Published in final edited form as:

Phonetica. 1981 ; 37(5-6): 285–305.

Variability of Vowel Formant Frequencies and the Quantal Theory of Speech: A First Report

David B. Pisoni

Indiana University, Bloomington, Ind. USA

Abstract

This paper reports the results of a study in which variability of formant frequencies for different vowels was examined with regard to several predictions derived from the quantal theory of speech. Two subjects were required to reproduce eight different steady-state synthetic vowels which were presented repeatedly in a randomized order. Spectral analysis was carried out on the vocal responses in order to obtain means and standard deviations of the vowel formant frequencies. In the spirit of the quantal theory, it was predicted that the points vowel, /i/, /a/ and /u/ would show lower standard deviations than the nonpoint vowels because these vowels are assumed to be produced at places in the vocal tract where small perturbations in articulation produce only minimal changes in the resulting formant frequencies. That is, these vowels are assumed to be quantal vowels. The results of this study provided little support for the hypothesis under consideration. A discussion of the outcome of the results as well as some speculation as to its failure to find support for the quantal theory is provided in the report. Several final comments are also offered about computer simulation studies of speech production and the need for additional empirical studies on vowel production with real talkers.

Introduction

Within the last few years considerable effort has been devoted to establishing well-defined anatomical, physiological and perceptual correlates for distinctive phonetic features in speech as well as attempting to rationalize various phonological processes that occur with some regularity in natural languages. Particular emphasis has been placed on the articulatory space used to form phonetic segments and features and the resulting acoustic attributes of speech sounds generated by these configurations [see Lindblom, 1972].

Although the vocal apparatus can theoretically assume a relatively large number of articulatory positions, only a small number of preferred or 'natural regions' of the articulatory space are actually used in phonological systems of languages. In an influential paper, Stevens [1972] has proposed that speech sounds produced in these so-called natural regions have certain 'quantal' properties or attributes that seem to be good candidates for the inventory of phonetic features used in language. According to Stevens, all phonetic features that occur in languages probably have their basis in acoustic attributes that have such quantal properties: perturbations in articulation at one of these natural regions produce only small changes in the acoustic output. Stevens also argues that these acoustic attributes are well-matched to the properties of the auditory system and therefore have some special status perceptually.

The basic line of reasoning behind Stevens quantal theory of speech production can be illustrated by reference to figure 1. Assume that one could manipulate some articulatory parameter continuously along a particular dimension. And further assume that one could obtain a measure of some specific acoustic parameter of the speech signal that would be controlled by changes in this articulatory parameter. As the articulatory parameter is varied

continuously one might expect to find continuous changes in the output of the acoustic parameter. However, what seems to happen is shown schematically in figure 1. There are places in this space where very small changes in the articulatory parameter, as in region II, produce large variations in the acoustic parameter whereas in other places, as in region III, large changes in the articulatory parameter produce only small changes in the acoustic output. Relations such as these between vocal tract shape and sound output have been studied in some detail by Stevens [1972, 1973] and others with the use of computer simulation models of the vocal tract [see also Lindblom and Sundberg, 1969; Liljencrants and Lindblom, 1972]. Small changes have been made systematically in one or more articulatory parameters and the resultant effects on the properties of sound output calculated quite precisely.

To see how Stevens carried out the simulation and arrived at these proposals, it will be helpful to briefly examine the resonance characteristics of the vocal tract. Figure 2 shows an approximation of the vocal tract shape for the vowel [a] in terms of a two-tube resonator [Fant, 1960]. The spectrum envelope produced by this configuration is shown below in the lower panel. The size of the pharyngeal cavity is constricted when compared with the size of the oral cavity. Using a vocal tract simulation model which calculates formant frequencies from area functions, Stevens found that over a wide range, variations in the lengths of the cavities (i.e., d_1 and d_2) and the cross-sectional areas (i.e., A_1 and A_2) did not affect the formant frequencies of the vowel to any large degree. The results of this simulation are shown in figure 3, where the calculated formant frequencies are plotted as a function of the length of d_1 . Notice that there is an area in the middle of this figure where variations in d_1 produce only small changes in the formant frequencies of F_1 and F_2 . On the other hand, there are regions at the extremes where a small variation in d_1 produces a much larger shift in the formant frequencies, especially the values of the second formant.

Stevens [1972, 1973] has employed these simulation techniques to study other vowels as well as several consonants and he concludes from these studies that there are certain places in the vocal tract which show quantal relations between articulation and the attributes of the sound output. Perturbations in articulation at these places apparently have only small effects on the acoustic signal and therefore, according to Stevens, would presumably affect perception only minimally.

One of the proposals that Stevens has made about these findings is that the sound segments used in languages are selected from a relatively small range of features or distinctive attributes. A careful examination of these acoustic attributes has revealed that they have a natural basis in terms of certain vocal tract configurations. These configurations occur in precisely those regions where variability in articulation can be tolerated without the sound output being appreciably affected.

