NAMRL 1300

INTERACTIVE VOICE TECHNOLOGY: VARIATIONS IN THE VOCAL UTTERANCES OF SPEAKERS PERFORMING A STRESS-INDUCING TASK

James D. Mosko, Kenneth N. Stevens, and Glenn R. Griffin

August 1983

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
PENSACOLA FLORIDA

Approved for public release; distribution unlimited

INTERACTIVE VOICE TECHNOLOGY: VARIATIONS IN THE VOCAL UTTERANCES OF SPEAKERS PERFORMING A STRESS-INDUCING TASK

James D. Mosko
Naval Aerospace Medical Research Laboratory

Kenneth N. Stevens
7 Hancock Place
Cambridge, Massachusetts 02139

and

Glenn R. Griffin
Naval Aerospace Medical Research Laboratory

62758N MF58528 MF5852801B 0001

Reviewed by
Ashton Graybiel, M.D.
Chief Scientific Advisor

Approved and Released by
Captain W. M. Houk, MC, USN
Commanding Officer

16 August 1983

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
NAVAL AIR STATION
PENSACOLA, FLORIDA 32508

SUMMARY PAGE

THE PROBLEM

If speech recognition systems are to be successfully implemented in advanced military aircraft and other military weapons platforms, they must be able to respond correctly to voice commands produced under wide-ranging operational conditions. Information is needed concerning the modifications of speech behavior to be expected under various operational conditions so that knowledge about such vocal variability can be incorporated into the design of appropriate recognition systems.

FINDINGS

Detailed acoustical analyses were conducted of the words produced by four speakers undergoing a motion disorientation-inducing performance task. The results indicated that the speakers differed markedly in the types and magnitudes of the changes that occurred in their speech. For at least some of the speakers, the stress-inducing experimental condition caused an increase in fundamental frequency, changes in the pattern of vocal fold vibration, shifts in vowel production, and changes in the relative amplitudes of sounds containing turbulence noise. All of the speakers showed greater variability in their production of the words in the experimental condition than in a more relaxed control situation. This variability was manifested in the acoustical characteristics of individual phonetic elements, particularly in speech sounds occurring in the vicinity of unstressed syllables. The kinds of changes and variability observed serve to emphasize the limitations of speech recognition systems based on template matching. There is need for a better understanding of these phonetic modifications and for developing ways of incorporating knowledge about these changes into speech recognition systems.

INTRODUCTION

Of the many possible military applications of speech recognition devices, the high-performance aircraft weapons platform appears to be one environment for beneficial integration. In this often "hands-busy", "eyes-busy" environment, alternate methods of systems query and control could enhance aircrew performance (1). This platform also offers many challenges to implementing speech recognition technology because the physical, cognitive, and psychological demands on the pilot are changing constantly. Often the changes are extreme, transitory, and unpredictable. Under such conditions the speech behavior of pilots will exhibit wide variations which can degrade voice communication performance with speech recognition devices. As plans develop to incorporate interactive voice systems in advanced military aircraft, it is important to determine how the speech of a pilot is likely to be modified in stressful operational situations.

Two of the authors have reported elsewhere (2, 3) the results of two experiments wherein attempts were made to document some of the vocal changes occurring in stressful conditions encountered in aviation -- high-noise levels (2) and motion disorientation (3). The conditions were simulated in the laboratory and variations in voice fundamental frequency, word token duration, and voice amplitude were reported. Since such variations rarely occur in isolation, it is important to understand other acoustic manifestations of such changes. In this report, we present a more detailed acoustical analysis of the vocal utterances of four selected speakers from one of the earlier experiments (3). The results of this analysis will serve as a guide for future work in defining the variations in speech produced by aviators in stress-inducing environments.

METHOD

The vocal utterances of four young adult male volunteer subjects, recorded as the subjects performed on the Visual-Vestibular Interaction Test (VVIT) (4), were subjected to detailed acoustical analyses. None of the subjects exhibited any hearing, speech, or physical condition to preclude their participation.

Description of VVIT.

In the VVIT individual subjects are seated in a blacked-out rotatable device facing a front lighted 17x17cm matrix. The matrix contains coordinate letters and digits on the left and upper margins, respectively, and randomly arranged digits in the body of the matrix. The subject receives an aurally presented coordinate set cue and his task is to locate the intersection of the coordinates in the body of the matrix, verbally report the digit at the intersection and the next two digits immediately below it in the same column. A typical cue is "A-2", and a typical verbal response is "six...five...nine." Each subject performs the task twice: once in a stationary, STATIC, mode and once in an oscillating, DYNAMIC, mode (sinusoidal oscillation, 0.02 Hz, 30 rpm peak).

The motion stimulus in the DYNAMIC mode involves two aspects of motion stress: 1) vestibular degradation of visual performance and 2) motion sickness. The first is generated by the vestibulo-ocular reflex (VOR) which tends to drive the eyes relative to the head-fixed display, thereby degrading visual target acquisition. With the sinusoidal stimulus, this aspect is definitely cyclic; the VOR and visual performance degradation reach a maximum twice in each cycle (corresponding roughly to the peak angular velocities), and diminish approximately to zero twice each cycle, as zero angular velocity is approached. In some individuals, exposure to this form of visual-vestibular interaction induces motion sickness which can build to the point of emesis within a 5-minute exposure (4).

In the present experiment, subjects were permitted to terminate the DYNAMIC mode if they felt severe motion sickness would result. Two subjects (speakers 2 and 4) completed both modes of the VVIT and two speakers (speakers 1 and 3) terminated the DYNAMIC mode prior to a complete trial.

Target Words.

The subjects were presented a target word preceding the coordinate cue. The subjects were instructed to repeat the target word and then to report the digits corresponding to the coordinate cue. A typical cueing sequence was "TACAN...A...2" and a typical subject response was "TACAN...seven...nine... two". The ten target words are listed in Table I. Two randomized lists of the target words and correct digit sequences are listed in Table II. Forty-three items per condition were used to conform to the time constraints of the VVIT.

Table I

List of target words used in the experiment

altitude	marker
bogey	monitor
contact	pattern
heading	radar
holding	tacan

ACOUSTIC ANALYSIS

The number of utterances available for acoustic analysis was approximately 400 under the CONTROL condition, 160 under the STATIC condition, and 120 under the DYNAMIC condition. Wideband spectrograms, 0 - 5000 Hz frequency range, were made of all of the words produced under the STATIC and DYNAMIC conditions, and for a selected subset of the words in the CONTROL condition. Various measurements and observations were made from all of these spectrograms,

Table II

Sequence of Words Presented to Speakers in STATIC and DYNAMIC Conditions

	STATIC CONDITION				DYNAMIC CONDITION		
	Cue	Carrier	Digits		Cue	Carrier	Digits
1.	A-Z	Pattern	3 4 2	1.	G-7	Radar	4 7 6
2.	H-10	Bogey	9 1 4	2.	D-5	Altitude	1 7 6
3.	L-7	Heading	7 6 5	3.	B-4	Holding	8 2 4
4.	K-9	Monitor	9 3 9	4.	L-1	Pattern	4 9 6
5.	C-9	Marker	1 2 9	5.	J-3	Bogey	4 3 4
6.	H-12	Altitude	2 7 3	6.	G-2	Heading	7 6 3
7.	J-3	Heading	4 3 4	7.	A-5	Radar	4 2 7
8.	E-10	Pattern	9 6 5	8.	K-12	Monitor	7 9 3
9.	B-4	Holding	8 2 4	9.	G-4	Contact	1 4 3
10.	A-10	Radar	8 1 3	10.	I-8	Tacan	4 8 7
11.	G-1	Bogey	3 1 5	11.	A-3	Holding	1 8 1
12.	L-2	Altitude	6 5 4	12.	I-4	Bogey	9 4 8
13.	J-10	Heading	3 4 2	13.	B-3	Altitude	4 4 3
14.	E-4	Radar	4 1 5	14.	L-2	Contact	6 5 4
15.	F-3	Marker	6 7 7	15.	C-1	Radar	8 3 1
16.	I-2	Bogey	8 3 6	16.	K-3	Tacan	7 6 6
17.	G-9	Heading	5 6 6	17.	L-5	Marker	5 4 9
18.	C-1	Tacan	8 3 1	18.	I-7	Bogey	2 8 3
19.	C-10	Monitor	2 5 1	19.	L-3	Altitude	5 3 7
20.	F-8	Marker	1 6 7	20.	E-12	Heading	6 8 5
21.	J-4	Holding	6 7 4	21.	K-1	Monitor	4 7 9
22.	L-1	Tacan	4 9 6	22.	I-12	Altitude	6 5 2
23.	K-12	Contact	7 9 3	23.	D-10	Monitor	9 8 7
24.	E-8	Radar	1 5 9	24.	D-2	Bogey	7 1 7
25.	G-5	Holding	6 3 1	25.	E-7	Marker	6 5 7
26.	I-9	Radar	5 2 1	26.	G-10	Heading	5 4 7
27.	B-1	Contact	3 6 9	27.	I-5	Contact	3 6 6
28.	D-3	Tacan	2 9 8	28.	B-12	Tacan	6 9 7
29.	H-4	Holding	1 4 2	29.	H-2	Radar	4 7 1
30.	D-9	Pattern	1 9 2	30.	F-9	Monitor	5 2 3
31.	K-1	Tacan	4 7 9	31.	H-10	Holding	9 1 4
32.	A-7	Monitor	1 3 4	32.	C-4	Contact	4 3 8
33.	G-10	Bogey	5 4 7	33.	L-4	Tacan	1 8 5
34.	E-7	Contact	6 5 7	34.	A-9	Marker	6 4 1
35.	H-12	Monitor	2 7 3	35.	L-12	Pattern	9 6 8
36.	I-4	Altitude	9 4 8	36.	J-10	Heading	3 4 2
37.	D-2	Marker	7 1 7	37.	E-4	Pattern	4 1 5
38.	A-2	Altitude	3 4 2	38.	H-7	Holding	1 4 7
39.	J-3	Contact	4 3 4	39.	I-9	Marker	5 2 1
40.	B-10	Pattern	4 3 1	40.	F-1	Pattern	9 8 5
41.	J-7	Bogey	4 2 8	41.	A-7	Radar	1 3 4
42.	D-8	Radar	6 2 9	42.	G-5	Bogey	6 3 1
43.	F-3	Monitor	6 7 7	43.	K-7	Monitor	6 8 6

as described in detail later. For selected utterances, additional analysis and observation was carried out using a computer. In order to do this additional analysis, speech waveforms were displayed in some cases, discrete Fourier transforms at selected points in the utterances were calculated and displayed, measurements of fundamental frequency were made, formant frequencies were determined using a linear predictive coding (LPC) procedure (5), and measurements of durations of certain speech events were made. All the computer-based data were obtained with the waveform low-pass filtered (4.8 kHz), sampled at 10 kHz, and first-differenced. Measurements of every property or parameter for each utterance were not attempted. A more limited goal was to make a sufficient number of measurements and observations to allow a cataloguing of the kinds of changes that occurred in the utterances under the various experimental conditions.

In order to guide the data analysis, we recognized that two broad categories of change can occur in an utterance when it is repeated by a speaker under different conditions. One of these is associated with overall changes in the "posture" or "state" of the speaker's speech production system, and these changes are reflected in certain average characteristics of the utterances. Thus, for example, a speaker may speak with greater effort by using a higher subglottal pressure. This higher pressure could lead to a greater overall amplitude of the speech sounds, a modification of the average spectrum of the glottal output, and a possible modification of sounds produced with abrupt release or with turbulence noise. Or, a greater overall tension of the vocal folds or a change in the characteristics of the vocal-fold surfaces could lead to a modification of the frequency, spectrum, and regularity of the glottal output.

