We shall point out that the acoustic specification of vowel sounds in terms of formant patterns owes its simplicity to the introduction of the *pole* concept in the description. This is often overlooked.

Among the sources of error in specification by sound spectrography some are inherent in the speech wave itself and others are caused by the analyzing instrument. Fig. I shows a few examples of common sonagraphic displays.

The accuracy in vowel formant measurements from sonagrams was found to be about 40 c/s in an experiment with synthetic vowels.

Is it possible to find a simple formula that automatically transforms spectral data into formant frequency values? Probably not. An alternative procedure is described which uses an inventory of *standard envelopes* or standard formant shapes which are applied to the spectrum under consideration.

THEORY

A fundamental concept in the acoustic description of a speech sound is that of the *spectral envelope.* Its constituent parts are the transfer function of the vocal tract and more or less constant factors such as source characteristics and radiation. Mathematically any envelope is most simply described in terms of *poles* and *zeros* which are points in the complex frequency plane. Poles correspond to resonances within the vocal tract and zeros to anti-resonances. The shape of a pole curve, or a zero curve, is completely specified by data on its frequency and bandwidth (Fig. IIa). The mathematical description of an envelope is equivalent to a decomposition of this envelope into a number of known functions, poles (and zeros), which, if multiplied by each other or summed on a dB scale, would restore the original shape of the envelope (Fig. IIb). Since bandwidths can be predicted from pole or formant frequencies, the measure to take on a zero-free vowel is simply the pole frequencies. More elaborate specifications would in ideal cases be redundant (1, 2).

In a harmonic spectrum (the [ɑ]-sample in Fig. I) the levels and the frequencies

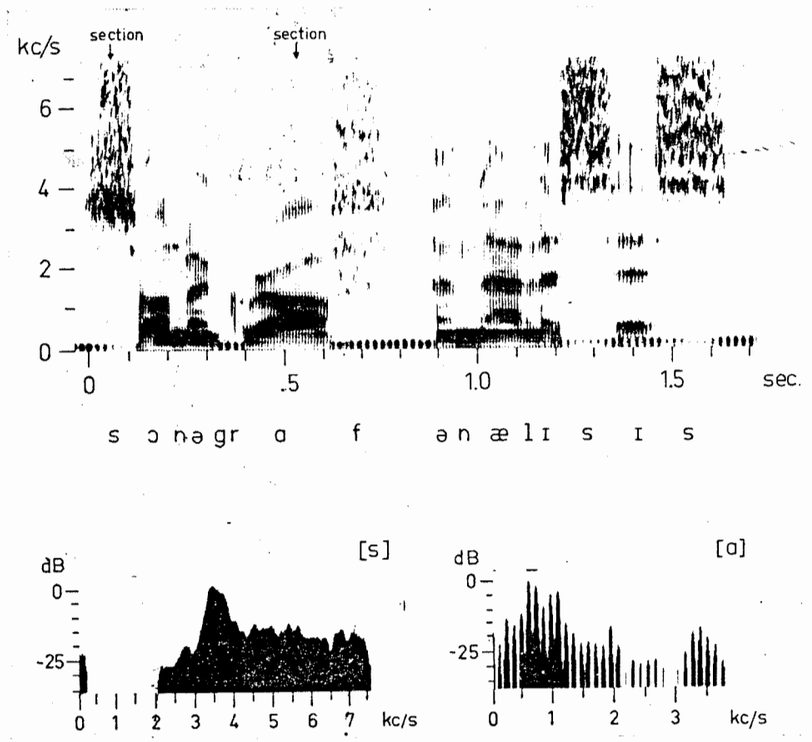


Fig. I. Typical Sona-Graph records. Wide-band spectrogram of the utterance "Sona-Graph analysis". Below wide-band section of the first [s] and narrow-band section of the [ɑ]

