Int. J. Man–Machine Studies (1972) **4**, 411–424

Formants in Automatic Speech Recognition

DAVID J. BROAD

Speech Communications Research Laboratory, Inc.,
35 West Micheltorena Street,
Santa Barbara, California 93101, U.S.A.

(*Received 4 July 1972*)

This paper concerns the use of formant frequency information in automatic speech recognition. The discussion is addressed to the physical significance of the formant and to how this relates to the phonetic concepts of *segment* and *equivalence* that are needed for the recognition of phonetic types. Specifically, the definition of the phone in terms of articulatory dynamics can be interpreted acoustically in terms of formant dynamics. Hence formant transition information can aid segmentation. Also, formant frequencies for given utterances by single speakers display remarkable interrepetition stability, while the speaker identity, phonetic type, and the phonetic, prosodic, and linguistic contexts are sources of non-random variability that should be included in a complete acoustic phonetic description of formant behavior.

Introduction

A formant is a damped sinusoidal component of the acoustic impulse response of the human vocal tract, i.e. a formant corresponds to a complex-conjugate pair of poles in the Laplace transform of the vocal tract response. Although frequency domain analyses such as sound spectrography have been used for measuring formant frequencies, it should be noted that the formant itself is not a frequency domain concept. The "frequency domain" is precisely the set of Fourier transforms of transformable time domain functions, and formant frequencies do not necessarily appear in the power spectrum of a speech signal, though they may determine its envelope. The frequency of the formant is its frequency of oscillation, and this frequency has no necessary reference to a frequency-domain representation of the speech signal. This point has sometimes led to confusion, even to the extent that the shortcomings of spectral analysis have been considered necessarily to be shortcomings of the formant concept as well. Formants had long been associated with the harmonic structure of speech, but early in the present century Scripture (1906) had already advanced the non-harmonic description of

formants in terms of damped sinusoids by means of extensive hand computations on waveforms measured from phonograph records.

Large scale acoustic phonetic studies using the sound spectrograph (Koenig, Dunn & Lacy, 1946) and related devices documented the characteristic formant structures of many speech sounds and prompted the hope that formant analysis would provide rapid solutions to various problems of speech technology, including visible speech, speech synthesis, speech transmission, and automatic speech recognition.

It was soon found, however, that the most obvious difficulty with any system based on formant analysis was the excessive difficulty of obtaining good automatic estimates of the formant frequencies. Gross misidentifications in the cases of closely spaced formants and high fundamental frequencies have been a special plague. Recent improvements in formant analysis, however, suggest that this obstacle can be overcome.

For automatic speech recognition (ASR) based on formant analysis a second major problem exists, namely, the formant patterns of speech display great variability as a function of speaker identity, phonetic context, and random variation. Not only are the patterns for given sounds variable, but the formants are in continuous motion and a basic problem is to specify how the continuous formant contours of voiced speech can be interpreted as sequences of discrete phonetic elements.

These difficulties have often led (or forced) workers in ASR to abandon formants in favor of parameters that are easier to measure and (hopefully) easier to interpret. Thus one school of thought holds that the formant is an obsolete or, at best, unworkable concept for recognition. The opposite school of thought, as advanced here, holds that the behavior of formants is complex only because speech is that way and that the formant provides one of the best ways we have for characterizing and understanding the structure of the speech signal. This idea stems from the importance of phonetic concepts in describing both the physical and the linguistic aspects of speech.

Phonetic Concepts are Basic to ASR

Speech is produced by sequences of co-ordinated movements that control the breath stream and its resulting sounds. These sounds, the consonants and vowels of speech, can be classified according to how they are produced by means of such physiological parameters as place of articulation, manner of articulation, lip shape, and tongue shape. Physiological phonetics undertakes the systematic description of all speech sounds through reference to the gestures that produce them. Phonetic notations based on physiology have an

exceedingly long history (consider that many ancient alphabets have a phonetic basis) and in their modern forms have proved to be useful tools for the investigation of the sound systems of the earth's spoken languages. The success of physiological phonetic systems supports the following double hypothesis about human speech:

(1) speech can be accurately described as a sequence of phonetic units, or *phones*, which can be defined by observable physical events;
(2) messages in a language are encoded as sequences of phones so that in appropriate contexts two utterances that are phonetically equivalent will also be semantically equivalent.