Using similar types of simulation procedures, Liljencrants and Lindblom [1972] have studied the vowel systems of a number of languages in order to predict their phonetic structure. Since human vowels represent only a small subset of the possible combinations of formant frequencies, it was of interest to see how the vowels of different languages were distributed in acoustic space. Liljencrants and Lindblom found that the vowel systems of quite diverse languages can be described in terms of a simple principle of 'maximal perceptual contrast' as defined by the linear distance in mel units between the points representative of two vowels. The whole vowel space for a particular language seems to be organized in terms of sound contrasts that are highly discriminable. This is true of languages having as few as 3 vowels to as many as 12. These two sets of findings are important because they suggest that the sound systems of natural languages may be matched, in some sense, to attributes of both perception and production.

It should be noted here that not all of the potential distinctive features that exist at the phonetic level are used distinctively in the phonological system of a particular language. Similarly, no natural language has as many phonemes in its phonological system as there are logically possible combinations of distinctive features. Thus, the sound systems of languages may have evolved the way they are for the following two reasons. First, the distinctions used in language are easily pronounceable for talkers thus permitting some variability in production. Secondly, these same articulatory distinctions are used to generate acoustic attributes that have highly distinctive properties resulting in signals that can be identified and discriminated under very poor listening conditions.

A basic assumption of the quantal theory of speech is that there are nonlinear relationships between articulation and acoustic output. According to Stevens, as a result of these nonlinearities a speaker 'manipulates his speech generating mechanism to select sounds with well-defined acoustic attributes that are relatively insensitive to small perturbations in articulation' [Stevens, 1972, p. 65]. While the quantal theory has a certain amount of face validity to it, and obviously may have important implications for a number of areas such as phonological development in young infants, the articulatory and acoustic basis of distinctive features, and physiological constraints on coarticulation phenomena, very little empirical work has actually been directed towards testing its basic assumptions. This is not too surprising at the present time because Stevens has not offered any specific predictions from the theory. Moreover, much of the work on the theory has been done under very idealized conditions involving numerous assumptions in order to implement the simulation procedures.

One obvious set of predictions that would seem to follow directly from the spirit of the quantal theory of speech deals with the stability or precision of the articulatory gestures underlying the production of different speech sounds. The precision of these gestures presumably would be reflected in the variability of the measured formant frequencies in the acoustic domain [Fant, 1960]. In connection with vowels, there has been some discussion in recent years that the point vowels /i/, /a/ and /u/ may be thought of as stable reference points in both articulatory and acoustic space [see Gerstman, 1968; Lieberman, 1970, 1976]. Indeed, the present investigation was undertaken in the hope of learning something about the variability associated with the production of a number of steady-state vowels and the possible relations that might be found between the variances of the measured formant frequencies [see also Broad, 1976; Broad and Wakita, 1976].

The specific hypothesis under consideration in this report was that some vowels, such as the point vowels, may be produced with inherently more stable articulatory configurations than other vowels and this articulatory stability or precision should be reflected in the pattern of variability of the formant frequencies over successive repetitions of a set of vowels produced by the same speaker. The study reported below used a production paradigm in which the subject was presented with one of eight synthetically generated vowel targets and then was required to repeat the vowel during a specific response interval. Reaction time measures were obtained and formant frequencies were calculated for each response.

Before proceeding to the experiment proper it would be appropriate to review briefly some of the previous work dealing specifically with the production paradigm and the studies which lead up to the present inquiry. Several investigations have been carried out using similar procedures, although they were not concerned with the same questions under consideration here. Some years ago Chistovich *et al.* [1966] carried out a study on the imitation of a set of synthetic vowels in order to determine whether the internal representation of a vowel was a continuous motor representation or a discrete or finite set representation. A single subject was presented with 12 synthetic vowels that followed a

continuous trajectory in the F_1 , F_2 , F_3 pathway from /a/ to /ε/ to /i/ and was required to perform three different experimental tasks. In the first task, the subject was required to mimic the synthetic vowel as accurately as possible after listening to the whole stimulus. In the second task, the subject had to shadow the vowel with the shortest delay possible. In the third task, the subject had to identify the vowel by writing down an appropriate Russian letter for the vowel phoneme. This task was carried out with a metal plate and a special writing pen so that latencies could be obtained. Latencies were also obtained for the first two tasks by examination of Mingograph records. Analyses of the spectra of the response vowels produced in the first two tasks were carried out with a 51-channel filter bank spectrum analyzer so that estimates of the frequencies of the first two formants could be obtained.

In order to determine whether the internal representation for vowels was continuous or discrete, Chistovich et al. [1966] examined three characteristics of the spectral data obtained in the first two tasks. First, they sought to determine the relationship between the mean formant frequencies of the response vowels and the formant frequencies of the stimulus or target vowels. Evidence of any nonlinearities in the form of a 'step-like quantizing function' could be taken as support for the discrete or categorization hypothesis concerning the internal representation of these vowels. The second analysis of the spectral data involved an examination of the standard deviations of the formant frequencies of the response vowels. Under the categorization hypothesis, the standard deviations of the formant frequencies should be minimal at the center of the category as defined by the presence of the step-like functions observed earlier and should increase toward the extremes of the category. That is, there should be greater precision or stability of the responses within rather than between response categories. Finally, Chistovich and her colleagues examined the topography or distribution of the formant frequencies of the response vowels. The presence of pronounced peaks in the distribution reflecting modal response values at particular formant frequencies could also be interpreted as support for the discrete transformation view of the internal representation of vowels.