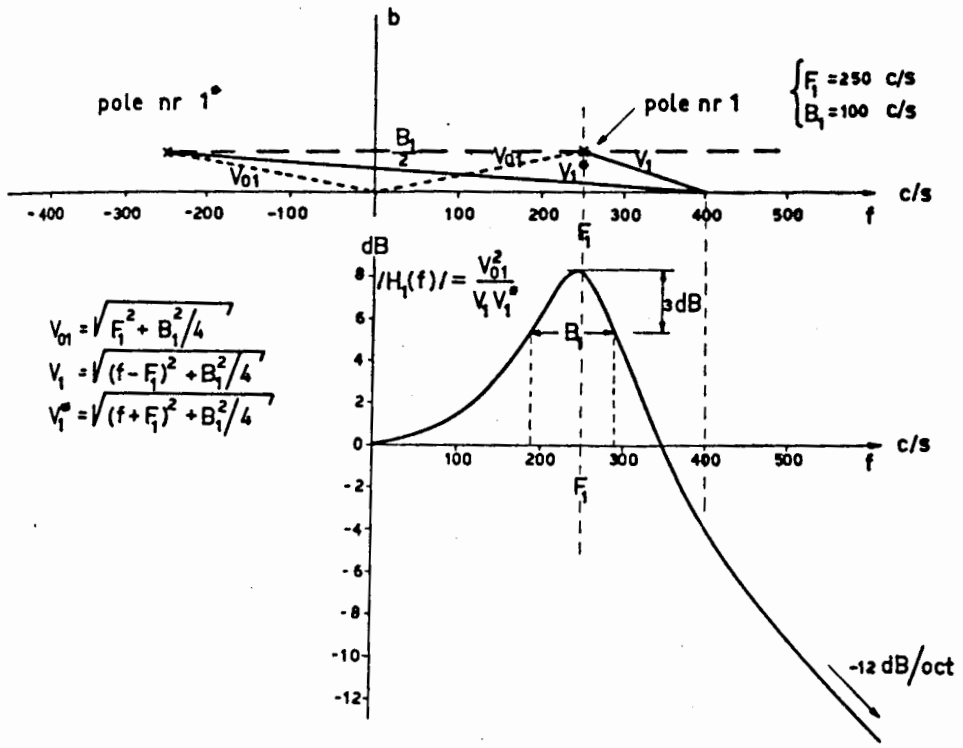
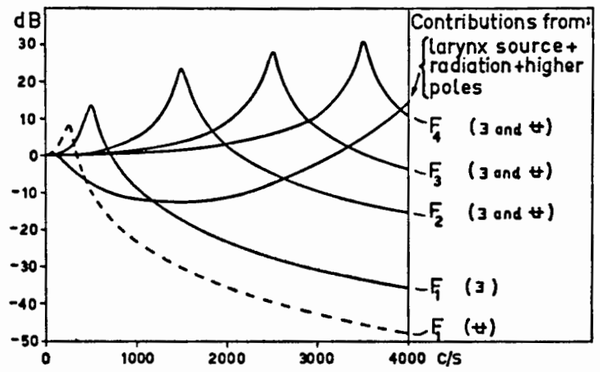


Fig. IIa Shows how a pole curve can be derived when pole frequency and bandwidth are given. (After Fant).

of the components define the envelope in a number of points; the lower the fundamental frequency the better defined the envelope. The position of the *strongest partial* within a *formant* or *energy maximum* is independent of that of the corresponding *envelope peak*; they may or may not, coincide depending upon the interrelations between pole frequency and fundamental frequency. The *pole* has no direct spectrographic manifestation.

It is thus clear that, when we measure the formant frequencies of a vowel we always aim at estimating the pole frequencies. Unless our measurements stand for poles they have no theoretical justification. On the other hand the difference between the frequency location of an envelope peak and the corresponding pole is negligible except in cases when two poles, or a pole and a zero, come very close together. In reality this implies that, if we succeed in determining the locations of the envelope peaks we shall, in most cases, have obtained a good estimate of the pole frequencies or the *F-pattern*. Given the poles we are able to reconstruct the envelope and thus our description covers all the relevant information that the spectrum may contain.



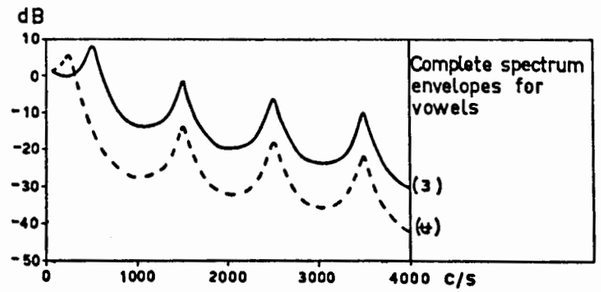


Fig. IIb. Analysis and synthesis of vowel envelopes. (After Fant).