Thus a phonetic notation can be viewed simultaneously as a reference to a set of physiological events in the speaker's vocal mechanism and as an abstract element of a linguistic code structure. The notation [it^{h}], for example, refers to a voiced high-front unrounded oral vowel followed by a voiceless alveolar aspirated plosive; at the same time it can refer to the lexical item it in English. The recognition of phonetic types is a significant subproblem in ASR. Some discussion of phonetic concepts in ASR was given previously (Board, 1972). Briefly, it is claimed that:

(1) phonetic transcriptions represent the speech signal efficiently while preserving the linguistically significant information. Implicit in this assertion is the assumption that some satisfactory way of presenting the prosodic information (i.e. fundamental voice frequency, intensity, and duration) is also included;
(2) phonetic representations are amenable to further linguistic processing such as the application of assimilation rules. Hence the variations in pronunciation due to context, speed of talking, etc, that an ASR system must handle can be described using phonetic notations;
(3) the physiological basis for phonetic transcriptions provides a link between the physical events of speech and the elements of the phonetic alphabet. Since an ASR system must at some stage interpret sequences of physical events as sequences of linguistic events, such a link is indispensable for ASR. Obviously, some knowledge of how articulatory events map into acoustic events is also required for physiological phonetics to be useful for ASR.

Several correspondences exist between physiological events and acoustic events in speech. Some simple relations include the correspondence between vocal fold vibration and a quasi-periodicity in the acoustic waveform and the correspondence between a plosive release and an acoustic burst. A most important relation is the one between formant frequencies and vocal tract shapes for voiced sounds produced with an open oral air path and closed

nasal air path (non-nasalized sounds). The vowels and the sonorant consonants are normally produced this way.

One of the basic problems in phonetic theory is the explication of the phone in terms of the dynamics of speech production. If ASR is to use the detection of phonetic units, then this basic question of phonetics becomes a practical concern. The situation is by no means simple, and naive conceptions of how phonetic segments occur in speech will not lead to satisfactory segmentation. In particular, phones do *not* in general occur as simple sequences of steady states concatenated through transitions. Some sounds, such as the glides [e^I] and [o^U], for example, are characterized by slow dynamic shifts and may achieve steady states only late in their production, or not at all. A task of acoustic phonetics is to specify how the articulatory events that signal the occurrence of phones may be detected in the acoustic speech waveform. To achieve this it is useful to model the speech waveform so that (1) its acoustic properties are accounted for, and (2) its relation to articulation can be specified. Speech synthesis provides one way of checking the first of these desiderata and an analysis of the physics of speech production is one way of checking the second.

Formants are Basic to Acoustic Phonetics

The relation between formant frequencies and vocal tract shapes has been known in an empirical way for some time. Joos (1948) and Potter & Peterson (1948) were some of the first to note the similarity between the distribution of vowels in the plane determined by the first two formant frequencies and the classical positions of the vowels on charts determined by horizontal and vertical places of articulation. Fant (1960) showed that the formant frequencies for vowels can be predicted accurately from the vocal tract shape as determined by X-rays. Schroeder (1967) then demonstrated a one-to-one correspondence between formant frequencies and the terms of the Fourier cosine series representation of the logarithmic normalized vocal tract area function. The vocal tract is modeled as an acoustic tube of length L whose cross-sectional area at x cm from the glottis is $A(x, t)$ at time t. Then the logarithmic area function is represented as the sum of its symmetric and antisymmetric components