The results obtained by Chistovich et al. in the mimicry task for all three criteria favored the view that the internal representation of vowels is discrete and involves categorization whereby the instructions to the articulators have undergone some form of transformation resulting in a reduction of information. While there were some differences in the spectral data for the mimicry and shadowing tasks, the overall results showed that the observed perceptual categories do not appear to reflect the phonemes of the speaker's language, but correspond more nearly to the allophones in her repertoire. The conclusion was based on the observation that there were more categories observed in the shadowing and mimicry tasks than in the writing task in which letter symbols were used as responses to represent the phonemes present in the subject's native language. Thus, the mimicry paradigm as used by Chistovich et al. appeared to provide much more detailed information about the phonetic knowledge that the speaker has access to as well as the means used to implement these distinctions in the articulatory domain. Similar findings have been reported more recently by Schouten [1977] for imitation of synthetic vowels by Dutch-English bilinguals.

In several more recent studies, Kent [1973, 1974, 1976] has also used the mimicry paradigm to study the imitation of synthetic steady-state vowels as well as the formant motions present in time-varying vocalic stimuli. Kent [1973] found in one study that English speakers showed evidence of a continuous representation for synthetic vowels with formant trajectories ranging from /u/ to /i/, but a more nearly discrete or categorical representation for vowels with trajectories ranging from /æ/ through /i/. The /u/ to /i/ stimulus continuum contains only two vowels that are phonologically distinctive in English whereas the /æ/ through /i/ continuum contains several English vowels including: /i/, /I/, /e/, /ε/ and /æ/. While there was a tendency in this study for the standard deviations of the formant

frequencies of the response vowels to show the expected pattern in terms of the categorization hypothesis described earlier, Kent reported relatively large individual differences across his four subjects. As expected, however, there were well-defined peaks in the distributions of the responses corresponding to phonological categories along the /æ/ to /i/ continuum.

In another study which compared the imitation of English vowels with non-English vowels by adults and children, Kent [1976] found that familiarity with the vowels had a large influence on the reliability of reproduction as indexed by the standard deviations of the formant frequencies of the response vowels. Moreover, Kent also reported that the children had standard deviations roughly twice as large as those obtained with adults. While Kent attributed this finding to the possible measurement errors associated with the higher fundamental frequency of children, these results may also reveal something about the relative precision of the articulatory control children have in producing vowels under these conditions.

Although these previous investigations have used a mimicry or imitation paradigm, the interest for the most part has been limited to questions concerning the internal representation of vowels and possible familiarity effects in speech production. After the present study was completed, Broad and Wakita [1976] reported results on variability of formant frequencies for a single talker. Their findings will be discussed in more detail below. In the present study, we were concerned with a somewhat different issue concerning the relative precision with which different vowels could be produced by talkers. More specifically, we were interested in determining whether we could find some support for the idea that some vowels are inherently more stable than others and consequently we would anticipate that their formant frequencies should show lower variances. If such a pattern of results could be observed in a production paradigm, the findings could be interpreted as empirical support for the quantal theory of speech which assumes well-defined and stable regions for some vowel articulations.

The independent variable of most interest in this study was the particular vowel phoneme itself. The vowel target stimuli that we used were selected to encompass the range of the vowel space of General American English including the so-called point vowels, /i/, /a/ and /u/, as well as the nonpoint vowels, /I/, /e/, /æ/, /ʌ/ and /ɔ/. Several dependent measures were obtained from computer analyses of the vocal responses. These included the mean frequencies of the first three formants and their respective standard deviations, the duration of the vowel response, and the latency to initiate the response. Some estimates of the reliability of these measurements as well as the reliability of the overall pattern of results were obtained by means of correlations between data obtained in two separate test sessions.

Method

Stimuli

The stimuli used in this experiment consisted of the following eight synthetically generated steady-state vowels: /i/, /I/, /e/, /æ/, /ʌ/, /a/, /ɔ/, and /u/. Each of the vowels had a duration of 300 ms with a pitch contour that fell linearly from 125 to 80 Hz over the duration of the stimulus. The formant frequencies of the stimuli were originally taken from the averaged measurements provided by Peterson and Barney [1952] for their male talkers although some slight changes were made in these values to improve the vowel quality at the time of synthesis. These values are given in table I. The stimuli were originally synthesized on a digitally controlled software speech synthesizer at the Research Laboratory of Electronics, MIT [Klatt, 1972]. Since this is a serial synthesizer, formant amplitudes were determined automatically. The eight stimuli were converted to analog form, recorded on audio tape and

then subsequently transferred to digital form on a PDP-11/05 computer in the Psychology Department at Indiana University where the experiment was carried out.