A second type of change that can occur when a speaker repeats an utterance under different conditions is a modification of particular phonetic elements within the utterance. In English (as in any language) a number of rules can be applied optionally to specify the way in which particular phonetic elements may be produced in specified phonetic environments. These rules generally apply to so-called redundant features (i.e., phonetic features which, by themselves, are not utilized for signaling a phonetic distinction in the language, but which can provide cues a listener might use in decoding an utterance), in addition to the cues associated with features that are distinctive. These types of optional rules are illustrated in the following examples:

1) A word-initial /b/, /d/, or /g/ in English can be optionally prevoiced (e.g., the initial /b/ in the word bogey);

2) An unstressed vowel between two voiceless consonants can be optionally produced as a voiceless vowel (e.g., the unstressed vowel in altitude);

3) An utterance-final stop consonant may or may not be released (e.g., the final /d/ in altitude);

4) A /t/ preceding an unstressed vowel may be optionally produced as a flap (e.g., in the word monitor).

As a preliminary to a detailed discussion of the data, and to indicate the acoustic structure of the words used in the experiment, we show in Figure 1 a spectrogram of each word produced by one of the speakers. Eight of the words are bisyllabic, all with stress on the first syllable, and two of the words have three syllables. The words contain a variety of stop and nasal consonants in initial, intervocalic, and final position, and both back and front vowels are represented in the corpus.

RESULTS AND DISCUSSION

AVERAGE PROPERTIES OF UTTERANCES

Fundamental frequency

In the utterances produced under the CONTROL condition, the fundamental frequency (F_0) was usually considerably lower than in the two experimental conditions. This difference presumably arose for several reasons: 1) the CONTROL utterances were generated in a phrase following the carrier word say; 2) these utterances were generated in rapid succession in the form of a list; and 3) in the experimental condition the words were followed by a number sequence. In view of these differences, we shall not compare F_0 patterns for CONTROL and experimental conditions, but only within the different experimental conditions.

The values of F_0 to be reported here were obtained by measuring the glottal period for three successive periods at a selected point in the word (usually 50 msec following the onset of the vowel selected for study) and averaging the reciprocals of these three numbers. There was some variation in the average F_0 from word to word throughout an experimental run, presumably because of variations in intrinsic F_0 from vowel to vowel, and because of influences of voicing characteristics of adjacent consonants. Consequently, in examining the effect of experimental conditions on F_0, we will either observe the F_0 for particular words throughout an experimental run, or we shall average over groups of words.

There were considerable differences among speakers in the way F_0 varied throughout the STATIC and DYNAMIC experimental runs. Selected data for the four speakers are presented in Figure 2. For Speaker 2, there was essentially no change in F_0 as the experiment proceeded through the STATIC and DYNAMIC conditions. Data for this speaker are shown for both stressed and unstressed vowels. For the other three speakers, there was some rise in F_0 throughout the experiment, particularly in the DYNAMIC condition. In the case of Speaker 1, for whom F_0 measurements were made in both stressed and unstressed vowels, the increase in F_0 for the DYNAMIC condition seemed to be more marked for the stressed vowels. The speaker showing the greatest F_0 rise appeared to be Speaker 3. Figure 3 displays the F_0 values for the stressed vowels of this speaker, averaged over successive groups of five words, throughout the experiment. There appears to be a gradual rise in F_0 as the

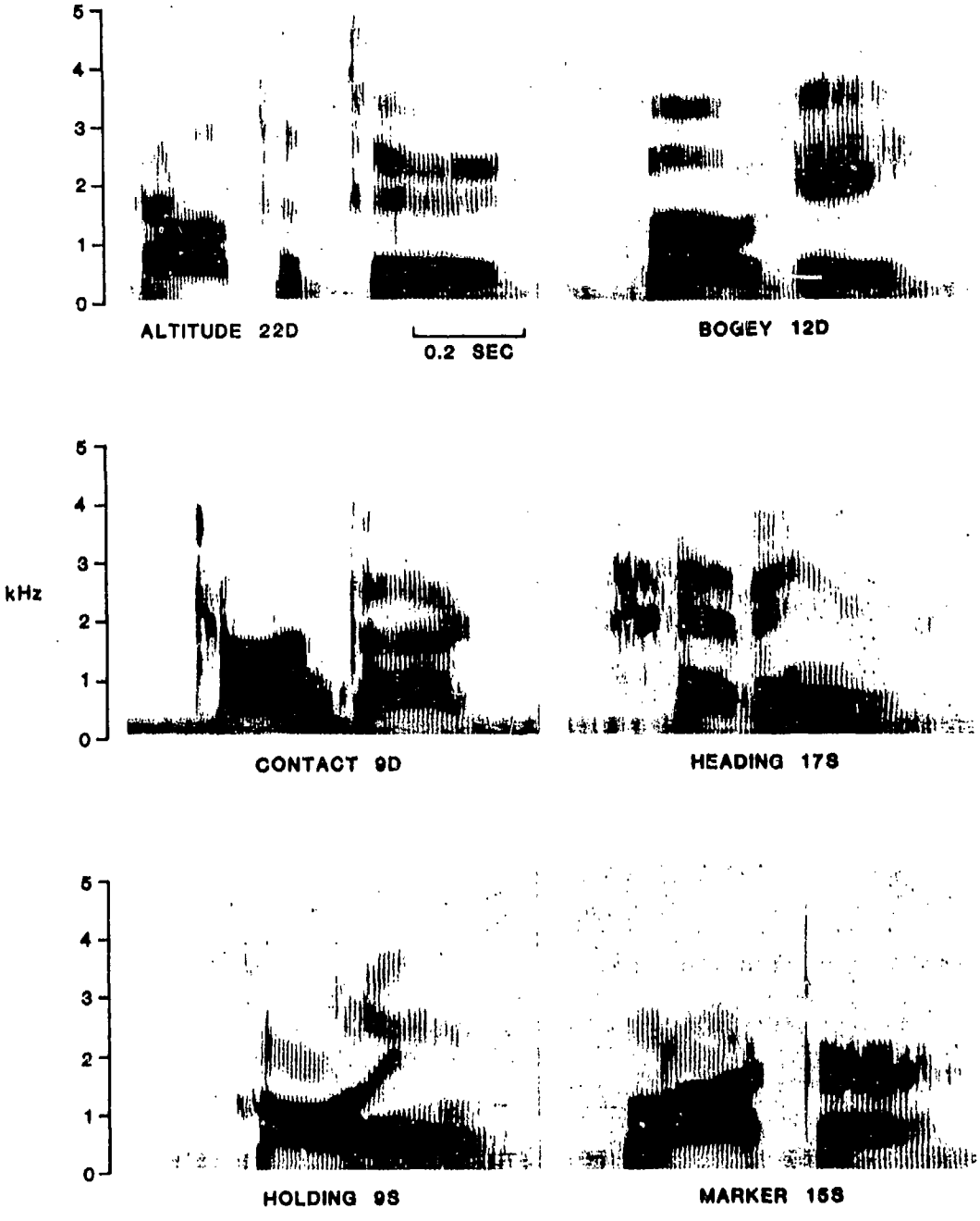

Figure 1 (Part a.). See caption Figure 1 (Part b).

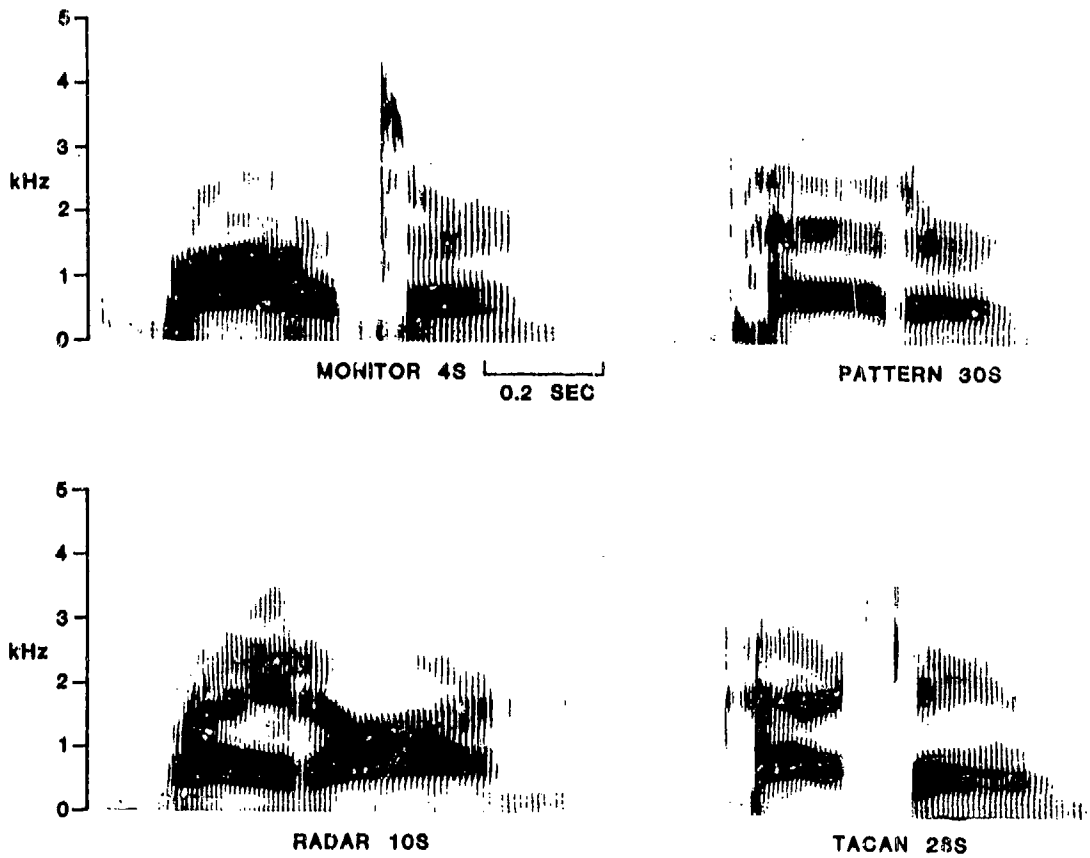

Figure 1 (Part b). Spectrograms of one example of each of the ten test words produced by Speaker 4. The number identifying each word is listed in Table 2, where S = STATIC and D = DYNAMIC.

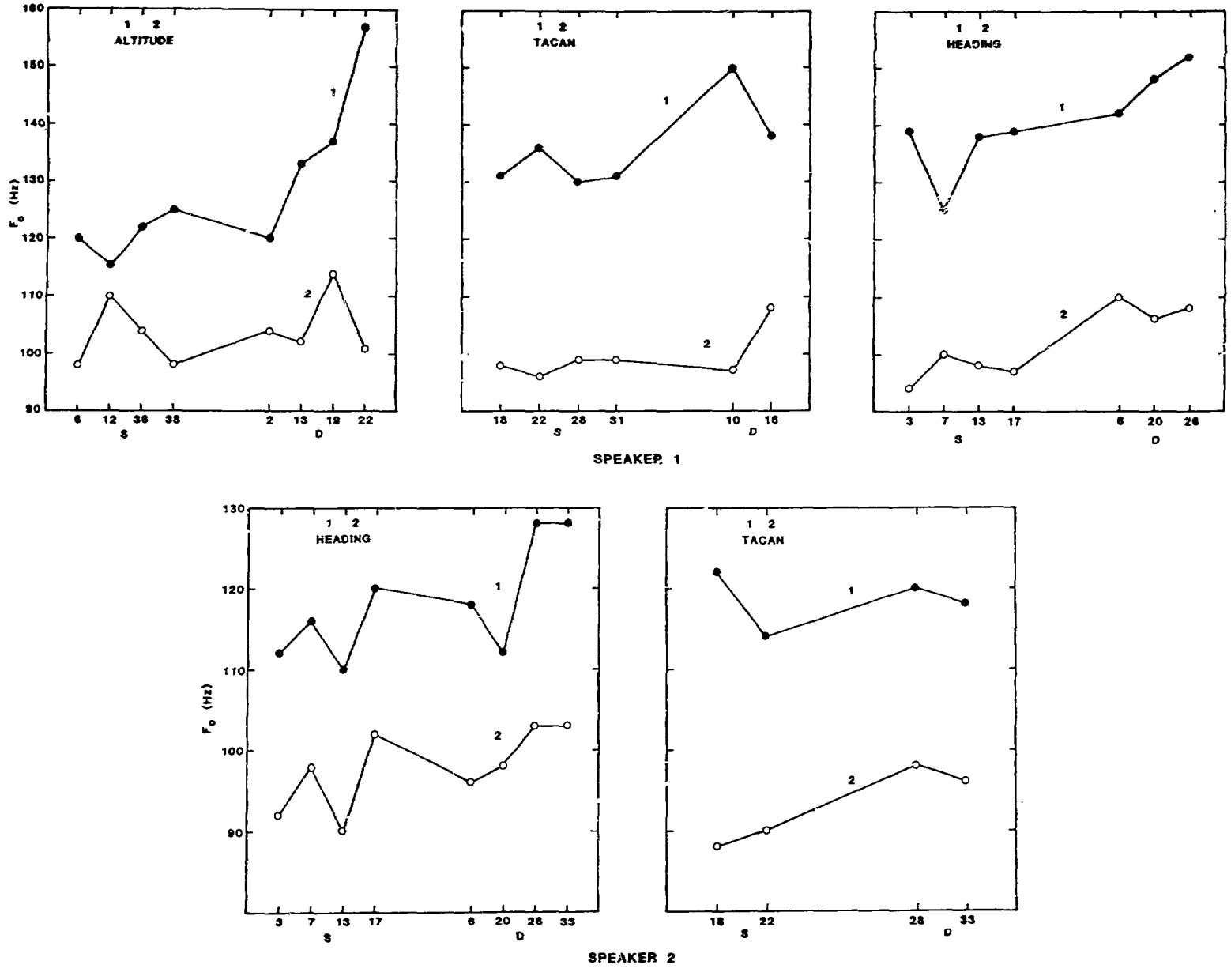

Figure 2 (Part a). See caption Figure 2 (Part b).