$$\ln A(x, t) = \ln A_s(x, t) + \ln A_a(x, t)\ 0 \leqslant x \leqslant L \qquad ..(1)$$

where the symmetry relations are

$$A_s(x, t) = A_s(L-x, t)$$
$$A_a(x, t) = -A_a(L-x, t).$$

Then Schroeder's result can be expressed

$$\ln A_a(x, t) = -2 \sum_{n=1}^{\infty} \frac{\Delta F_n(t)}{F_{n0}} \cos [(2n-1)\pi x/L] \qquad \ldots(2)$$

where F_{n0} is the nth resonance frequency of a uniform tract of length L (i.e. $F_{n0} = (2n-1)c/4L$, where c is the speed of sound), and $\Delta F_n(t) = F_n(t) - F_{n0}$ where $F_n(t)$ is the value of the nth formant frequency at time t. According to Schroeder's derivation, the symmetric component $\ln A_s(x, t)$ is to a first-order approximation determined by the input impedance at the lips and is not sensitive to changes in the formant frequencies. Hence the formant frequency information apparently does not completely specify the vocal tract shape. Atal (1970) has discussed a possible way of overcoming this if formant bandwidth information is also available. Of course the relation just mentioned gives only a partial acoustic specification of articulation. First, the symmetric component $A_s(x, t)$ is omitted. Second, even if the entire area function $A(x, t)$ were known, there remains a final transformation to obtain the shapes and positions of the articulators. Lindbolm & Sundberg (1971) propose a model for deriving vocal tract shapes from a characterization of the articulators. This is a step forward solving the inverse problem: given a tract shape, where are the articulators?

Although the relation between formant frequencies and articulation is at present only incompletely known, what we do know is sufficient to begin the pursuit of some questions that are basic to recognition, particularly the question of segmentation.

The Formant and Segmentation

The speech wave can be segmented according to its basic speech wave types, e.g. quasi-periodic, quasi-random, or quiescent waveforms. For example, the word *fee*, denoted [fi] phonetically, is segmentable into an initial quasi-random waveform followed by a quasi-periodic waveform. Sometimes, however, a single phone consists of a succession of more than one wave type, e.g. the voiceless aspirated plosive [p^h] is a succession of quiescent, transient, and quasi-random waveforms. The more difficult case for phonetic segmentation occurs when two or more successive phones have the same wave type. The most common example of this situation is the sequence consisting of two or more voiced phones, such as the utterance we were away, denoted [w ɪ w ɹ ə w e ɪ] phonetically, which has a quasi-periodic waveform throughout all its seven phones.

Figure 1 shows a plot of the first four formant frequencies for this utterance. The problem is to deduce the segmentation from this information. While some of the segments are fairly obvious, such as the [ɪ] and the [ɹ], the total segmentation is not at all clear. To illustrate this point, consider a naive segmentation derived from the major extrema in the formant contours. By this method we obtain, moving from left to right, the initial [w] from the min (F_2), the [ɪ] from max (F_2), the next [w] from min(F_2), the [ɹ] from min(F_3) and max(F_2), the third [w] from min(F_2) and min (F_1), and the terminal [ɪ] of the glide [e$^{\text{ɪ}}$] from max(F_2) and min(F_1). This criterion works

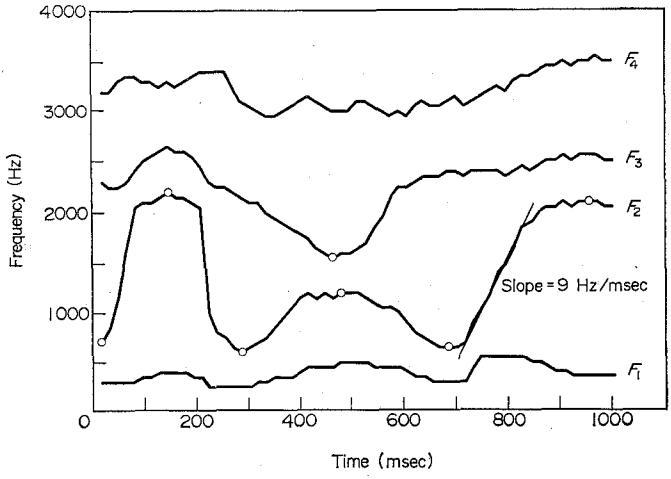

FIG. 1. The first four formant frequencies for the utterance we were away.