SOURCES OF ERROR INHERENT IN THE SPEECH WAVE

In considering the factors that may jeopardize the precision of formant frequency measurements we find difficulties inherent in the structure of the speech wave itself and limitations imposed by the Sona-Graph. Several investigators (3, 4, 12) have brought attention to the following sources of error in vowel analysis.

(1) The higher the fundamental frequency the less information on the envelope shape. Fig. IIIa exemplifies this difficulty. We see that the partials in this idealized vowel spectrum indicate only one spectral peak whereas the envelope displays two. It is interesting to note that, in spite of the identical envelopes, the ear would probably not equate Fig. IIIa and Fig. IIIb with respect to phonetic quality; Fig. IIIb could be transcribed [ɑ] whereas Fig. IIIa would sound more [ɔ]- like (5).

(2) The more asymmetrical a formant, the more distant the envelope peak may be from the strongest partial within the formant. In Fig. IIIb the envelope is fairly well defined by the relatively dense pattern of partials. There is, however, a certain degree of asymmetry in the second formant. The asymmetry is also obvious in F1 of

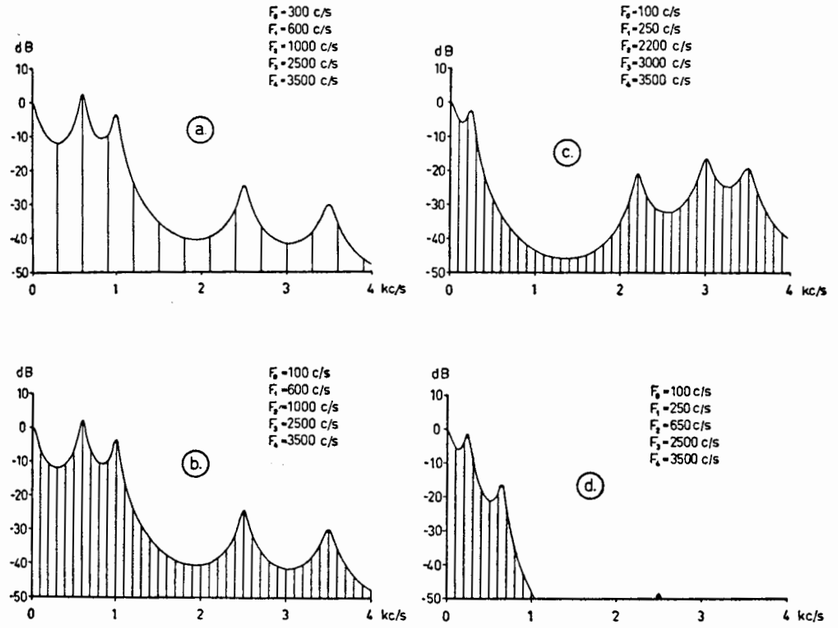


Fig. III illustrates cases where envelopes may be hard to draw on basis of the particular configurations of partials. The difficulty may be caused by too high a fundamental component (a), asymmetry in formants (b, c, d), and by lack of information in the energy distribution in high frequency regions (d). The envelope curves were produced by a computer which was programmed to make a mathematical synthesis of four elementary resonance curves, source function, radiation and higher-pole correction.

Fig. IIIc and F1 and F2 of Fig. IIId. F1 of a close [i]-sound often exhibits negative skewness. The use of a weighting formula for deriving the formant frequency[1] would in these cases fail to give a good result.

(3) In close vowels only one slope may be visible in the first formant.

(4) The first two formants of back vowels are often badly defined since they are usually close together. In these cases it might be worth while to consult the wide-band spectrogram which displays the continuity of the F-pattern.

(5) Close back vowels have only a slight amount of energy in the upper formants. Considerable high-frequency pre-emphasis may be needed to make them appear. This is evident in Fig. IIId.