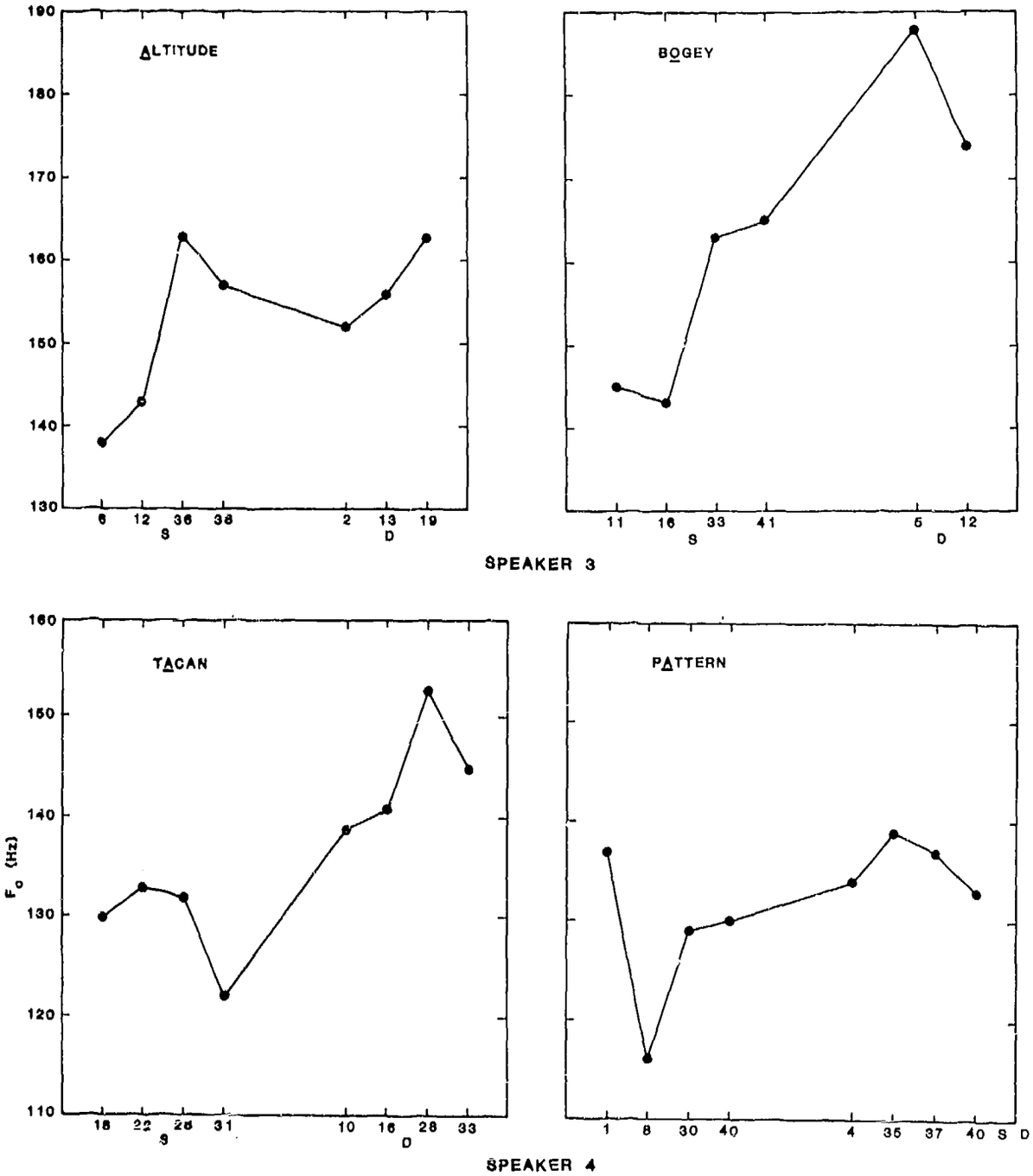

Figure 2 (Part b). Examples of fundamental frequency changes in selected words throughout an experimental session for each of the four speakers. Each point represents the mean F_o for a single token over three successive glottal periods, sampled at a point 50 msec from the onset of the vowel. The numbers on the abscissa represent token numbers (from Table 2) for the STATIC (S) and DYNAMIC (D) conditions. In the case of Speakers 1 and 2, data are shown for both the stressed vowel (solid dots) and the final vowel (open circles). These data show that the stress-inducing conditions of the experiment have different effects on different speakers, and that the effect on the F_o of the stressed vowels tends to be greater than the unstressed vowels.

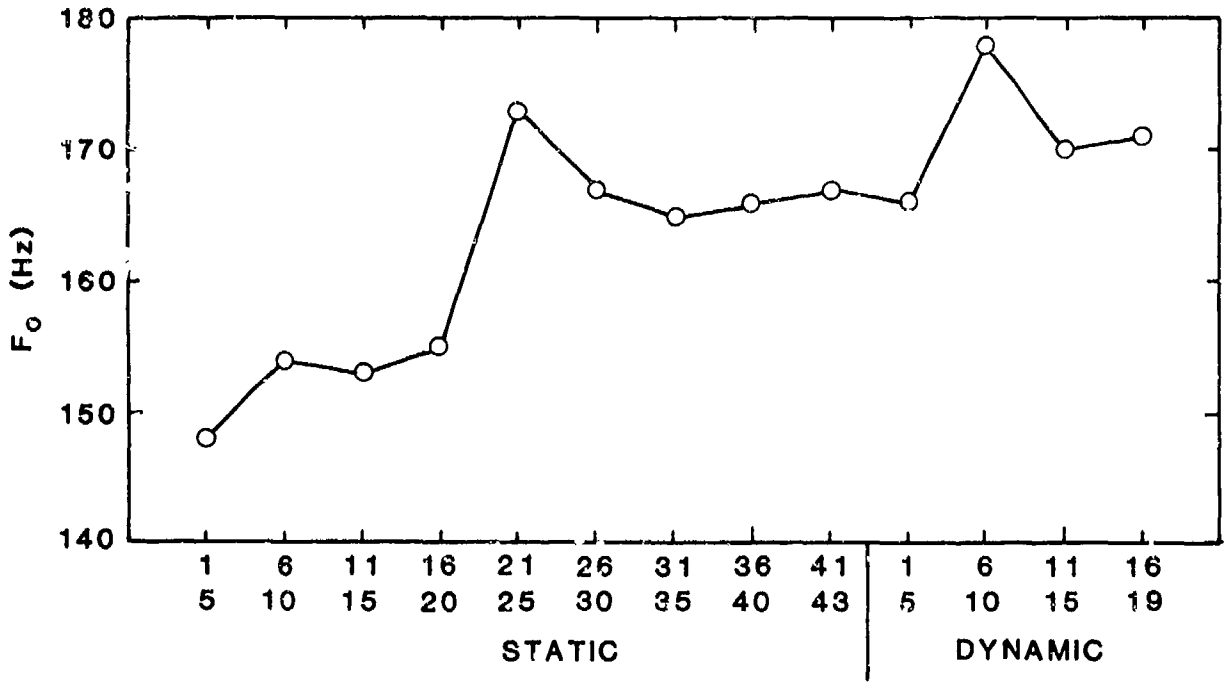

Figure 3. Fundamental frequency for stressed vowels in words produced by Speaker 3 in the course of an experimental session. F_0 values for each vowel were obtained by procedures indicated in Figure 2 and in the text. Each point usually represents the average of these F_0 values over the vowels in five successive words in the experiment, as indicated on the abscissa.

speaker progresses through the STATIC condition, and F_0 remains at the higher value in the DYNAMIC condition until the experiment is terminated at item 19D. The local peaks in the curve could represent brief intervals in which a transient change occurred in the speaker, possibly as a result of increased stress.

These increases in F_0 for some speakers as the experiment progressed presumably reflect either a rise in vocal-fold tension or a rise in subglottal pressure, or an overall increase in tension that occurs throughout the articulatory, laryngeal, and lower respiratory systems.

State and configuration of vocal folds

The principal acoustic consequence of a change in the tension of the vocal folds is a change in the frequency of vocal-fold vibration. However, another consequence of a modification in the state of the vocal-fold surfaces or in the configuration of the vocal folds is an alteration in the waveform of the volume velocity source at the glottis. This kind of change is more difficult to quantify, since the glottal waveform is filtered by the vocal tract, and the sound wave provides only indirect evidence for this waveform.

For one speaker (Speaker 4) informal listening to the recording for the STATIC and DYNAMIC conditions indicated that a change in voice quality occurred as the utterances for the DYNAMIC condition were produced. An attempt was made to establish some kind of acoustic manifestation of this change, through observation of the waveforms or spectra of the vowels. Figure 4 shows the waveform and the spectrum sampled 50 msec from onset of voicing for the vowel /æ/ in the word pattern produced by Speaker 4. The two utterances of the word were at the beginning of the STATIC condition, and near the end of the DYNAMIC condition. Also shown in the figure are spectrograms of these two words.

Probably the most striking difference between the two utterances is in the waveform. For the DYNAMIC condition, there is a more rapid decay in the first-formant oscillation in the early part of the glottal cycle, and the oscillation is almost extinguished in the later part of the cycle (presumably during the most open phase of the glottal cycle). In the spectra, the main difference is in the relative amplitude of the lower harmonics, particularly the first. Thus for utterance 40D, the amplitude of the first harmonic is about 12 dB below the peak in the first formant (F1), whereas in item 1S this difference is about 18 dB. In the spectrogram there is evidence for a "filling in" of spectral energy at low frequencies, below F1, in item 40D.

These acoustic differences can be ascribed to changes in the glottal configuration for the two utterances. In the case of 40D, it is probably that the vocal folds are more abducted, so there is never a complete closure of the glottis during the cycle. This configuration would lead to greater acoustic losses in the region of F1, as indicated by the more rapid decay of the waveform.