well in the sense that everything it detects is a segment (or a significant part of a segment in the case of the glide [e$^{\text{ɪ}}$], but its shortcoming is the events it overlooks, i.e., the initial vowel of *away* and the [e] of the final glide.

To refine the segmentation, let us review some concepts from phonetic theory, which have been formalized in detail by Peterson & Shoup (1966). These authors distinguish two types of *articulatory states*. The first, which is intuitively the more familiar type, is the articulatory *steady state* which involves a "minimum in the absolute magnitude of the average rate of change of the position and shape of a supraglottal articulator . . .". According to this definition it is not necessary for an articulator to pass through a stationary point to be in a steady state; it is only necessary for it to slow down between intervals of faster movement (the transitions). The second type of articulatory state is the *controlled articulatory movement*, which is

the movement of an articulator that is "slow relative to most of the movements of the articulator and in which the average change . . . is relatively constant and regular throughout". To relate these concepts to the acoustic domain, we may use equation (2) to arrive at estimates of the average rate of change of an articulatory configuration. A number of approaches are possible since the physiological phonetic theory does not explicitly define the "position and shape" function for an articulator. One possibility that leads to a simple relation is to differentiate equation (2) with respect to t, square both sides, and integrate with respect to x over the length L of the vocal tract:

$$V^2 = \int_0^L \left[\frac{\partial}{\partial t} \ln A(x, t)/A_0\right]^2 dx = \frac{32L^3}{c^2} \sum_{n=1}^{\infty} \left[\frac{1}{2n-1} \frac{dF_n}{dt}\right]^2 \quad \ldots(3)$$

where c is the speed of sound and A_0 is the average area over the length of the tract. This function is plotted in Fig. 2 for the utterance we were away. Only the first four formants were used in the summation.

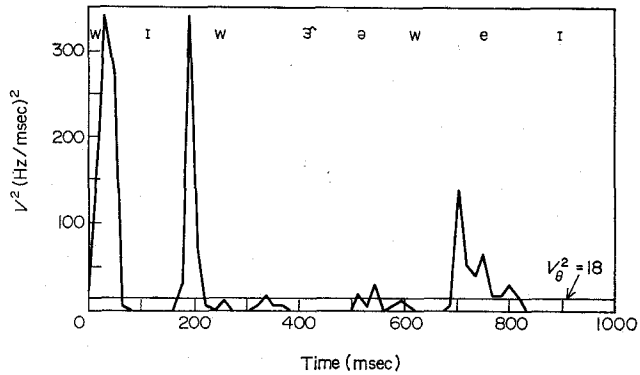

FIG. 2. The average weighted square formant velocity $V^2(t)$ for the utterance we were away of Fig. 1.

Somewhat arbitrarily setting an *ad hoc* threshold of $V^2 = 18$ Hz^2/msec^2, and defining an interval where $V^2 > 18$ as a transition results in the set of six transitions shown. The first five of these are less than 50 msec in duration while the sixth lasts for about 125 msec. It seems reasonable to consider this latter transition as the result of a controlled articulatory movement in the sense of the definition mentioned above. To illustrate the fairly regular change in this transition, Fig. 1 includes a straight line of slope 9 Hz/msec that fits the second formant transition quite well over the interval. In the Peterson & Shoup phonetic theory, a phone may be defined around either an articulatory steady state or around a controlled articulatory movement.

This formulation permits the [e^{I}] glide that we want to identify as a phone to be distinguished from the ordinary transitions between phones. The study of transitions, glides, and diphthongs by Lehiste & Peterson (1961) is informative in this context.