Subjects

Two subjects were used in the present experiment. Both were right-handed, native speakers of English and reported no history of a speech or hearing disorder. One subject (B.S.) was a laboratory technician employed in the Psychology Department, the other subject (M.W.) was a graduate student in psychology. Subject B.S. was born in Northwestern Indiana and lived most of his life there. Subject M.W. spent most of his life in Southern California. Both subjects were naïve with respect to the hypothesis under consideration in this study. They were paid at a rate of \$ 3.50 per hour for their services.

Procedure

The subjects were run one at a time in a single-walled sound-attenuated room (Industrial Acoustics Model 401). The experiment required two sessions lasting about 1 h held on two successive days. The subject was seated comfortably in front of a high-quality microphone (Electro-Voice Model 664) and was told to keep his head positioned so that a distance of 10 inches could be maintained between his lips and the microphone. The stimuli were presented through earphones (Telephonics TDH-39) at a comfortable listening level which was the same for each subject. All of the target stimuli were stored in digital form on disk and presented in real-time during the course of the experimental sessions. The subject's response as well as the target vowel output from the computer were recorded at 7 ½ ips on separate channels of a two-channel tape deck (Ampex AG-500) for later spectral analysis. Response latency was recorded by means of a voice-operated relay interfaced appropriately to the computer. All stimulus timing, warning and response lights were controlled automatically by the computer.

There were three parts to the experiment, (1) familiarization, (2) practice and (3) test. During the familiarization phase, subjects were presented with the set of eight target vowels in order six times in a row with a 4-second interval between successive stimuli. They were told to listen carefully to each vowel and to study the following list which was present in front of them since it would help them to learn to distinguish each vowel sound:

	Vowel sound	Example of word with vowel sound
(1)	'EE'	heed
(2)	'IH'	hid
(3)	'EH'	head
(4)	'AE'	had
(5)	'UH'	hud
(6)	'AH'	hod
(7)	'AW'	hawed
(8)	'UU'	who'd

The second phase of the experiment consisted of having subjects practice producing the appropriate vowel as a response to the target vowels which were presented randomly for forty trials. Data were not collected during this phase. Subjects received the following set of detailed instructions describing the production task and how their response was to be produced:

'This is an experiment dealing with the way people imitate vowel-like sounds. On each trial you will hear through your earphones a synthetic vowel like the ones you heard earlier. They will be presented in a random order one at a time. Your task is to try to mimic the sound as best as you can. That is, say out loud the same vowel as you would produce it. Remember, the sounds you will hear are synthetic and approximations to the real vowels we want you to produce. There will be a "warning" light on the console in front of you signaling when the trial will begin. When you see the light go on you should get ready to utter the vowel that will come up next. Approximately 1 second after the light goes on the target vowel sound will come through your earphones. As soon as you know which vowel sound it is you should get ready to initiate your utterance. When the "go" light comes on you should make your response. After each vowel is presented you will have 1 second in which to initiate your mimic of the vowel. If you take longer than this red light on the panel in front of you will flash and we will proceed to the next trial. You should consider each trial independently of previous trials. The stimuli are arranged in a completely random order, so you should not try to predict what stimuli will be coming up next. There is a microphone in front of you and you should be sure that your lips are always the same distance from the microphone. It is important that you produce your vowel responses as naturally as you can. It is not necessary for you to exert any extra effort since the microphone is very sensitive and we want you to do this as naturally as possible.'

The last part of the experiment, the test phase, was identical to the practice session except that subjects received four blocks of 80 trials with a 1-min rest period between successive blocks. When subjects returned for the second test session on the next day, they received only a repetition of the third phase consisting of an additional four blocks of 80 trials. Within each block of 80 trials, each of the eight stimulus vowels appeared randomly ten times each with the only restriction that no stimulus ever followed itself in the test sequence. Thus, each block of 80 trials could be considered a completely independent replication of the basic experiment with ten trials per stimulus vowel. Stimuli within a block were separated by 3.5 s.

Spectral Analysis

Spectral data from the responses were obtained by means of linear prediction analysis [see McCandless, 1974]. The audio tapes for each subject were taken to MIT's Lincoln Laboratories where a highly interactive computer facility for studies of the acoustic properties of speech could be used for the analysis [for a description of this facility see Zue, 1976]. With the kind help and collaboration of Dr. Victor Zue we were able to obtain formant tracks of the first three formants for each response and a computer printout of the numerical values of these formant frequencies sampled every 5 ms. Armed with these printouts we obtained values of the first three formants for each response vowel and its duration as well as estimates of the fundamental frequency. The measurements contained in these printouts formed the basis for all of the subsequent data analysis reported below.

Results

Because of the large number of responses obtained in the present investigation and the amount of time required to perform the spectral analysis, only a small portion of the available data from this experiment were processed. For analysis purposes we chose the first block of 80 trials in each test session. That is, only analyses of the data obtained on trials 1 through 80 and trials 321 through 400 will be reported below.