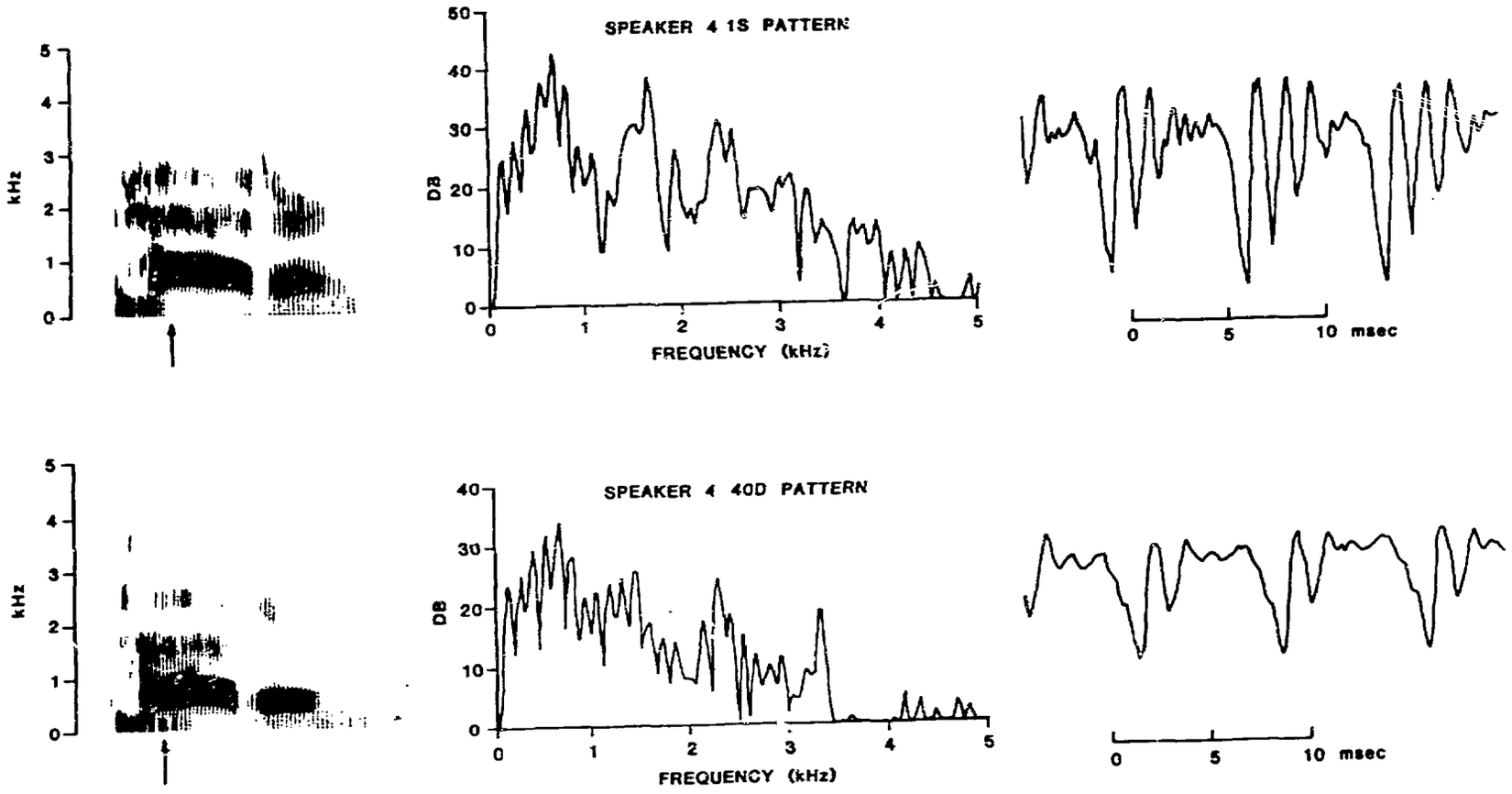

Figure 4. Illustration of the change that appears to occur in the glottal waveform in the course of an experimental session for Speaker 4. Data for two tokens of the word pattern are shown: one produced at the beginning of the STATIC condition (top), and the other near the end of the DYNAMIC condition (bottom). Spectrograms of the two words are shown at the left. The middle panels display Fourier spectra (26 msec window) sampled at the points indicated by the arrows under the spectrograms. At the right are shown segments of waveforms of the vowels sampled at the same points. These examples are interpreted as evidence for changes in the pattern of vibration of the vocal folds as the experiment proceeds.

The present study did not examine this kind of modification in detail for every speaker. For Speakers 1, 2, and 3, informal observations of waveforms and spectra of the type shown in Figure 4, together with informal auditory evaluations, failed to show consistent changes in the glottal waveform or voice quality that were as marked as those illustrated for Speaker 4 in Figure 4. The figure illustrates, however, the nature of the changes that can occur in the acoustic characteristics of a vowel when there are modifications in the manner of vibration of the vocal folds.

Formant frequencies

The frequencies of the first three or four formants were measured at selected points in a number of the utterances produced by each of the speakers in the CONTROL, STATIC, and DYNAMIC conditions. An example of the kind of data that emerged from these measurements for Speaker 4 is shown in Figure 5a. The formant frequencies in each utterance of the vowel /æ/ in pattern were measured (using an LPC algorithm) at two points located approximately 40 and 60 msec from the onset of voicing. Each point in Figure 5a represents an average of these two values for F1 and formant 2 (F2) for one utterance. This figure shows that the formant frequencies for utterances of this word in the CONTROL condition are tightly clustered, indicating that the successive repetitions of the word are very similar. The scattering of the points suggests that there is greater variability in the utterances for the STATIC condition, and still greater variability for the DYNAMIC condition. Although it is difficult to draw firm conclusions on the basis of so few utterances, there is a tendency for F2 (and, to some extent, F1) to be higher for the experimental conditions than for the CONTROL condition. A possible interpretation is that the larynx is positioned slightly lower for the more relaxed CONTROL condition, leading to a longer vocal tract and hence lower formant frequencies. Similar tendencies can be observed for the first vowel in the word tacan for Speaker 4, shown in Figure 5b, although in this case the points for the CONTROL condition are not quite as tightly clustered.

Examples of the formant frequency data for the other speakers are given in the various panels of Figure 6. The formant values were obtained in the same way as described above, i.e., each point represents the average of two measurements spaced 20 msec apart. In the stressed vowel in the word tacan, Speaker 1 shows essentially the same amount of variability in F2 for the CONTROL and the experimental conditions. For the /ɛ/ in heading, on the other hand, the points for the CONTROL condition are more tightly clustered. For both vowels, F1 seems to be lower for the CONTROL condition than for the experimental conditions. In the two vowels of altitude for Speaker 1, there is again substantial variability in F2 for the experimental conditions. For all the utterances of Speaker 1 there is no obvious trend in the data when STATIC and DYNAMIC conditions are compared.

The limited amount of data collected for Speaker 2 suggest that he was less variable in his vowel productions during the experimental conditions than were the other speakers. Speaker 3 shows more variability, particularly in the stressed vowel in altitude. Apparently the influence of the following /l/ was different from one repetition to the next.

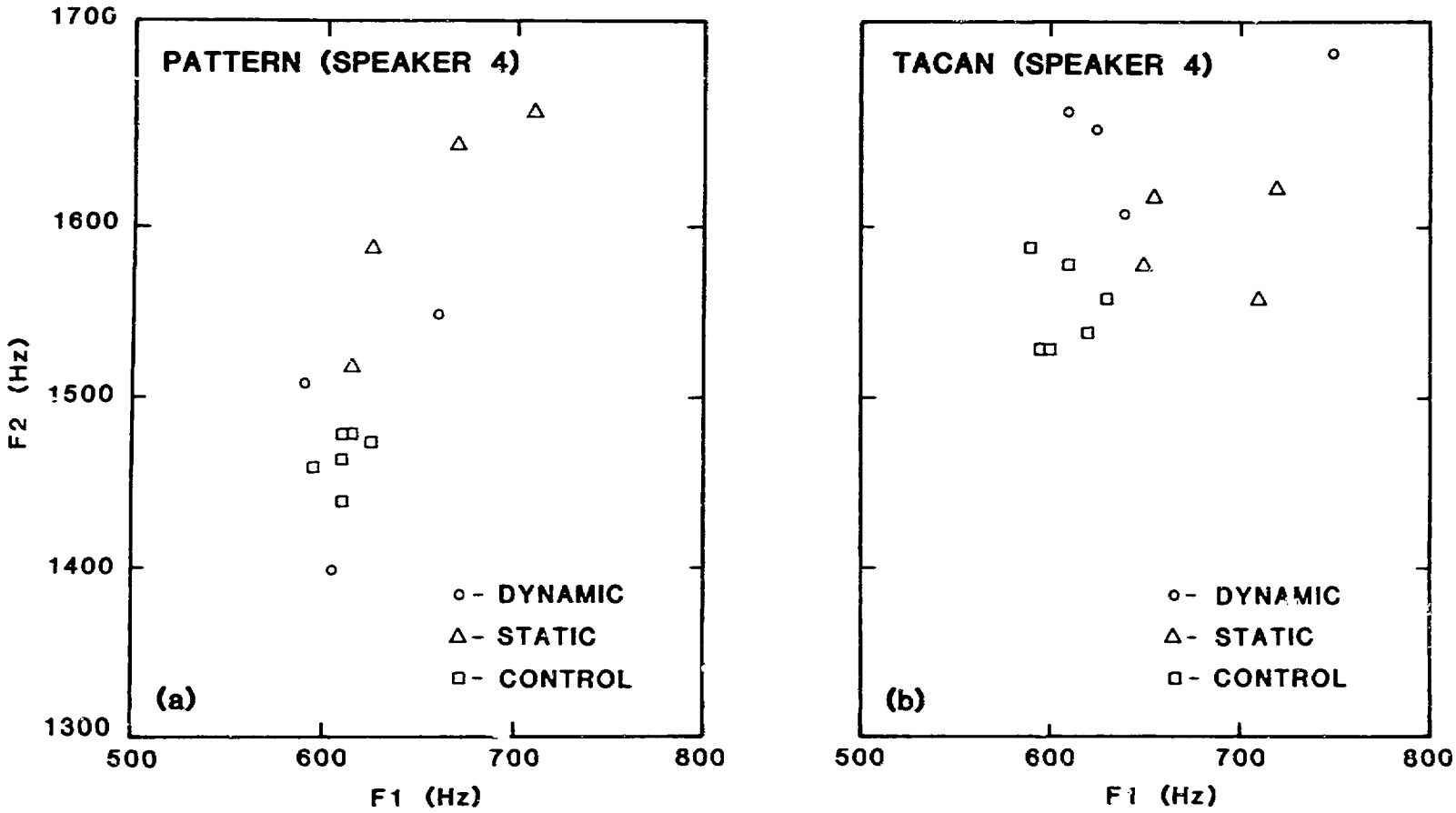

Figure 5. Data on formant frequencies F1 and F2 for the stressed vowel in two words produced by Speaker 4 in the CONTROL, STATIC, and DYNAMIC conditions. Formant frequencies represent averages of two measurements spaced about 20 msec apart, centered 50 msec from vowel onset. There appears to be less variability in vowel production for the CONTROL condition than for the experimental conditions, as well as a tendency for F1 and F2 to increase in the stress-inducing situation.