It is also of interest that the above criterion for defining a transition detects two transitions between the [ɹ] and the third [w]. It would be pleasing to claim that the transitions into and out of the intermediate unstressed vowel have been detected by this means. A slightly different analysis sheds more light on this question.

Equation (3) is a measure of the average acoustically detectable movement throughout the vocal tract, but the definitions of articulatory dynamics in the phonetic theory refer to specific articulators. For example, it is possible for the lips to be in an articulatory steady state while simultaneously the tongue engages in an articulatory transition. This suggests a formulation different from equation (3), namely, that equation (2) should be differentiated with respect to time directly and the quantity.

$$V_x(t) = \frac{\partial}{\partial t} \ln A_a(x, t)/A_0 \qquad \ldots(4)$$

should be examined for various values of x, e.g. $V_{17\ cm}(t)$ could be an estimate of the absolute rate of change of the lip opening. Figure 3 shows plots of $V_{9\ cm}(t)$ and $V_{17\ cm}(t)$ for the utterance we were away. For the most part these curves give the same information as given in Fig. 2. The interesting exception is that in Fig. 3 the region around the initial vowel of away shows $V_{9\ cm}$ detecting a transition at $t = 570$ msec while $V_{17\ cm}$ is simultaneously near zero, while 65 msec later, $V_{9\ cm}$ is quite small and $V_{17\ cm}$ is detecting a transition. This result is seen in Fig. 1 in the fact that the rise in F_3 out of the [ɹ] is essentially complete by approximately 60 or 70 msec before F_2 completes its descent into the [w]. A somewhat similar segmentation results from considering separately the F_2 and F_3 transitions. The significance of this result is that these transitions permit the detection of the initial segment of away even though it would be difficult to isolate any target configuration in this interval by a casual examination of Fig. 1. Physiologically the vowel apparently occurs between the gestures for releasing the retroflexion of the tongue apex for the [ɹ] and for initiating the lip rounding and velarization for the [w].

The formant velocities discussed here are similar to the *spectral derivative* discussed by Rabiner, Schafer & Flanagan (1971) in connection with rules for the generation of transitions in a formant synthesizer. It is not suggested that the above methods provide ideal automatic segmentation criteria. For

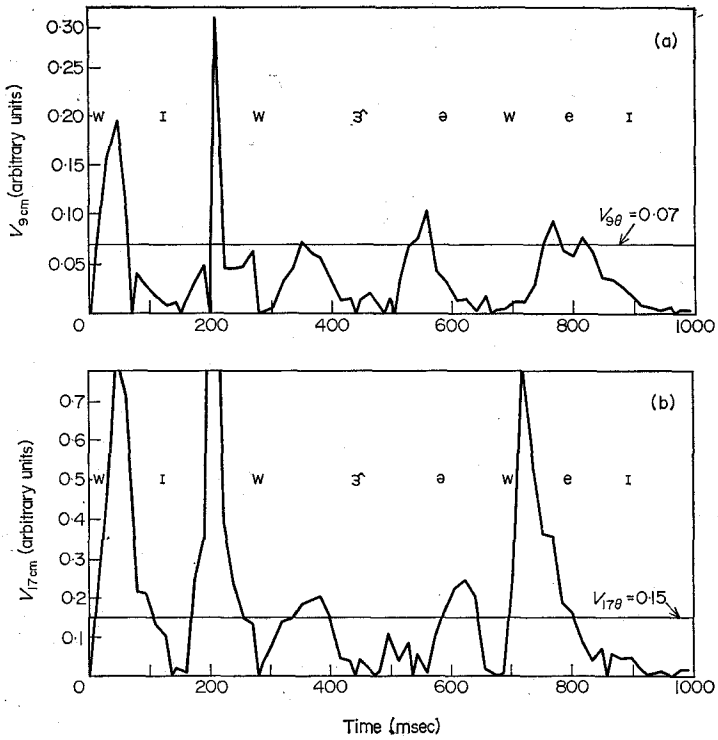

FIG. 3. The absolute weighted formant speeds corresponding to the vocal tract area changes at 9 cm from the glottis (top graph) and at 17 cm from the glottis (bottom graph) for the utterance of Fig. 1.

example, we might expect an improvement if the formant contours were smoothed prior to differentiation. The above examples do suggest, however, that basic phonetic concepts that define the segment in the articulatory domain can guide the search for useful segmentation criteria by reference to the intimate connection between formant frequencies and articulation.