Formant frequencies for the first three formants of each response were obtained by selecting a point 50 ms from the onset of the vowel and recording the numerical values for F_1 , F_2 and F_3 provided in the computer printout. Some care was taken at the time these values were

recorded to insure that they did not represent spurious peaks in the formant tracks or marked deviations from the immediately surrounding values of the formant frequencies. In all cases, the values selected at 50 ms seemed to be chosen from fairly stable regions located centrally in each vowel.

The mean formant frequencies for the first and second formants averaged over both experimental sessions are shown in figure 4 separately for each subject. The filled circles in this figure show the formant frequencies for the synthetic target vowels whereas the open triangles and squares show the response vowels for each subject respectively. The F_1 – F_2 vowel space for both subjects shows a reasonably good approximation to the space of the original target vowels, although there are some noticeable differences that can be observed in this figure. For example, both subjects show some deviation for the high front vowel /i/ and the vowels /I/ and /u/. In each of these cases, the frequency of F_1 is higher than in the target vowels. The mean formant frequencies of F_1 , F_2 and F_3 for each vowel broken down separately by test session are given in table II.

Product-moment correlations were carried out on the means from the two test sessions to determine whether the observed values were reliable and whether the pattern of the mean values was stable. In each case, extremely high and statistically significant correlations were obtained for both subjects on the mean formant frequencies for F_1 , F_2 and F_3 across test sessions. The correlations are shown in table III.

In order to examine the relative stability or precision with which the different vowels were produced in this experiment, we also calculated the standard deviations of the formant frequencies for each vowel. Figure 5 shows the standard deviations of the first two formants averaged over the two test sessions for each subject as a function of the target vowel. The variability of F_1 is somewhat lower than F_2 , although this would be anticipated since the range over which the frequency of F_1 can vary is much smaller than F_2 . Nevertheless, inspection of the standard deviations of F_2 shows that some vowels are more variable than others. This is particularly true for subject B.S., shown in the left-hand panel, who displays more variability for the back vowels than front vowels. Unfortunately, the same pattern of results is not observed in the data from subject M.W., shown on the right. The standard deviations of the formant frequencies of F_1 , F_2 and F_3 for each vowel are also given in table II separately for each test session.

To assess the reliability of the pattern of variances associated with the production of different vowels across test sessions, product-moment correlations were also calculated. These values are shown separately for each subject in table III. In contrast to the correlations obtained for the means, the correlations for the standard deviations were, in each case, very low and nonsignificant, suggesting that the pattern of variability across vowels is not stable from session to session. Correlations were also carried out across subjects for each formant frequency taken separately. These correlations were also quite low and nonsignificant suggesting that both subjects in this experiment were apparently employing quite different patterns of articulatory control in producing the same vowel responses. We did find, however, significant correlations, at least for B.S., between the frequency of F_1 and its standard deviation ($r = +0.753$, $p < 0.05$) and the frequency of F_2 and its standard deviation ($r = +0.752$, $p < 0.05$). Both effects may be seen in figure 5 in the left-hand panel.

We also collected latency data during the production of these vowels with the hope that some interesting pattern of results might emerge. Unfortunately, because of programming problems and limitations on the computer's cycle time, we had to instruct subjects to wait until a 'go' light was present before executing their response. There is no doubt that this gives an unrealistic estimate of the response latency and this will be corrected in future

work. Nevertheless, table IV shows the means and standard deviations for response latency in each of the two sessions and the average over both sessions. While there were some differences between different vowels, they do not appear to be reliable across test sessions nor even across subjects, which is probably to be expected based on the procedure used. However, it is clear that there is a substantial difference in the latencies for both subjects. The mean response latency across all vowels for subject B.S. was 256 ms whereas the mean latency for M.W. was 670 ms.

Although the synthetic vowels that were used as target stimuli were all synthesized with the same overall duration of 300 ms, it was of some interest to examine the durations of the vowel responses produced by the two subjects. The means and standard deviations for duration of the vowel response are shown in table V. Inspection of this table shows that while there are some slight differences in vowel length across the different vowel targets, for the most part, no systematic pattern appears to emerge, at least according to traditional criteria on vowel duration in the phonetic literature. However, the pattern of durations observed here does seem to be reliable across both test sessions, since very high correlations were observed for both subjects. It is also of some interest to note here that in all cases the two subjects produced responses of greater duration than the target stimuli, often by as much as 100 ms. It may be recalled from the instructions described earlier that we did not explicitly provide any information to the subjects concerning the natural inherent durations of these vowels and simply left this as a free variable.