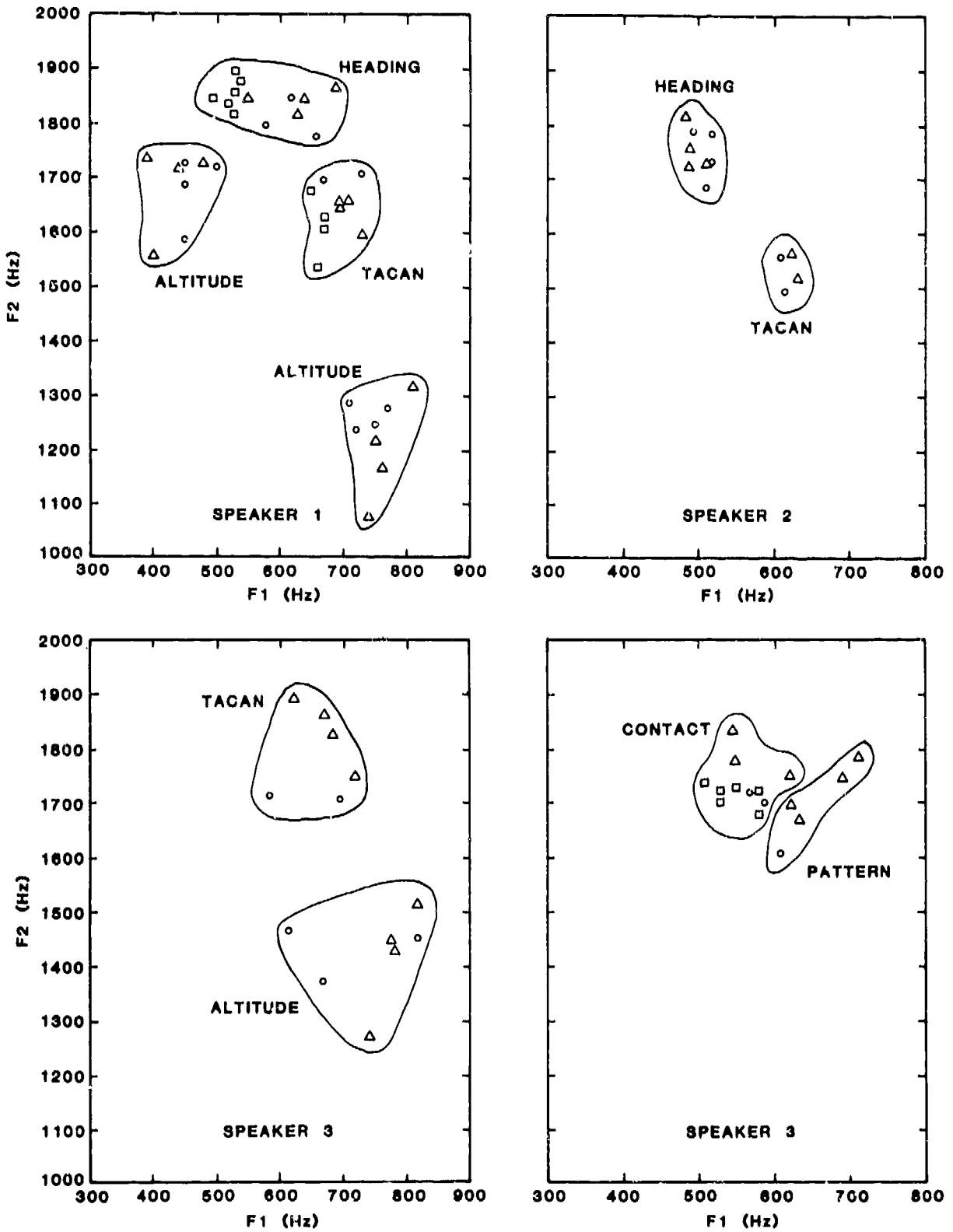

Figure 6. Examples of formant frequency data from several words produced by Speakers 1, 2, and 3 in the CONTROL condition (squares), the STATIC condition (triangles), and the DYNAMIC (circles). Formant measurements are obtained by the method indicated in Figure 5.

The principal conclusions to be drawn from the formant-frequency data are: 1) there is more variability in vowel formant frequencies in successive productions of words in the STATIC and DYNAMIC conditions than in the CONTROL condition, with F1 showing a range up to 100 to 200 Hz, and F2 a range up to 200 to 300 Hz; 2) for some speakers and some vowels there are systematic shifts in formant frequencies from the CONTROL condition to the experimental conditions, suggesting that there is a shift in "posture", possibly a change in positioning of the laryngeal structures.

Frication and aspiration noise; stop bursts

A number of speech sounds are produced by creating turbulence at a constriction in the vocal tract, thereby causing random noise to be generated in the vicinity of the constriction. The noise is usually called frication noise if it is generated primarily at a narrow constriction in the oral cavity, and is called aspiration noise if the vocal tract is relatively unconstricted, the glottis is somewhat open, and noise is produced primarily in the vicinity of the glottis. When a stop consonant is released, a burst of frication noise is usually generated in the vicinity of the constriction, and this may be accompanied by a transient acoustic excitation of the vocal tract as the intra-oral pressure is abruptly released.

These various kinds of sources of excitation of the vocal tract could potentially undergo some modification as the speaker is exposed to a stressful situation. For example, a change in the respiratory force, giving rise to a modification in the subglottal pressure, could influence the amplitude and spectrum of the turbulence noise. The detailed configuration of the constriction and the state of the surfaces forming the constriction could also influence the turbulence noise. The properties of the burst for a stop consonant could be affected by the abruptness of the stop release, which in turn may be determined by the state of the surfaces of the structures forming the constriction.

To investigate these possible effects, several measurements were made on noise portions of selected utterances in the corpus. One set of measurements was made on the noise bursts at the /t/ and /k/ releases in the word tacan. The method for obtaining quantitative measures of the burst is illustrated in Figure 7. In the case of the /t/ burst, the spectrum was sampled with a 25.6 Hamming window centered 5 msec following the release, and another spectrum was obtained with a similar window centered at the onset of the third glottal period following the beginning of normal voicing. Both spectra were smoothed using a 14-pole LPC procedure. The amplitude of the spectral peak nearest the fourth formant region in the smoothed spectrum was determined for each spectrum, and the differences between these amplitudes for the burst and for the vowel onset were obtained. A similar procedure was followed for the /k/ burst and the vowel immediately following this burst, except that in this case the amplitudes that were measured were the spectral peaks in the third formant region.

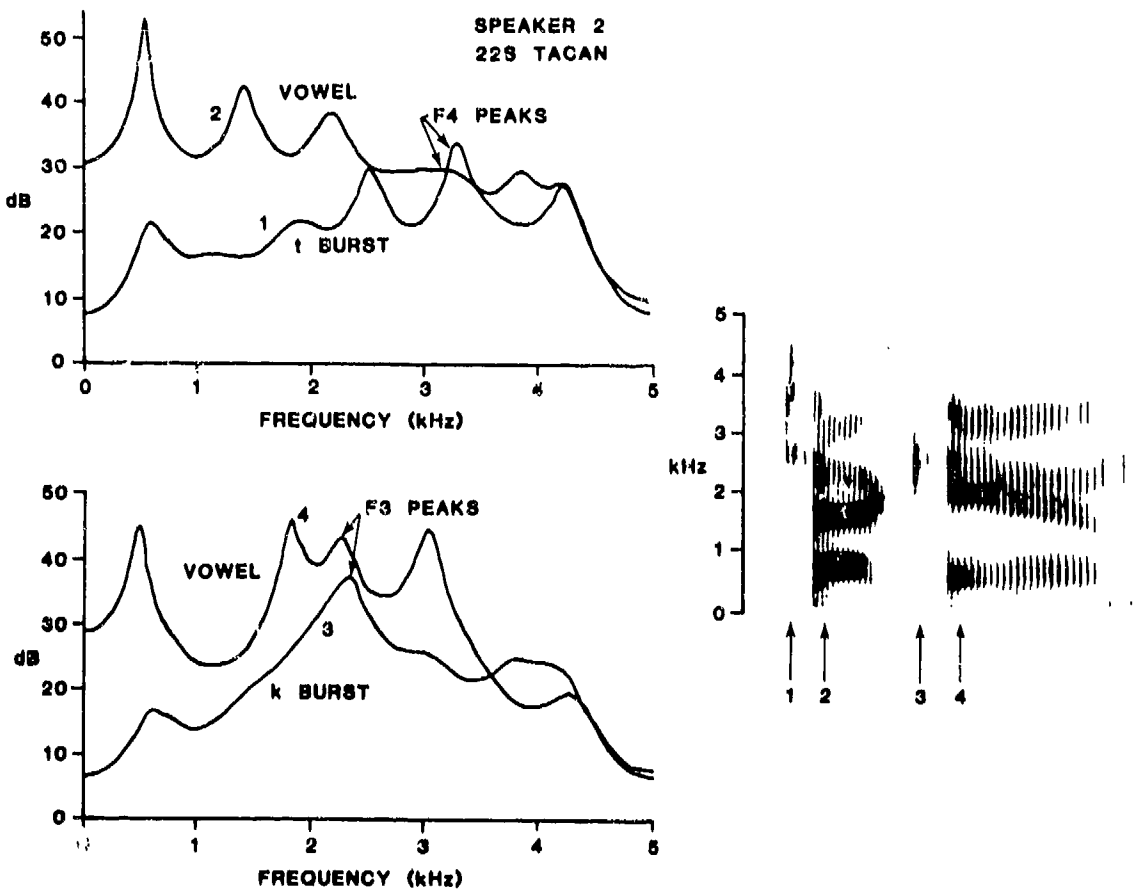

Figure 7. Illustration of the measurement of burst amplitude for the two stop consonants in the word tacan. The upper left panel shows LPC spectra sampled 1) immediately following (5 msec) the onset of the /t/ burst, and 2) at the third pitch period of the following vowel. The lower left panel shows similar spectra for the /k/ burst in the same word. The sampling points are indicated in the spectrogram at the right. The relative burst amplitude for the /t/ burst is defined as the amplitude of the F4 peak of the burst relative to that in the spectrum of the vowel (i.e., the F4 peaks shown in the upper left panel). In the case of the /k/ burst, the measure used is based on the amplitudes of the F3 peaks, as indicated.

Data for Speaker 2 from the CONTROL and experimental situations are summarized in Table 3. Two trends are evident in these data when one examines the amplitude of the burst relative to the adjacent vowel (the differences in the table). One is that the amplitudes of the bursts in relation to the vowels are, on the average, greater in the experimental conditions than in the CONTROL conditions. This difference is especially evident for the /t/ bursts. The other is that there is more variability in this relative burst amplitude in the experimental utterances than in the CONTROL utterances. A conclusion to be drawn from these data is that this speaker, in the experimental condition, is using a more abrupt release for the stop consonants (possibly with a higher respiratory effort) and he is more variable in the way he implements these articulatory and respiratory gestures.

Table III

Amplitudes of Spectral Peaks (in dB) in Burst and in Adjacent Vowel for the Stop Consonants in Several Repetitions of the Word Tacan by Speaker 2

Control Utterances	F4 peak in /t/	F4 peak in first /æ/	Difference	F3 peak in /k/	F3 peak in second /æ/	Difference
1	30	26	4	26	40	-14
2	29	31	-2	32	37	-5
3	29	33	-4	21	41	-20
4	29	30	-1	27	39	-12
5	29	34	-5	31	37	-6
6	27	33	-6	30	37	-7
7	30	37	-7	27	37	-10
8	30	35	-5	22	35	-13
mean			-3.2			-10.9
s.d.			3.3			4.6
Experimental Utterances						
18S	33	28	5	22	44	-22
22S	33	30	3	36	42	-6
28S	30	29	1	34	40	-6
31S	35	27	8	26	39	-13
10D	34	27	7	41	41	0
16D	33	26	7	31	41	-10
28D	38	29	9	42	41	1
33D	29	33	-4	38	44	-6
mean			4.5			-7.8
s.d.			4.1			6.9

There were substantial individual differences between speakers, both in the relative burst amplitudes for the two stops in tacan and in the amount of variability from one repetition to the next in the experimental conditions. For example, Table 4 shows that Speaker 4 had a relatively weak /t/ burst (in the F4 region) in relation to the F4 amplitude of the vowel, but was quite

Table IV

Mean and Standard Deviation of Relative Burst Amplitude (in dB) of /t/ and /k/ in tacan for Repetitions in Experimental Conditions, for Three Different Speakers. Method of Measuring Relative Amplitude is Shown in Figure 7.

	Speaker 2	Speaker 3	Speaker 4
Relative amplitude of /t/ burst	4.5	-5.8	-9.1
s.d.	4.1	5.8	3.3
Relative amplitude of /k/ burst	-7.8	9.3	3.9
s.d.	6.9	6.3	5.2

consistent in producing this burst amplitude (s.d. of 3.3 dB). On the other hand, this speaker (as well as Speaker 3) had a much higher amplitude /k/ burst than did Speaker 2.