The Formant and Phonetic Equivalence

Following the division of the speech wave into segments, the identification of the phonetic type of each segment remains. What is the "correct" phonetic transcription of an utterance? Is it the concatenation of the dictionary pronunciations of the word sequence? Quite obviously not, because of the many phonetic variations that are possible for an utterance. Is the transcrip-

tion, then, to be defined by a phonetician's response to the utterance? Here we have the problem that the phonetician may be influenced by his native phonemic system, by what he expects to hear, etc. These questions may be restated: When are two phones to be considered phonetically equivalent? Peterson & Barney (1952) address this question through reference to speaker intentions and listener identifications of words. Peterson (1952) then extended the study by referring to speakers' imitations of standard sounds. Peterson & Shoup (1966) define phonetic equivalence in terms of specific physiological parameter values, but sufficient physiological records of utterances for a direct objective check on transcriptions are exceedingly difficult, if not impossible, to obtain in significant quantity.

To illustrate the problems raised by the above observations, consider again the utterance we were away. If the sequence of "steady states" is observed, it is found that the first and second formant frequencies always fall in identifiable areas of the F_1F_2 plane plotted by Peterson & Barney. If the positions of the steady states are taken to "recognize" the segments, a transcription of the form [u ɪ u ʊ ʊ u ɪ] results. By addition of third formant information, the first [ʊ] could doubtless be corrected to [ɪ]. Also, since extended experience with sound spectrograms shows that [w] usually has a much weaker third formant than [u], it might be possible to effect the somewhat better transcription [w ɪ w ɹ ʊ w ɪ]. Also, by taking into account the glide that was identified in the preceding section, the final [ɪ] might be corrected to [e^{ɪ}] or [ɛ^{ɪ}]. Nevertheless, certain difficulties remain. First, even if context is taken into account, the vowel of we is almost certainly identified as [ɪ] and not the [i] we might expect or, for that matter, that we would probably hear if asked to transcribe the utterance. This problem would apparently have to be handled linguistically as, e.g. specifying that [wɪ] can sometimes be a permissible pronunciation of we. The 64-msec duration of the initial vowel of away, as defined by the segmentation of the preceding section, is quite short, and could on this basis be identified as an unstressed vowel. That is, since /ə/ is always unstressed, it may be distinguished prosodically even if the phonetic value, e.g. [ʊ], is not so close to that of [ə]. These considerations illustrate two points:

(1) formant frequency information is not the exclusive determiner of phonetic value and other parameters, such as segment duration and formant amplitudes, will undoubtedly be required for full phonetic recognition;
(2) phonemes may frequently be realized by allophones that are not phonetically equivalent to their canonical allophones.

Thus, even if perfect phonetic recognition is possible, the conversion to meaningful word sequences is not to be accomplished in general by a simple "dictionary" matching between input phonetic strings and pronunciation entries.