Discussion

We began this report with a brief discussion of the quantal theory of speech and its basic assumption that nonlinearities can be found between articulation and acoustic output, particularly measurements of the formant frequencies of vowels. In keeping with the spirit of the quantal theory as it might be applied to the production of vowels, we predicted that some vowels, particularly the point vowels, would show greater stability or precision in articulation which, in turn, would be reflected in the variabilities of the measured formant frequencies of vowels. To a first approximation, the results obtained for two subjects in the present production task do not provide very convincing support for this hypothesis. No clear pattern of variances appeared to emerge for the eight different vowels studied. While the mean values of the formant frequencies for the responses were highly reliable over the two test sessions analyzed, no such reliability was observed for the overall pattern of variances associated with these means. Moreover, comparisons between the two subjects revealed quite different patterns of variances for the formant frequencies suggesting that these subjects appeared to carry out the production task with somewhat different strategies. Indeed, these two subjects also showed rather substantial differences in their overall latencies in initiating their vowel response.

What can be concluded from the results obtained in this study with regard to the predictions of the quantal theory? First, there is probably some reason to be suspicious of the pattern of variances obtained from the data analyzed so far, primarily because this result does not appear to be very reliable across test sessions. Secondly, it may very well be the case that the outcome of this experiment is dependent, to a large extent, on the experimental procedure used, especially requiring our subjects to wait for some brief period of time for the 'go light' before initiating their vocal response. Some modifications have already been made in the hardware in our laboratory which will permit a replication of this experiment without requiring a delay in the vocal response. The results of this study should then provide some information about whether the vowel responses should be produced spontaneously without delay. We also plan to place the vowel targets in various phonetic environments, since these have been shown to exert rather strong effects, not only on the formant frequencies of the

vowels themselves, but also on the pattern of variabilities associated with the formant frequencies [see Stevens and House, 1963].

Finally, it is quite possible that the predictions derived from the quantal theory may be simply wrong in the context of production experiments such as this or ones similar to it. It should be pointed out here that much of the evidence used to support the quantal theory of speech and its predictions with regard to the formant frequencies of vowels and their presumed stability has been obtained under highly idealized conditions in the context of simulation experiments [see, however, Stevens, 1975]. As far as we know, no empirical data have been collected with real talkers under conditions approximating those assumed by the model that Stevens has used. Obviously, for the theory to have any claim to reality some empirically motivated set of criteria would have to be specified and subsequent data collected on the issue.

Some support for the predictions of the quantal theory, at least with regard to vowels, has already been reported in the literature some years ago although in a somewhat different context. For example, in an extensive investigation on the effects of consonantal context on vowel formant frequencies, Stevens and House [1963] reported standard deviations for F_1 and F_2 for eight different vowels averaged over 14 different consonantal contexts for three talkers. Examination of these results indicated that tense vowels have substantially lower variances than lax vowels. Moreover, /i/ and /a/, two of the vowels under special consideration in the present study showed substantially smaller variances for F_2 than the other vowels. The variance of F_2 for the vowel /u/ was the highest, a fact easily accounted for by the addition of pronounced lip rounding in most consonantal contexts. After describing these data, Stevens and House [1963; p. 120] even pointed out, at the time, that 'on the basis of these data it is evident that some vowel articulations are more stable than others'. It may well be the case that the quantal theory is a reasonable account of some aspects of vowel articulation, but only when more constraints are placed on the surrounding phonetic environment of the vowel as it might appear in spontaneously produced utterances or in continuous fluent speech. Thus, the theory may be correct but not necessarily because of the current evidence obtained in the simulation work carried out by Stevens.

As noted earlier in this report, Broad and Wakita [1976] have also examined the formant variability of vowels produced by one female talker. These vowels were originally produced as instructional materials for use in phonetics courses and consequently provided a relatively large data base consisting of some 800 vowel tokens. Although the data were collected under somewhat different experimental conditions, these results were consistent with the basic findings obtained in the present study. Broad and Wakita also failed to find any systematic relation between formant variability and the articulatory features of the vowels produced by their talker. Moreover, while certain vowel configurations were found to be highly reproducible, they were also not able to find any relation between formant variability and nativeness of the vowel tokens, the formant frequencies or the formant bandwidths of the vowels produced by this talker.

In summary, the results obtained in the present vowel production experiment provide little support for the original prediction that point vowels are inherently more stable than nonpoint vowels, at least with regard to measuring the variability of the formant frequencies. The pattern of results observed in this study, as well as the procedures used in the production paradigm, suggested several modifications for future work which employs the pattern of variances of formant frequencies as an index of the precision or stability of vowel production. Other related work is currently underway in our laboratory on these problems and additional reports should be forthcoming.

Acknowledgments

This paper is based on research carried out while the author was on leave at the Research Laboratory of Electronics, Massachusetts Institute of Technology. I am grateful to Dr. Victor Zue for his kind help and cooperation with regard to the spectral analysis of the data and to Prof. Kenneth N. Stevens for his interest and advice in this work. The research was supported, in part, by NIH grants NS-07040, and NS-04332 to MIT. Additional support was provided by NIMH grant MH-24027 and NIH grant NS-12179 to Indiana University in Bloomington.