(6) Zeros, i.e., anti-resonances, appear as spectral minima. They originate from

[1] E.g., the one suggested by Potter and Steinberg (6), $F = \frac{\Sigma \omega_i f_i}{\Sigma \omega_i}$ where ω refers to the level and f to the frequency of the ith component within the spectral maximum.

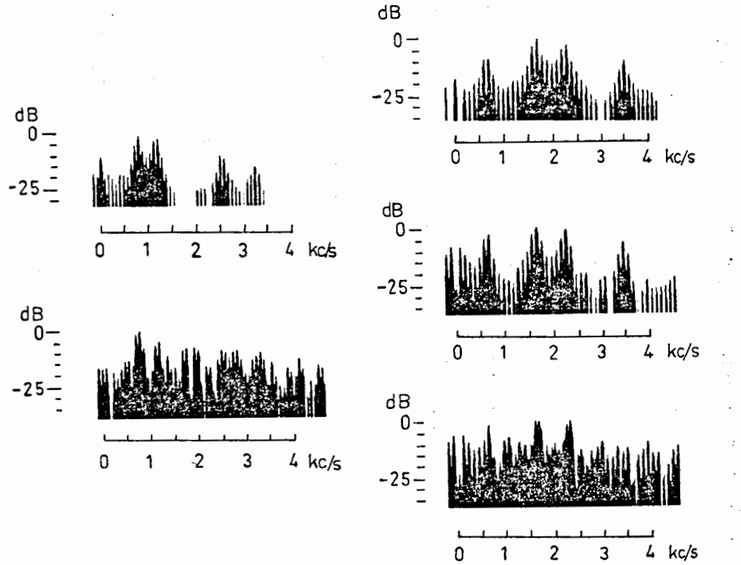


Fig. IV. Samples of steady-state synthetic vowels approaching [a] (left column) and [æ] (right column). The top sections pertain to non-distorted versions. The spectra below exemplify various degrees of distortion.

the source (cf. Flanagan (7)) or a superimposed nasalization. If there is a pole in the vicinity of such a zero the frequency of the corresponding envelope peak may deviate considerably from that of the pole. Moreover, the existence of a zero may give rise to extra formants, i.e., "spurious" formants which do not belong to the F-pattern proper.

(7) In non-stationary intervals the time position of the sample must be chosen more or less arbitrarily. Perceptual experiments may throw some light upon this problem (8).

INSTRUMENTAL SOURCES OF ERROR

There are several instrumental factors to bear in mind when attempting to minimize irrelevant contributions to the speech sample under analysis.

Distortion

Spurious formants may often be artefacts produced by the Sona-Graph. The upper spectra of Fig. IV pertain to relatively pure samples of synthetic [a] and [æ].

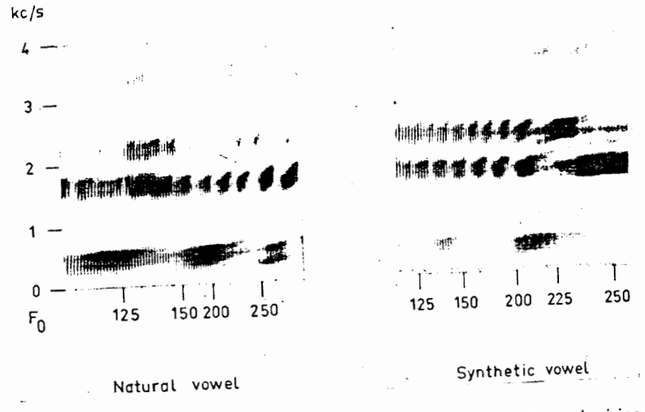


Fig. V. Natural vowel (left) and synthetic vowel (right) on continuously rising fundamental frequency. Subjective estimations of formant frequencies are likely to be influenced by the specific patterns of partials within the formants.

The others display various degrees of distortion. These intermodulation effects could be avoided, however, by keeping the position of the Sona-Graph input control fixed at a high level of amplification and by regulating the level before the Sona-Graph, e.g., on the attached tape-recorder. The critical value on the VU-meter has to be determined by experience.

Time location and sample durations of sections

The time position of a section is apt to vary somewhat owing to the mechanical properties of the microswitch. This uncertainty is about ± 10 msec.