The above considerations should establish that the phonetic interpretation of formant patterns is not completely trivial. Indeed, the impression may even be one of an embarrassing quagmire. It is instructive at this point to consider some of the experimental evidence that is available about the sources

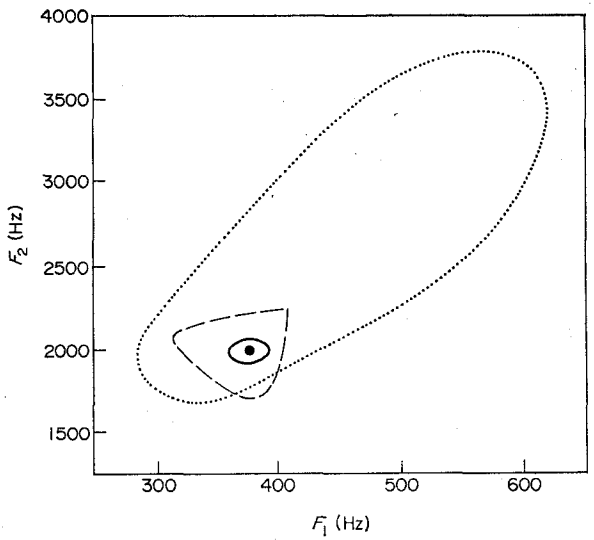

FIG. 4. The [ɪ] region of the F_1F_2 plane. · · · · · Peterson–Barney region for [ɪ] in context [h_d]; —— Broad–Fertig location for [ɪ] in context [h_d] including 1 σ ellipse; – – – Broad–Fertig region for [ɪ] in all contexts.

of variability in formant patterns and attempt to draw from inferences regarding the possibilities for their interpretation.

Figure 4 shows a portion of the plane defined by the first and second formant frequencies as axes. The large loop is the region that Peterson & Barney found to enclose their data for the vowel [ɪ] pronounced by 76 speakers, including men, women, and children, in the phonetic context [hɪd] (the word hid). Generally, owing to differences in vocal tract size, the children's [ɪ]'s tend to occupy the upper right end of the region, the men's the lower left, and the women's to occupy an intermediate position. Broad & Fertig (1970) measured the formant frequencies for the same vowel [ɪ] produced in 576 different consonantal contexts by a single male speaker.

The dot in Fig. 4 shows the average position in the F_1F_2 plane for the Peterson & Barney context [hɪd] in their study. Since three repetitions of each context were measured, information on the intrinsic variability of utterances was also available. Thus the small ellipse around the Broad–Fertig [hɪd] defines the one-standard-deviation region for these data.

The dashed curve encloses the space measured for all 576 consonantal contexts. The region for all contexts is almost a proper subset of the Peterson–Barney region, but it is easy to imagine this happy circumstance as being

TABLE 1

Inter-repetition standard deviations for the first three formant frequencies reported from various sources in the literature

	s.d. (Hz)		
Source	F_1	F_2	F_3
Potter & Steinberg (1950) [Sustained vowels]	20–40	40–70	60–90
Öhman (1966) [VCV utterances in Swedish]	35	75	70
Broad & Fertig (1970) [C/ɪ/C syllables]	16–22	63–86	76–86
	Difference limen (Hz)		
Flanagan (1955) [Synthesized vowels]	12–27	20–90	—

different if some other speaker whose production of hid was closer to the border had been chosen. Nevertheless, Fig. 4 shows that at least for the particular case studied the variability attributable to speaker identity is greater than that attributable to consonantal context which, in turn, is greater than that attributable to random variation. The phonetic value of the vowel itself dominates all these effects.

The magnitude of the random variation merits some further remarks. Table 1 shows the values obtained by various investigators for the standard deviations of the first three formant frequencies for repetitions of the same

utterance by the same speaker. The values reported in the different studies are all comparable to each other. These values are also of the same order as those reported by Flanagan (1955) for the difference limens for human perception of formant frequencies. These are shown on the bottom line of Table 1. It is also interesting that all the values in Table 1 are of the same order or less than the typical corresponding formant bandwidths. Hence individual speakers are evidently very consistent in the exact formant patterns they produce for given utterances; we may even hypothesize that in repeating the same utterance a speaker will a majority of the time reproduce his behavior so closely that no perceptible difference in formant frequencies results. We may also infer that only a small part of the total variability of formant patterns is attributable to true random variation in a speaker's behavior and that by far most of the variability can be accounted for by the phonetic type of the segment, the speaker's individual characteristics, and the phonetic and linguistic context, in about this order.