In order to illustrate some other aspects of turbulence noise generation, measurements were made of aspiration noise in the initial /h/ in heading, and of the /p/ burst in pattern. Measurements of the /p/ burst were made in a manner similar to the /t/ and /k/ bursts in tacan (Figure 7). The amplitudes of the spectral peaks in burst and in vowel in the F3 region were used as the measure. In the case of the /h/, the noise spectrum was sampled at a point 15 msec prior to the onset of voicing in /ɛ/. The spectrum was smoothed (as before) and the amplitude at the peak in the F3 region was measured. The following vowel spectrum was sampled at the third glottal period, and again the F3 peak was measured. The data for three speakers are summarized in Table 5. Again we see large individual differences in the amplitude of the noise in relation to the vowel and in the amount of variability shown by the different speakers. Speaker 4 has the greatest relative amplitude of the noise, but also shows considerable variability in the noise generation mechanism.

Table V

Amplitudes (in dB) of Spectral peaks in Noise and in Following Vowel for /h/ in heading and /p/ in pattern, for Experimental Conditions. The Numbers in Parentheses represent the Number of tokens Measured.

	Speaker 2	Speaker 3	Speaker 4
Amplitude of F3 peak in /h/	20	16	26
Amplitude of F3 peak in /ɛ/	38	29	33
Difference	-18(8)	-13(3)	-7(8)
s.d. of difference	3.6	0.5	5.9
Amplitude of F3 peak in /p/	19	15	24
Amplitude of F3 peak in /æ/	34	35	30
Difference	-15(8)	-20(5)	-6(8)
s.d. of difference	4.3	3.9	6.4

In summary, then, these samples of data from bursts and aspiration noise show rather large fluctuations in the relative amplitude of the noise during the experimental conditions in which the speakers are placed under stress. Data from one speaker suggest that the variability is less when the test words are being repeated under more relaxed conditions, and that the mean values of the noise amplitude are different under these conditions. The data also indicate rather large individual differences in relative spectral amplitude of the noise for a given speech sound from one speaker to another.

Timing and segment durations

A number of measurements were made of the time between various acoustic boundaries in several of the words. The aim was to determine whether there were systematic changes in the temporal characteristics of the utterances between the CONTROL condition and the two experimental conditions. The measurements included the durations of stressed vowels in syllable-initial position, durations of vowels with secondary stress in syllable-final

position, voice-onset time for voiceless stop consonants, and the duration of the stop gap for medial voiceless stop consonants.

Some of the results are summarized in Table 6. In general, there were no systematic differences in durations for the STATIC and DYNAMIC conditions, and consequently the data for these two experimental conditions were averaged. The main effects emerging from these data are: 1) final vowels in the words are consistently longer in the experimental conditions than in the CONTROL condition; and 2) there is greater variability in vowel durations in the experimental conditions. These two results are presumably a consequence of the fact that the CONTROL words were read rather rapidly as a list (each word preceded by say), with the speaker in a relatively stable physiological state. The data also show greater variability in final vowel durations than in initial vowel durations, both in the experimental and CONTROL conditions. This result is not unexpected, since the amount of syllable-final lengthening occurring in this situation is presumably not well controlled. Apparently, as long as some final lengthening is produced by a speaker to mark the end of an utterance, the actual amount of lengthening is not crucial.

Further detailed examination of the data in Table 6 indicates trends that characterize the timing of some speakers but not others. For example, there is a tendency for Speaker 3 to lengthen voice-onset time (VOT) in the experimental conditions, whereas Speaker 1 tends to shorten it. There are also substantial differences in the way the consonant /k/ preceding the second vowel in tacan is produced. For two speakers (3 and 4), the VOT is long relative to that for the initial /t/, whereas for the other two speakers, the reverse is true.

In summary, these data on timing provide additional evidence for the observation that speakers show more variability when they are placed under stress in the experimental conditions than in the more relaxed CONTROL condition. There is little evidence (beyond the syllable-final lengthening), however, for significant shifts in timing strategies between the various CONTROL and experimental situations. Some shifts in durations are exhibited by some speakers when they are in the experimental situation, but these changes in timing are not very large.

MODIFICATIONS OF INDIVIDUAL PHONETIC ELEMENTS

A number of modifications were observed in individual phonetic elements in repetitions of words in the STATIC and DYNAMIC conditions. These modifications are usually the result of application of optional low-level phonetic rules in English. Apparently when a speaker is producing utterances under various amounts of stress, fluctuations in the application of these rules give rise to phonetic modifications. Most of the phonetic modifications described below were observed on the spectrograms made of the utterances.

Medial unstressed vowel in "altitude"

The medial unstressed vowel in the word altitude lies between two

Table VI

Average Durations of Various Speech Events for Words Produced in CONTROL and Experimental Situations. Values are in Milliseconds. Numbers in Parentheses are Number of Tokens Measured. For Experimental Conditions, Data from STATIC and DYNAMIC Situations are Averaged.

		CONTROL		EXPERIMENTAL	
		Duration	s.d.	Duration	s.d.
Stressed vowels					
Speaker 1 /æ/	tacan	126	6 (4)	131	9 (5)
/æ/	pattern	135	7 (6)	157	6 (4)
/oɪ/	holding	148	9 (9)	168	16 (4)
Speaker 2 /æ/	tacan	115	6 (5)	115	9 (7)
/oɪ/	holding	172	8 (5)	158	19 (8)
Speaker 3 /æ/	tacan	146	20 (5)	142	17 (5)
Speaker 4 /æ/	tacan	134	6 (4)	157	13 (8)
/oɪ/	holding	201	7 (4)	209	17 (8)
	mean	147	9	155	13
Final vowels					
Speaker 1 /æ/	contact	173	9 (4)	204	30 (7)
/ɪŋ/	holding	134	11 (8)	239	48 (6)
Speaker 2 /æ/	tacan	239	13 (5)	307	26 (7)
/ɪŋ/	holding	106	28 (5)	234	23 (7)
Speaker 3 /æ/	tacan	219	28 (5)	236	27 (6)
Speaker 4 /æ/	tacan	260	6 (4)	284	29 (8)
/ɪŋ/	holding	157	11 (4)	235	30 (8)
	mean	184	15	248	30
Stop gap					
Speaker 1 /k/	tacan	71	0 (4)	85	9 (6)
Speaker 2 /k/	tacan	65	8 (5)	68	6 (7)
Speaker 3 /k/	tacan	74	6 (5)	88	4 (5)
Speaker 4 /k/	tacan	88	8 (4)	81	7 (8)
	mean	74	6	81	7
Voice-onset time					
Speaker 1 /t/	tacan	67	5 (4)	55	14 (5)
/k/	tacan	29	3 (4)	24	7 (5)
/p/	pattern	41	10 (6)	46	13 (4)
/k/	contact	60	10 (4)	49	6 (4)
Speaker 2 /t/	tacan	57	8 (5)	50	7 (7)
/k/	tacan	39	16 (5)	45	5 (7)
Speaker 3 /t/	tacan	58	6 (5)	62	20 (5)
/k/	tacan	74	6 (5)	88	4 (5)
Speaker 4 /t/	tacan	54	12 (4)	46	10 (8)
/k/	tacan	88	8 (4)	81	7 (8)
	mean	57	8	55	9

voiceless consonants and, in English, such a vowel is subject to an optional rule: devoicing. Examples of two utterances of this word by the same speaker - one with the vowel voiced and the other devoiced - are shown in Figure 8. Of the 28 utterances of this word by the four speakers in the STATIC and DYNAMIC conditions, the medial vowel was voiceless in seven. There was no significant difference in the number of devoiced vowels in the two experimental conditions. Only one speaker (Speaker 4) was consistent in voicing all of these unstressed vowels. The least consistent speaker was Speaker 3, who devoiced the vowel in one-half of the utterances. In cases where the vowel was voiced, there was often considerable variability in the duration of voicing. For example, in the five utterances by Speaker 1 in which the vowel was voiced, the number of glottal vibrations occurring during the open phase of the vowel varied from two to five.

Prevoicing of initial voiced stop in "bogey"

An initial voiced stop in English may be prevoiced. In the list of words used in this study there was just one initial voiced stop - the /b/ in bogey. Of the 34 utterances examined for the two experimental conditions, seven showed prevoicing. Only one speaker (Speaker 3) failed to prevoice any of the utterances. There was a slight tendency for more of the /b/'s to be prevoiced in the DYNAMIC condition (five prevoicings versus two in the STATIC condition). Of the utterances that were prevoiced, there was some variation in the amplitude and duration of the prevoicing.

Release of final stops

Two of the test words ended in stop consonants (altitude and contact), and these words provided an opportunity to examine the speakers' habits with regard to final stop release. Of the 57 versions of these two words that were examined, the final stop was released in 39. Most of the unreleased stops were the voiced final stop in altitude. Some speakers consistently released the stop in contact but no speaker was consistent in releasing the stop in altitude. There was no significant difference in the incidence of final stop releases for the STATIC and DYNAMIC conditions. When the final stop was released, there was considerable variability in the amplitude of the burst at release. Figure [illegible] shows spectrograms of three versions of contact produced by Speaker 1. In 34S, there is no stop release, in 23S there is a weak release, and in 14D a strong release. Furthermore, in 23S the release is clearly an alveolar release, whereas in 14D, the burst shows that the release is from the velar position. For this word with a final stop cluster, then, we observe this additional variability in the place of consonant release. In none of the utterances of this word were there two successive releases corresponding to the two different places of articulation for the consonant sequence.

Velar stop before an unstressed vowel

In the word bogey the velar stop /g/ occurs before an unstressed vowel. Observation of spectrograms of this word indicated, for some utterances of

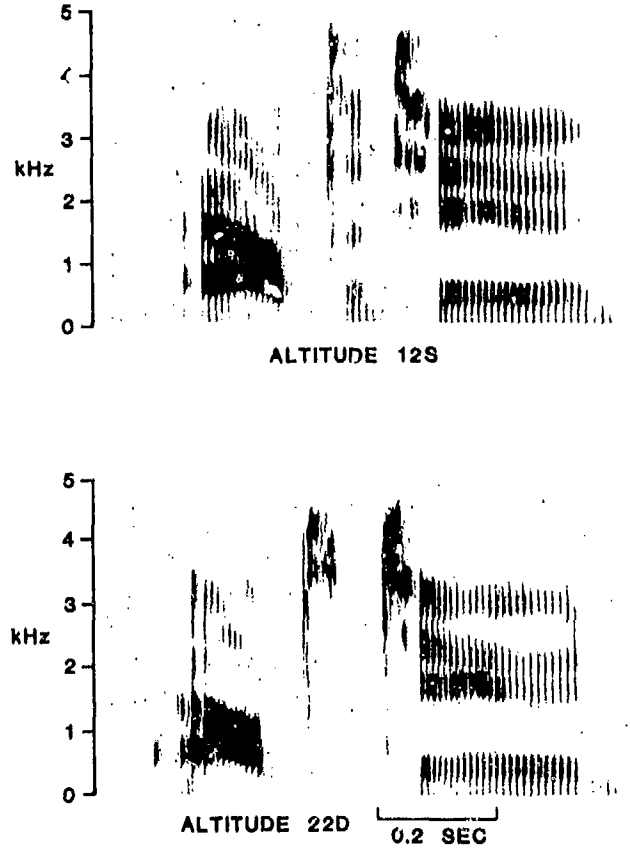

Figure 8. Spectrograms of the word altitude produced by Speaker 2 during the experimental session. These spectrograms illustrate a version of the word in which the unstressed vowel is devoiced (lower panel) and one in which there is a brief interval of voicing in the vowel (upper panel).