The process of integration involves a weighting of the energy according to a "memory" function. The shape of this function determines the contribution of energy from points along the negative time axis up to the one selected for sectioning (4). The integration time is dependent upon the width of the analyzing filter and the integrating circuitry. The sample durations of narrow-band and wide-band sections are about 30 msec. and 5 msec. respectively.

When sections are made of random noise segment it is important that the time constant of the integrating circuitry is much larger than that of the band-pass filter. In the Kay-Electric Sona-Graph this requirement is not fulfilled and random irregularities appear in wide-band, and still more in narrow-band, spectral sections of e.g., voiceless fricatives. These irregularities may be mistaken for "true" spectral maxima and minima (9).

Among the causes of *spurious formants* we have thus found nasalization and pecularities in the source and instrumental factors such as distortion and limitations in the smoothing filter in the Sona-Graph sectioning device.

Pre-emphasis

A prefiltering of $+6$ dB/octave may give rise to certain inaccuracies in formant measurements. As a general rule a bass attenuation means that the lower a formant is in frequency the more its true shape will be mutilated and the further up in the spectrum we will find its peak. Too steep a bass cut will be a serious source of error in the case of analysis with a "600"-c/s filter.[2] This method which requires the use of a magnified frequency scale, produces a spectrographic display which makes accurate measuring very difficult. Experimental results show errors of 150 c/s in judgments of F_1 of synthetic [i]. On the other hand the time resolution of this analysis is accurate and the method lends itself readily to purposes other than precise frequency measurements.

[2] A reduction of the speed of the input signal by a factor of two results in halving all the frequencies of the speech. The 300-c/s filter will pass twice as many partials as it would if the signal were played back at normal speed. Its effective width is thus 600 c/s. This type of analysis has been recommended for the study of high-pitched voices.

Calibration

The instability of the instrument makes facilities for individual calibration of spectrograms desirable. Errors introduced in the manual stages of the measuring procedure and by unsatisfactory indication of zero line could to some extent be minimized by the use of harmonic spectra in which frequencies can be measured as fractions of F_0. The fundamental component may be extracted with good accuracy by averaging over e.g., ten partials in a narrow-band spectrogram.

EXPERIMENTAL RESULTS

In order to evaluate the accuracy of formant frequency measurements five experienced investigators were asked to estimate the F-patterns of synthetic vowels on sonagrams. No special instructions were given. The subjects just adhered to "current praxis".

Irrespective of whether the method was wide-band spectrogram, wide-band or narrow-band section, the mean error in a fairly large number of measurements was about 40 c/s for male voices.[3]

The major variable influencing judgments seemed to be fundamental frequency. With increasing fundamental frequency the error tended to become larger than 40 c/s. Only rarely did it exceed $F_0/4$ however. This figure is in many cases larger than the difference limen for vowel formant frequencies, which is roughly of the order of 3% (10, 11).

There is a clear tendency for the investigator to be influenced by the strongest component within a formant. This is readily understood when looking at Fig. V, which shows a synthetic and a natural vowel both on a rising pitch. We see the wavy motion of the strongest partials through the resonances. The extent of the motion is an indicator of the errors which might occur in formant frequency measurements.

PROCEDURE

We may now ask a number of questions. For instance, what method of analysis should we choose in order to obtain a maximally precise estimation of the F-pattern of a vowel? Is it possible to state a procedure by which, in a simple way, reliable formant frequency data could be extracted from a Sona-Graph record or any similar spectrographic representation?

It seems reasonable that, for the study of vowels, harmonic spectra will best serve our purposes. In view of the complex variables that influence the shape of the spectral envelope we must conclude that the prospects of finding a formula that will

[3] Standard deviation 40 c/s.

be of general application and automatically give us the frequency of the pole are highly unfavorable.

Unless the formant in question is symmetrical the information contained in a spectrographic configuration of partials within this formant will be insufficient. A correction factor for skewness or asymmetry could perhaps be inserted into our formula but this would entail too great complications.

It appears likely that an investigator would profit more by having an *inventory of standard envelopes* at his disposal from which, on basis of his knowledge and experience, he could select the more probable patterns and apply them to the harmonic spectrum. The procedure suggested is a graphical *spectrum matching* technique.