It also is worth noting that the known values for intrinsic variation in formant patterns give a criterion for checking the adequacy of descriptive models for formant patterns. When there exists a discrepancy between an expected formant frequency and the actual formant frequency that exceeds a few standard deviations, for example, there is probably some regular individual, phonetic, or linguistic effect that is not being taken into account. So far the formant patterns in only very limited combinations of speaker: phone type : context : speaking situation have been studied. Although a given speaker exhibits remarkable consistency, there is much to learn about defining phonetic equivalences between speakers. The results of Klein, Plomp & Pols (1970) indicate that to a large extent vowel equivalence between speakers can be described by linear transformations of the F_1F_2 plane. The description of the formant patterns for even a single speaker, however, remains a major problem for research. Ultimately, it is desired that the formant frequency contours should be described as functions of the current segment, the preceding and following segments, the rate of speaking, and the prosodic environment.

The Directorate of Mathematical and Information Sciences of the United States Air Force Office of Scientific Research (AFSC) supported this research under contract F44620-69-C-0078.

References

ATAL, B. S. (1970). Determination of the vocal tract shape directly from the speech wave. *J. acoust. Soc. Am.*, **47**, 65.

BROAD, D. J. (1972). Basic directions in automatic speech recognition. *Int. J. Man–Machine Studies*, **4**, 105.

BROAD, D. J. & FERTIG, R. H. (1970). Formant-frequency trajectories in selected CVC syllable nuclei. *J. acoust. Soc. Am.*, **47**, 1572.

FANT, G. (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

FLANAGAN, J. L. (1955). A difference limen for vowel formant frequency. *J. acoust. Soc. Am.*, **27**, 613.

JOOS, M. (1948). Acoustic phonetics. *Language, Monogr. Suppl.*, **24.**

KLEIN, W., PLOMP, R. & POLS, L. C. W. (1970). Vowel spectra, vowel spaces, and vowel identification. *J. acoust. Soc. Am.*, **48**, 999.

KOENIG, W., DUNN, H. K. & LACY, L. Y. (1946). The sound spectrograph. *J. acoust. Soc. Am.*, **18**, 19.

LEHISTE, I. & PETERSON, G. E. (1961). Transitions, glides, and diphthongs. *J. acoust. Soc. Am.*, **33**, 268.

LINDBLOM, B. E. F. & SUNDBERG, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *J. acoust. Soc. Am.*, **50**, 1166.

ÖHMAN, S. E. G. (1966). Coarticulation in VCV utterances: spectrographic measurements. *J. acoust. Soc. Am.*, **39**, 151.

PETERSON, G. E. (1952). The information bearing elements of speech. *J. acoust. Soc. Am.*, **24**, 629.

PETERSON, G. E. & BARNEY, H. L. (1952). Control methods used in a study of the vowels. *J. acoust. Soc. Am.*, **24**, 175.

PETERSON, G. E. & SHOUP, J. E. (1966). A physiological theory of phonetics. *J. Speech Hear. Res.*, **9**, 5.

POTTER, R. & PETERSON, G. E. (1948). The representation of vowels and their movements. *J. acoust. Soc. Am.*, **20**, 528.

POTTER, R. K. & STEINBERG, J. C. (1950). Toward the specification of speech. *J. acoust. Soc. Am.*, **22**, 807.

RABINER, L. R., SCHAFER, R. W. & FLANAGAN, J. L. (1971). Computer synthesis of speech by concatenation of formant-coded words. *Bell. Sys. Tech. J.*, **50**, 1541.

SCHROEDER, M. R. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *J. acoust. Soc. Am.*, **41**, 1002.

SCRIPTURE, E. W. (1906). *Researches in Experimental Phonetics: The Study of Speech Curves.* Washington, D.C.: Carnegie Institution of Washington.