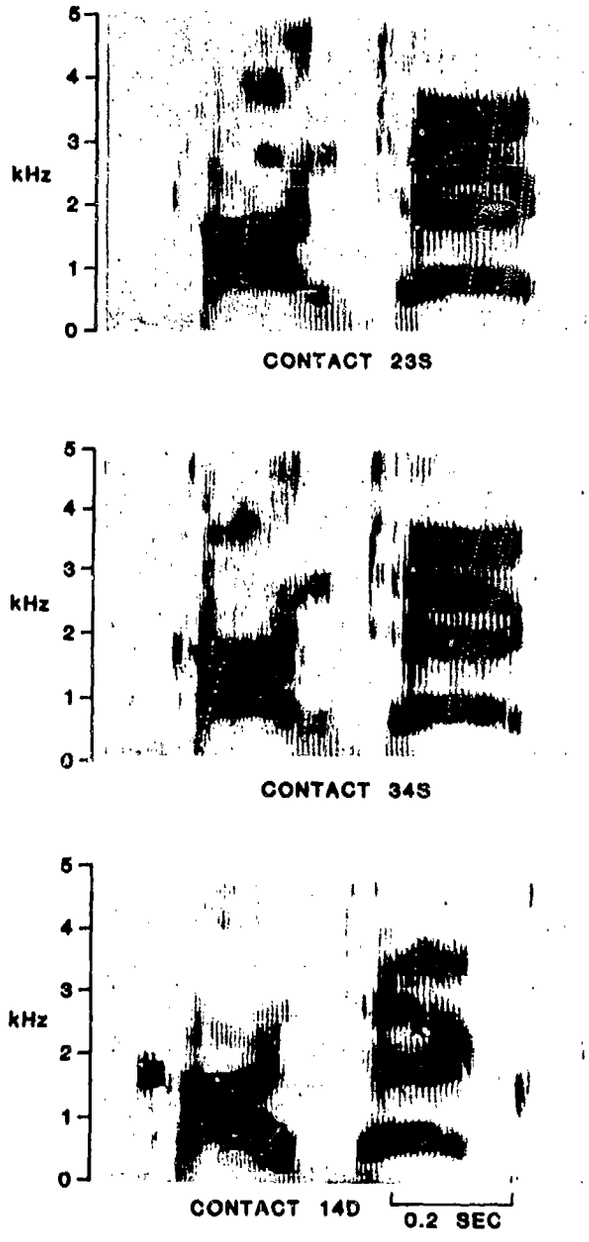

Figure 9. Spectrograms of three versions of the word contact produced by Speaker 1 during the experimental session. These spectrograms illustrate three kinds of release of the final consonants: an alveolar release (top panel), no release (middle panel), and a velar release (bottom panel). Also evident on these spectrograms are varying amounts of nasalization in the vowel preceding the nasal consonant: substantial nasalization (bottom panel), and a small amount of nasalization (top panel).

the word, the /g/ was produced without a complete closure of the vocal tract, so that a velar continuant resulted. Intermediate cases between a stop and a continuant were also observed. This "weakening" of a velar stop is not unexpected, since there is no opposition in English between a velar stop and a velar fricative, and consequently there is not a strong motivation to represent this distinction in the sound wave.

Two examples of the word bogey produced by Speaker 2 are shown in Figure 10. In one of these utterances (5D), there appears to be a complete velar closure, whereas in the other (12D), the continuation of the second formant is evidence that the closure was not complete.

Of the 33 versions of bogey examined in the two experimental conditions, observation of the spectrograms indicated that about one-half of the velars were produced with a complete closure and a release burst, and about one-half showed no evidence of a complete closure. All speakers produced utterances of both types, and there was not a significant difference in the incidence of the two types in the STATIC and DYNAMIC conditions.

Nasalization of vowels preceding nasal consonants

In English and many other languages when a nasal consonant follows a vowel, some anticipatory nasalization is produced in the vowel. The amount and duration of this nasalization depends to some extent on the phonetic context, but some variation is to be expected within the utterances of one speaker and from one speaker to another.

The list of test words contains a number of examples of vowels followed by nasal consonants: contact, heading, holding, monitor, pattern, and tacan. The acoustic correlates of nasalization are not understood sufficiently to make it possible to establish with any precision from the sound wave the time at which velopharyngeal opening occurs, or the degree of velopharyngeal opening. The general acoustic correlates of nasalization for vowels are 1) a weakening of the first formant, 2) a broadening of the first formant, 3) introduction of an additional "nasal" formant in the vicinity of the first formant, and 4) a shift in the frequency of the first formant. From observation of spectrograms, we have attempted to determine the time at which nasalization becomes evident in each of the words listed above. This procedure is clearly subject to some uncertainty, and the results must be interpreted with this in mind.

Examples of words in which a final vowel precedes a nasal consonant are shown in Figure 11. In the case of 7S, there is evidence of nasalization in the vowel immediately following the release, whereas in 6D acoustic evidence for nasalization does not appear until about 50 msec following the /d/ release. Evidence for variability in nasalization of a vowel preceding a nasal in intervocalic position can be observed in the spectrograms of the word contact, shown earlier in Figure 9. In 23S there is little nasalization in the vowel, and the onset of the nasal consonant is rather abrupt. For 14D, on the other hand, nasalization can be seen over about one-half of the vowel

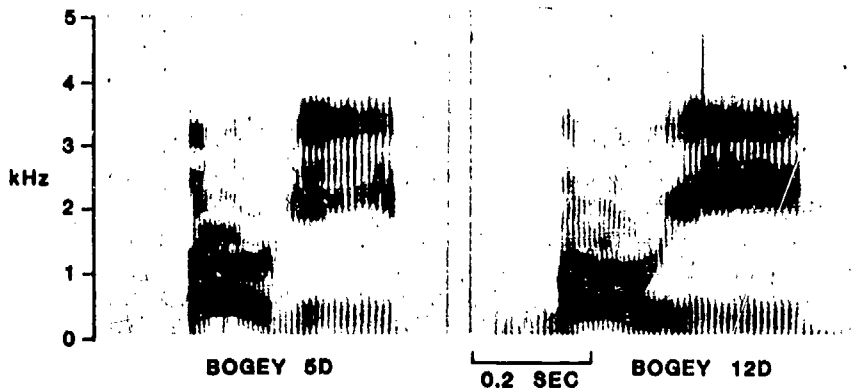

Figure 10. Spectrograms of two versions of the word bogey produced by Speaker 2 during the experimental session. The version at the left illustrates a token in which the pre-unstressed /g/ is produced with essentially complete vocal-tract closure, whereas for the right-hand utterance complete closure is not achieved.

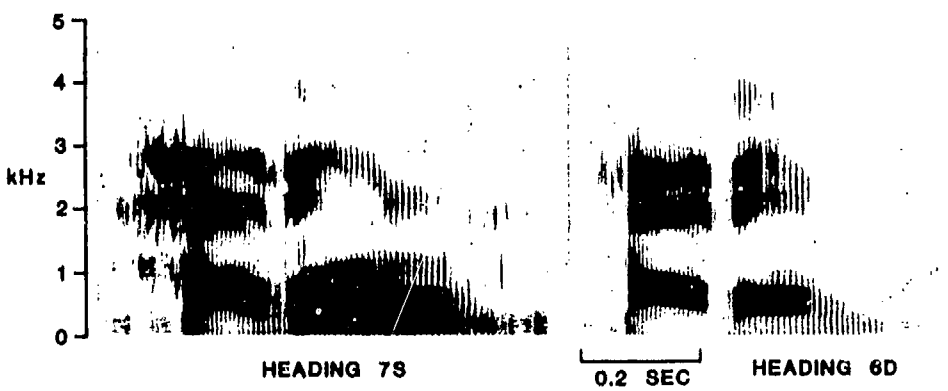

Figure 11. Spectrograms of two versions of the word heading produced by Speaker 4 during the experimental sessions. These spectrograms illustrate different degrees of nasalization for the final vowel, with the token at the left showing the greater amount of nasalization. Other examples of vowel nasalization can be seen in Figure 9.

duration (as evidenced by the appearance of an additional resonance at low frequencies) but for 34S there is an intermediate degree of nasalization.

There are large individual differences in the average time from consonant release to onset of nasalization in syllable-final vowels (such as heading, tacan, etc.). These average times are about 38, 54, 17, and 22 msec, respectively, for Speakers 1, 2, 3, and 4. Within a given speaker there are also substantial differences in onset of nasalization (as in the examples in Figures 9 and 11), but there are no consistent differences for the DYNAMIC compared with the STATIC conditions. The kind of variability that exists for a particular speaker is illustrated in Figure 12. For all of the utterances with a final nasal consonant produced by Speaker 4 (32 in all), the percent of the final vowel duration that was nasalized (as determined by spectrographic observation) was measured and displayed as a distribution in the figure. The distribution of this percentage across all 32 utterances is very broad, indicating that this speaker shows a large amount of variability in timing the onset of nasalization in these vowels when he is operating in the stress-producing experimental conditions.

Weakening of alveolar stops and nasals in pre-unstressed position

Several of the phonetic modifications discussed above are related to the "weakening" that occurs in the components of a syllable when it is unstressed. When an alveolar stop or nasal precedes an unstressed vowel, this weakening can take the form of flapping of the consonant. Within the list of utterances that were used in this study, there are six places where this optional weakening of an alveolar consonant can occur. These are listed as follows: heading, holding, monitor, pattern, and radar.

Examination of the spectrograms showed that there were at least three different ways of producing the stop consonants in these words: 1) as a full-fledged stop with a well-defined burst; 2) as a flap with a brief closure interval (less than about 20 msec); 3) as a continuant with no evidence of closure. Examples of two of these manifestations of pre-unstressed stop consonants are given in Figure 13, and the third can be seen in Figure 11, item 6D. The stop in monitor was almost always produced with a full-fledged stop closure by the four speakers (except on one occasion by Speaker 2). All the other alveolar stops in pre-unstressed position were produced in various ways by the different speakers, and no speaker was completely consistent in producing one of these words in the same way each time during the experimental session.

Likewise the /n/ in monitor was produced in several ways: 1) a continuant with no evidence of nasal closure, 2) a nasal flap, 3) a non-nasal flap, i.e., with no velopharyngeal opening, and 4) a full-fledged nasal consonant. Examples of some of these manifestations of this pre-unstressed nasal consonant are shown in Figure 14. Most speakers produced more than one of these versions of /n/ in the course of the experimental session.

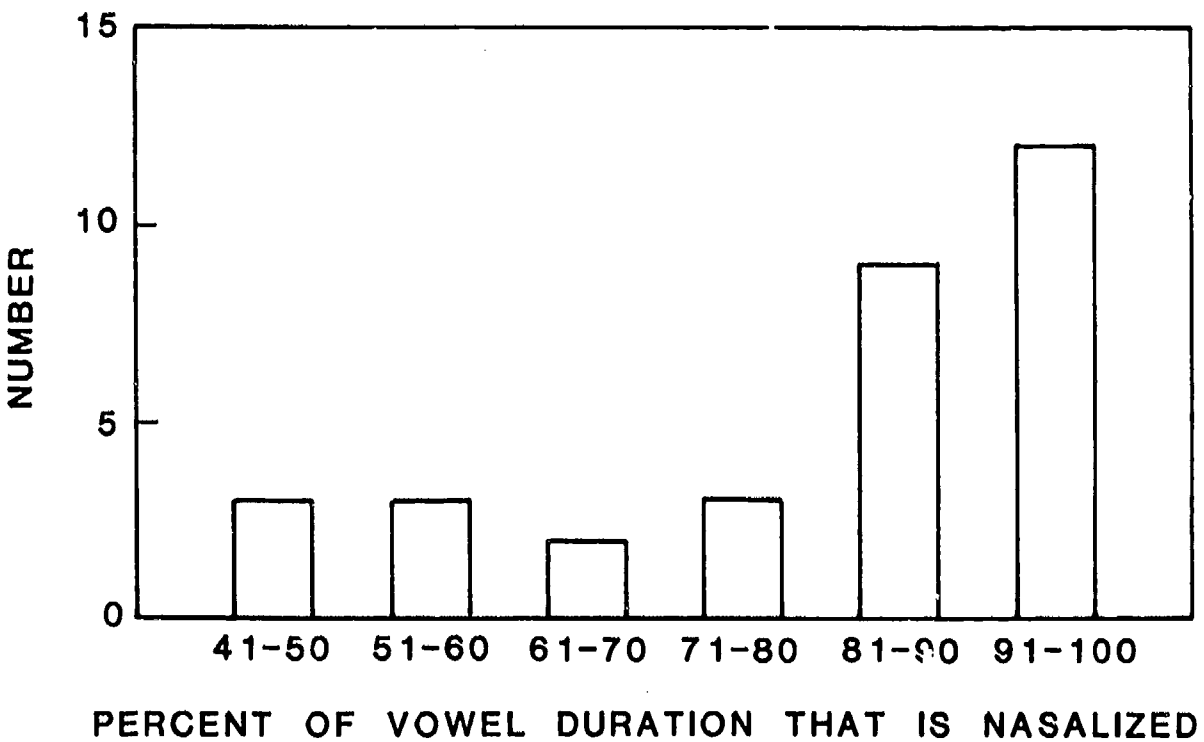

Figure 12. Degree of nasalization exhibited by Speaker 4 in vowels occurring in final syllables with a following nasal consonant. For each of 32 such vowels (in the words heading, holding, pattern, and tacan), the percent of the vowel duration that was nasalized was estimated from the spectrograms. The figure represents a distribution of these 32 measurements.