(1) Estimate approximately the pole (–zero) pattern "by eye" and "by ear", i.e., if our vowel under examination approaches [ɑ] our previous experience of analyzing and synthesizing similar sounds tells us that, if it is a male [ɑ], it is likely to have its first formant F_1 at about 650 c/s, F_2 at 1000 c/s, F_3 at 2600 c/s, and F_4 at 3250 c/s. This is a very rough estimate which may be adjusted by preliminary spectral measurements.

(2) We then select the standard envelope from the inventory whose pole (–zero) pattern corresponds most closely to the one just estimated. The F-pattern of the standard envelopes may of course differ somewhat from that of the sample under consideration as we have made up our inventory of a small number of points strategically chosen in the vowel hyperspace. The important point here though is that, whereas there are an infinite number of combinations of parameter values, formant shapes remain fairly constant for a lot of points in the vowel space. Thus it does not matter if the F-pattern of our standard envelope deviates somewhat from that of the analyzed vowel as the shape of each standard formant will in all probability be fairly congruent with the corresponding formant shape in the spectrum under scrutiny.

(3) Enclose the harmonics of the formant to be measured with the selected standard formant(s) (printed on transparent material). Move the standard formant along the frequency scale till an optimal fit has been obtained and read off the frequency value of its envelope peak.

In the case of two poles, or a pole and zero, close together the frequencies of the peaks and valleys could be adjusted so as to approximate more accurately the actual pole (and zero) frequencies.

The development of an inventory of standard envelopes should involve systematic variation of all the relevant parameters responsible for the shape of spectral envelopes of speech sounds. This means primarily the first four formants, formant bandwidths, source function, and "higher pole correction". This inventory will continually be supplemented and revised as our knowledge of speech spectra grows.

Speech Transmission Laboratory
Royal Institute of Technology
Stockholm

LITERATURE

(1) Fant, G., *Acoustic Theory of Speech Production* (Mouton & Co., 's-Gravenhage, 1960), 323 pp.
(2) Fant, G., "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies", *For Roman Jakobson* (Mouton & Co., 's-Gravenhage, 1956), 109–120.
(3) Ladefoged, P., *The Perception of Vowel Sounds*, Ph.D. Thesis, University of Edinburgh (Edinburgh, 1959).
(4) Fant, G., "Acoustic Analysis and Synthesis of Speech with Applications to Swedish", *Ericsson Technics*, 15, No. 1 (1959), 3–108.
(5) Miller, R. L., "Auditory Tests with Synthetic Vowels", *J. Acoust. Soc. Am.* 25 (1953), 114–121.
(6) Potter, R. K., and Steinberg, J. C., "Toward the Specification of Speech", *J. Acoust. Soc. Am.*, 22 (1950), 807–820.
(7) Flanagan, J. L., "Some Influences of the Glottal Wave upon Vowel Quality", invited paper presented at the Fourth International Congress of Phonetic Sciences, Helsinki, September 4–9, 1961; published in this volume.
(8) Brady, P. T., "Perception of Sounds Generated by Time-Variant Resonant Circuits", Massachusetts Institute of Technology, Research Laboratory of Electronics, *Quarterly Progress Report* No. 58, July 15, 1960, 218–220; and MIT, RLE, *Quarterly Progress Report* No. 59, October 15, 1960, 134–138.
(9) Speech Transmission Laboratory, Royal Institute of Technology, (Stockholm), *Quarterly Progress and Status Report*, 1/1960, October 15, 1960, 11–13.
(10) Flanagan, J. L., "Difference Limen for Vowel Formant Frequency", *J. Acoust. Soc. Am.*, 27 (1955), 613–617.
(11) Flanagan, J. L., "Estimates of the Maximum Precision Necessary in Quantizing Certain Dimensions of Vowel Sounds", *J. Acoust. Soc. Am.* 29 (1957), 533–534.
(12) Peterson, G. E., "Vowel Formant Measurements", *J. of Speech and Hearing Research* 2 (1959), 173–183.