References

- Broad DJ. Toward defining acoustic-phonetic equivalence for vowels. *Phonetica*. 1976; 33:401–424.
- Broad, DJ.; Wakita, H. Some statistics on vowel formant variability. San Diego, Calif: American Association of Phonetic Sciences; 1976.
- Chistovich L, Fant G, de Serpao-Leitao A, Tjernlund P. Mimicking of synthetic vowels. *Quart. Prog. Status Rep., Speech Transmission Lab., Roy. Inst. Technol., Stockholm*. 1966; No. 2:1–18.
- Fant, G. Acoustic theory of speech production. The Hague: Mouton; 1960.
- Gerstman LJ. Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust.* AU-16. 1968:78–80.
- Kent RD. The imitation of synthetic vowels and some implications for speech memory. *Phonetica*. 1973; 28:1–25. [PubMed: 4751546]
- Kent RD. Auditory-motor formant tracking: a study of speech imitation. *J. Speech Hear. Res.* 1974; 17:203–222. [PubMed: 4836041]
- Kent, RD. Imitation of synthesized vowels by children and adults. 92nd Meeting of the Acoustical Society of America; San Diego, Calif. 1976.
- Klatt, DH. Acoustic theory of terminal analog speech synthesis. *Proc. 1972 Int. Conf. Speech Communication and Processing*; Boston, Mass. 1972. p. 131-135. *IEEE No. 72, CHO, 567–7 AE*
- Lieberman P. Towards a unified phonetic theory. *Linguistic Inquiry*. 1970; 1:307–322.
- Lieberman P. Phonetic features and physiology: a reappraisal. *J. Phonet.* 1976; 4:91–112.
- Liljencrants J, Lindblom B. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language*. 1972; 48:839–862.
- Lindblom, BEF. Phonetics and the description of language. In: Rigault, Charbonneau, editor. *Proc. 7th Int. Congr. Phonetic Sciences*; Mouton; The Hague. 1972. p. 63-97.
- Lindblom B, Sundberg J. A quantitative model of vowel production on the distinctive features of Swedish vowels. *Quart. Prog. Status Rep., Speech Transmission Lab., No. 1*. 1969:14–30.
- McCandless SS. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Trans. Acoustics Speech Signal Processing*. 1974; 22:135–141.
- Peterson GE, Barney HL. Control methods used in a study of the vowels. *J. acoust. Soc. Am.* 1952; 24:175–184.
- Schouten MEH. Imitation of synthetic vowels by bilinguals. *J. Phonet.* 1977; 5:273–283.
- Stevens, KN. The quantal nature of speech: evidence from articulatory acoustic data. In: David; Denes, editors. *Human communication: a unified view*. New York: McGraw-Hill; 1972.
- Stevens, KN. *Quart. Prog. Rep. No. 108*. Massachusetts: Research Laboratory of Electronics, MIT; 1973. Further theoretical and experimental bases for quantal places of articulation for consonants; p. 247-254.
- Stevens, KN. Quantal configurations for vowels. 89th Meeting of the Acoustical Society of America; Austin, Tex.. 1975.
- Stevens KN, House AS. Perturbation of vowel articulations by consonantal context: an acoustical study. *J. Speech Hear. Res.* 1963; 6:111–128. [PubMed: 13984031]
- Zue, VW. doctoral thesis. Massachusetts Institute of Technology; 1976. Acoustic characteristics of stop consonants: a controlled study. unpublished

Fig. 1.
Hypothetical quantal relations between a parameter that describes some aspect of articulation and the resulting acoustic parameter of speech [adapted from Stevens, 1972].

Fig. 2.

The upper panel shows a two-tube resonator approximating the vocal tract configuration for the vowel [a]. A_1 and A_2 represent the cross-sectional areas of the pharyngeal and oral cavities, respectively. The lower panel shows an approximation of the spectrum envelope for the vowel [a] produced by the idealized configuration above. The peaks in the function represent the center frequencies of the formants [adapted from Stevens, 1972].

Fig. 3. Results of a vocal tract simulation showing the relation between the frequencies of the first and second formants (F_1 and F_2) when the length of the back tube, d_1 , is varied between 5 and 15 cm [adapted from Stevens, 1972].

Fig. 4.

Mean values of the first- and second-formant frequencies (in Hz) for two subjects, B.S. and M.W., obtained in the production task. The filled circles show the formant frequencies for the synthetic target vowels, whereas the solid lines in each panel are drawn through the vowel space of the observed responses.

Fig. 5.
Standard deviations (in Hz) of the mean formant frequencies for F₁ (●) and F₂ (Δ) for the eight target vowels shown separately for each subject.