DISCUSSION

Eli Fischer-Jørgensen:

Mr. Lindblom takes it for granted that what we want to find by formant measurements are the pole frequencies (or resonance peaks of the vocal tract). Most phoneticians in measuring formant frequencies will however probably try to find the peaks of the physical spectrum (e.g. of a section taken on the sonagraph). There will however, according to Gunnar Fant, be some differences between these peaks and the pole frequencies.

(1) The falling slope of the larynx spectrum will have the effect that the peaks of the section will be slightly lower than the poles, and this is especially true of the lower poles,

(2) When two poles are close together, the overlapping of their curves will cause the peaks of the resultant physical pattern to come still closer together. This is also true of a low F1 peak in relation to its negative pole. This means that one should make slight corrections for low F1 peaks and for close formant peaks.

It is evident that these corrections should be made if the aim is to find the relations between formant peaks and the configuration of the vocal tract. It is less evident, if the aim is to look for relations between physical stimulus and auditory reaction. –

But it is of course unpractical to have two different notions of formant frequency, and it is possible that the inaccuracy is less disturbing for the physical-auditory relation than for the physiological-physical relation. Both Gunnar Fant and Gordon Peterson seem to prefer identifying formant frequency with pole frequency. At any rate it is important that we all measure according to the same principles. It would be very useful to get practical directions for the corrections to be made.

In Reply:

It is important, I think, that we remain faithful to the mathematical model developed for the acoustic description of vowel sounds whatever the purpose of our investigations. Thus, in the model, there are not "two different notions of formant frequency". Formant frequency is synonymous only with pole frequency. The frequency of a pole alone, gives us the main shape of the corresponding elementary resonance curve since bandwidths are fairly predictable. Several pole frequencies enable us to reconstruct the spectral envelope. Once the spectral envelope and the fundamental frequency are known the relative amplitudes and the frequencies of the individual harmonics are uniquely specified. In other words, we can substitute the lowest four pole frequencies for the amplitudes and frequencies of a large number of partials. A formant frequency measurement is a pole frequency measurement.

Now, where do we find the pole on our spectrographic record? Our best policy is to make an attempt at estimating the frequency of the corresponding envelope peak. As Eli Fischer-Jørgensen points out there may be differences in frequency between a pole and its peak. Even when such factors as those mentioned by Eli Fischer-Jørgensen are taken into account, e.g., the slope of the source function and the distance between two close resonances, it is likely that we should hardly ever meet with differences that exceed 5–10 c/s. To all intents and purposes the number of c/s that an envelope peak deviates from its pole may be regarded as negligible. In comparison with the accuracy that we can achieve in locating peaks this figure is usually extremely small and the answer to our question is that we find the pole at the envelope peak. It should be noted that the frequency of the envelope peak is not the same as the centre of gravity of the spectral maximum or formant. As was said earlier asymmetry in formants renders this measure inaccurate. Thus the world of phonetics is not divided into those who content themselves with envelope peaks and those who, being more exacting, measure pole frequencies. We all make pole frequency measurements when trying to estimate the location of envelope peaks.

So far our discussion has treated of ideal cases i.e., non-aspirative, non-nasalized vowels without anomalies in the larynx spectrum. To complete the picture it should be added that, whereas a well-defined envelope peak is usually indicative of a pole, the pole is not always manifested as a peak in the transfer function; zeros in the vicinity and a large bandwidth may spoil the peakedness of a resonance. Instead of

a peak the envelope just has a plateau. A formant manifests a transfer function pole as a *potential* spectral energy maximum. The success of pole-zero matching techniques shows however, that, in spite of these complications, the model is basically sound.

In discussing physical-auditory relations we should think of our vowel sound as a complex stimulus in which, in ideal cases, the distribution of energy is most simply described by the lowest four poles and the fundamental frequency component. Considering what is known so far about auditory processes it is not justifiable to exchange the pole specification for a description that considers arbitrarily the "formants" in isolation. *Not the formants but the entire spectrum constitutes our stimulus.*

To sum up, the procedure of formant frequency measurements is theoretically a search for pole frequencies. In practice poles and peaks coincide. Thus good accuracy is obtained if we succeed in finding envelope peaks. As a help to finding envelope peaks I suggest the use of standard envelopes mathematically derived as mentioned in my paper.