\$watermark-text

\$watermark-text

\$watermark-text

Table I

Formant frequencies of synthetic target vowel (in Hz)

	i	I	e	æ	Λ	a	ɔ	u
F ₁	270	374	530	660	640	730	570	300
F ₂	2,290	2,070	1,840	1,720	1,190	1,090	840	870
F ₃	3,010	2,666	2,480	2,410	2,390	2,440	2,410	2,240

Table II
Means and standard deviations (in Hz) for F₁, F₂ and F₃ by sessions for two subjects

Subject B.S.										Subject M.W.							
Vowel target	i	I	e	æ	Λ	a	ɔ	u	i	I	e	æ	Λ	a	ɔ	u	
Means																	
X̄F ₁	(1)	304	476	560	658	600	775	590	322	320	434	501	610	560	742	576	336
	(2)	317	468	552	650	600	778	608	352	342	421	484	580	593	769	558	348
Mean		311	472	556	654	600	777	600	337	331	428	493	595	577	756	567	342
X̄F ₂	(1)	2,259	1,923	1,861	1,775	1,201	1,141	933	889	2,039	1,907	1,791	1,784	1,237	1,040	871	979
	(2)	2,246	1,996	1,839	1,713	1,250	1,179	999	998	2,140	1,920	1,824	1,664	1,235	1,086	897	855
Mean		2,253	1,960	1,850	1,744	1,226	1,160	966	945	2,090	1,914	1,808	1,724	1,236	1,063	884	917
X̄F ₃	(1)	2,954	2,561	2,550	2,466	2,427	2,502	2,431	2,349	3,122	2,594	2,486	2,427	2,153	2,345	2,234	1,998
	(2)	2,900	2,537	2,570	2,474	2,451	2,404	2,332	2,292	2,715	2,512	2,438	2,395	2,329	2,353	2,294	2,060
Mean		2,927	2,549	2,560	2,470	2,439	2,453	2,382	2,321	2,919	2,553	2,462	2,411	2,241	2,349	2,264	2,029
Standard deviations																	
σF ₁	(1)	23	31	17	32	18	49	9	16	15	21	35	53	77	26	40	20
	(2)	12	8	22	38	19	44	35	2	16	30	18	28	19	31	19	9
Mean		18	20	20	35	18	47	22	9	16	26	27	40	48	29	25	15
σF ₂	(1)	86	63	45	74	33	43	57	35	48	63	51	91	61	27	32	57
	(2)	67	144	107	66	33	39	53	75	108	35	45	71	33	80	30	91
Mean		77	104	76	70	33	41	55	56	78	49	48	81	47	54	31	74
σF ₃	(1)	146	76	56	113	60	83	79	127	51	48	43	55	129	28	61	50
	(2)	91	53	148	23	193	30	43	97	183	39	30	44	54	53	32	88
Mean		119	65	102	68	127	57	61	112	117	44	37	50	92	41	47	69

Table IIIProduct-moment correlations for means and standard deviations for F₁, F₂ and F₃ across test sessions

		Subject B.S.	Subject M.W.
Means	F ₁	+ 0.997 *	+ 0.986 *
	F ₂	+ 0.996 *	+ 0.883 *
	F ₃	+ 0.965 *	+ 0.891 *
Standard deviations	F ₁	+ 0.427	+ 0.080
	F ₂	+ 0.229	– 0.008
	F ₃	– 0.344	– 0.063

* With $df = 6$ these correlations are significant beyond the 0.005 level.

\$watermark-text

\$watermark-text

\$watermark-text

Table IV

Response Latency (in ms) by test session for two subjects

	Vowel target									Means
	i	I	e	æ	ʌ	a	ɔ	u		
Subject B.S.										
Session 1										
mean	239	259	259	243	246	252	234	213	243	
σ	102	63	50	64	75	57	53	45	63	
Session 2										
mean	234	305	243	226	244	232	219	222	240	
σ	93	71	32	31	26	28	44	43	46	
All sessions*										
mean	228	296	255	250	270	274	243	238	256	
σ	60	68	45	48	71	69	45	54	58	
Subject M.W.										
Session 1										
mean	676	652	635	584	608	552	568	598	609	
σ	74	74	96	83	44	70	58	59	70	
Session 2										
mean	711	727	658	656	709	677	651	714	687	
σ	75	61	78	76	54	57	97	91	74	
All sessions*										
mean	724	698	657	657	672	645	621	692	671	
σ	85	75	81	83	78	79	109	94	86	

* Each mean is based on 80 responses per vowel obtained over all sessions including trials 1–640.

\$watermark-text

\$watermark-text

\$watermark-text

Table V

Vowel response duration (in ms) by test session for two subjects

	Vowel target									
	i	I	e	æ	ʌ	a	ɔ	u		
<i>Subject B.S.</i>										
Session 1										
mean	415	416	404	380	418	415	446	475		
σ	37	24	30	35	25	58	27	29		
Session 2										
mean	401	401	397	385	400	412	410	437		
σ	36	35	35	20	17	21	23	32		
Overall										
mean	408	409	401	383	409	414	428	456		
σ	37	30	33	28	21	40	25	31		
<i>Subject M.W.</i>										
Session 1										
mean	387	403	406	422	427	433	426	387		
σ	23	21	37	24	27	25	35	24		
Session 2										
mean	378	408	406	412	411	431	420	370		
σ	17	19	31	18	21	25	24	41		
Overall										
mean	383	406	406	417	419	432	423	379		
σ	20	20	34	21	24	25	30	33		