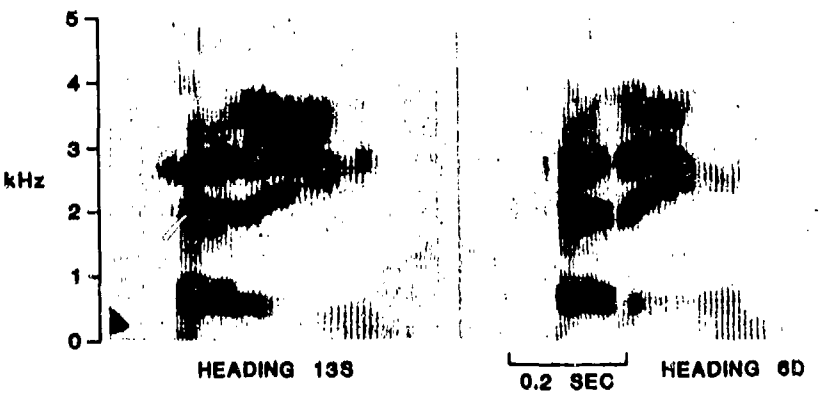

Figure 13. Spectrograms of two versions of the word heading produced by Speaker 1 during the experimental session. For the token at the right, the /d/ is flapped, whereas in the left-hand token there is little evidence for an alveolar closure. A more complete /d/ closure is illustrated in the right-hand spectrogram in Figure 11.

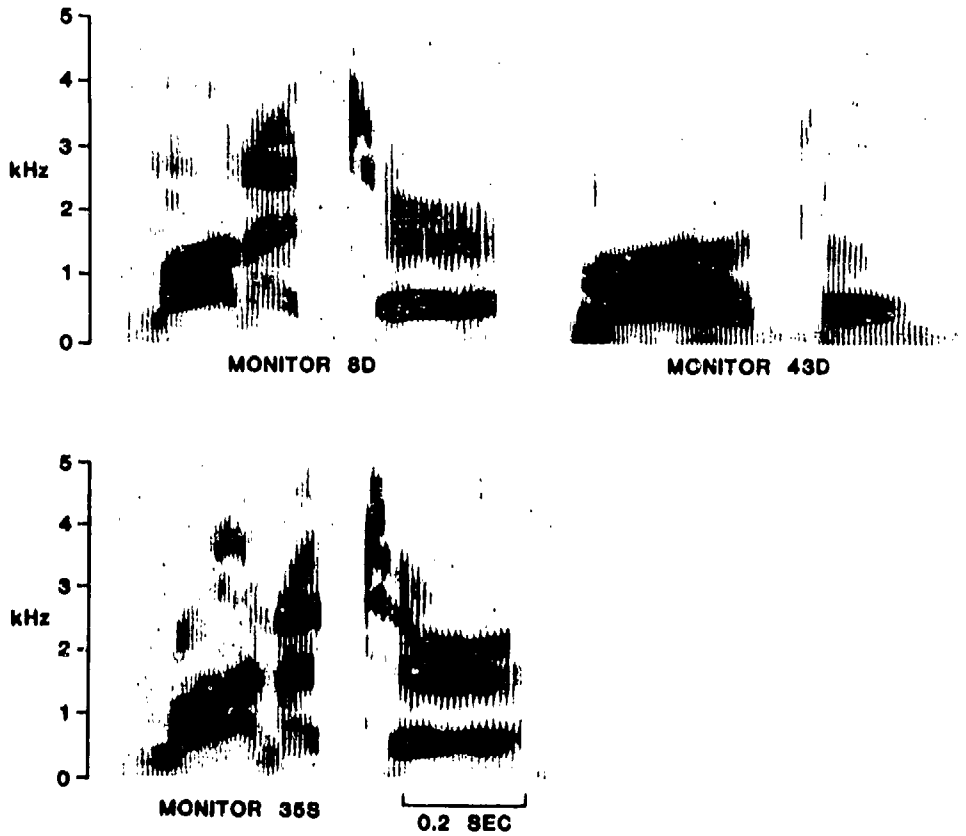

Figure 14. Spectrograms of three versions of the word monitor, illustrating three different acoustic manifestations of the pre-unstressed /n/. For the upper right token (produced by Speaker 4), there is no /n/ closure, and the only evidence for the nasal consonant is some nasalization during the vowel (slight weakening of F1). The other two tokens (produced by Speaker 1) show examples of a complete /n/ closure (lower left) and of a flapped /n/ (upper left).

CONCLUSIONS

Two general observations can be made from the data presented: 1) some speakers will show consistent shifts in the properties of their speech when a task requires them to perform under stress; and 2) most speakers will exhibit more variability in the properties of their speech sounds when performing under stress than they will when performing under more relaxed and controlled conditions.

The first observation points to some consistent acoustic changes for which data should be extracted in future experiments:

1) shifts in fundamental frequency in an upward direction, particulary for stressed vowels;

2) changes in the waveform of the glottal pulses, leading to modifications in vowel spectra, particularly in the lowest one or two harmonics;

3) shifts in the first two formants of vowels, primarily in the direction of increased F1 and F2 for the samples observed in this analyses; and

4) changes in the amplitude of turbulence noise in relation to the adjacent vowel for certain stop consonants; this change appears to be in the direction of increased relative amplitude of the noise for stressful situations.

The second observation appears to be more germane to the development of interactive voice systems, and that is the significant variability of words produced by a speaker in a stress-inducing situation. Apparently, because a number of production options are available for particular speech sounds in varying phonetic contexts, a speaker is less consistent in the choice of options in conditions of stress. For stressed syllables, options arise because certain acoustic characteristics are not required to make phonetic distinctions in English, and are free to vary. Similarly, "weakening" in the production of certain phonetic elements occurs because the production of the elements are made with less force or with a less precise realization of the ideal target states for the articulators.

We observed examples of such options in the production of stress or relatively strong syllables: 1) pre-voicing or lack of pre-voicing for word-medial voiced consonants; 2) the degree of nasalization that precedes a vowel; and 3) the amount by which a final vowel is lengthened.

For the task involved in this study, "weakening" or a decrease in the force or preciseness of articulation probably accounts for most of the variability observed. Significant variability in four types of weakening were observed: 1) variability in the amount of devoicing of an unstressed vowel between two voiceless consonants; 2) variability in the release of stop consonants; 3) variability in the degree to which a pre-unstressed

velar stop becomes a continuant; and 4) variability in the degree to which pre-unstressed alveolar nasals or stops become flapped or become even weaker than flaps.

Three implications can be drawn from these results for the use of speech recognition systems in aircraft. First, and most obvious, is the importance of "training" recognition devices while the operator is not in a relaxed, stress-free environment. A common practice among users of recognizers is to train the device in the work space and during task performance. Updating the reference patterns at time intervals after initial training is also a common procedure. These practices increase the robustness of the device for changes in the acoustic structure of the speakers' utterances and changes in the ambient environment. Using similar procedures for aircraft environments will be difficult to realize since the variability of the environment and task requirements exceeds what is encountered in most common work spaces. Obtaining the number of training tokens of each utterance required to account for each speaker's variability in each different stressful situation would be unwieldly and tedious.

A second implication suggests training the users of interactive voice systems to speak in a uniform manner under all conditions, even conditions of stress. Users would be made aware of the phonetic modifications likely to occur under different conditions and would be trained to minimize them. Again, such a procedure will be hard to realize in aircraft because of the rapidly changing physical, cognitive, communicative, and psychological demands on aviators.

A third implication suggests that the most effective approach to the problem of changes in the speech signal due to stress is to design a speech recognition device capable of dealing with speech variations more directly. Since any recognizer designed for use in aircraft will use a finite vocabulary, grammar, and syntax, an integral part of such a device would be a set of rules specifying the possible acoustic modifications given the phonetic description of the word. A set of rules would indicate, for example, that prevoicing of an initial voiced stop consonant is an optional property providing evidence for the voicing feature, but is not a necessary acoustic correlate of this class of consonants in English. Within the context of the aviation environment, a large number of such rules would be incorporated into the recognition system to indicate alternative realizations of a word. Some rules would be specific, but many would be general in form, and would describe phonetic tendencies applicable to a large number of lexical items.

In summary, the kinds of changes and variability observed in the vocal utterances of speakers in a stress-producing situation serve to emphasize the limitations of speech recognition systems based on template matching. There is a need for a better understanding of these phonetic modifications and for developing ways of incorporating knowledge about these changes into speech recognition systems.

REFERENCES

1. Nye, J. M., Human factors analysis of speech recognition systems. Speech Technology, Vol. 1, No. 2, 50-57, 1982.

2. Mosko, J. D., Remington, R. W., and Griffin, G. R., Phonological variants in medial stop consonants under simulated operational environments: implications for voice activated controls in aircraft. In: Aural Communication in Aircraft. Advisory Group for Aerospace Research and Development (AGARD) Conference Proceedings 311. London: Technical Editing and Reproduction Ltd., Pp 3-1 - 3-4, 1981.

3. Mosko, J.D., and Griffin, G. R., Clear speech: A strategem for improving radio communications and automatic speech recognition in noise. In: Aural Communication in Aircraft. Advisory Group for Aerospace Research and Development (AGARD) Conference Proceedings 311. London: Technical Editing and Reproduction Ltd., Pp 4-1 - 4-4, 1981.

4. Lentz, J. M., and Guedry, Jr., F. E., Motion sickness susceptibility: A retrospective comparison of laboratory tests. Aviat. Space and Environ. Med., 49:1281-1288, 1978.

5. Markel, J. D. and Gray, A. H., Linear Prediction of Speech. New York: Springer-Verlag, 1976.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NAMRL - 1300	2. GOVT ACCESSION NO.: AD-A135 932
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: Interactive Voice Technology: Variations in the Vocal Utterances of Speakers Performing a Stress-Inducing Task	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): James D. Mosko, Kenneth N. Stevens, and Glenn R. Griffin	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Aerospace Medical Research Laboratory, Naval Air Station, Pensacola, Florida 32508	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62758N MF58523 MF5852801B 0001
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Medical Research and Development Command, Naval Medical Command, National Capital Region, Bethesda, Maryland 20814	12. REPORT DATE: 16 August 1983
	13. NUMBER OF PAGES: 36
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

James D. Mosko, Ph.D. is currently at the Rome Air Development Center, Griffis Air Force Base, New York

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Speech Recognition
Interactive voice technology
Speech communication
Stress effects on speech
Motion stress
Vocal variability
Acoustical analysis of speech

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

As plans develop to incorporate interactive voice systems in advanced military aircraft, it is important to determine how the speech of a pilot is likely to be modified in stressful operational situations. Speech recognition devices must be designed to respond correctly to commands produced under a variety of conditions.

Acoustical analyses were conducted of words produced by four speakers in

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

a motion stress-inducing situation. The aim of the analyses was to document the kinds of changes that occur in the vocal utterances of speakers who are exposed to motion stress and to comment on the implications of these results for the design and development of voice interactive systems.

The speakers differed markedly in the types and magnitudes of the changes that occurred in their speech. For some speakers, the stress-inducing experimental condition caused an increase in fundamental frequency, changes in the pattern of vocal fold vibration, shifts in vowel production and changes in the relative amplitudes of sounds containing turbulence noise. All speakers showed greater variability in the experimental condition than in a more relaxed control situation. The variability was manifested in the acoustical characteristics of individual phonetic elements, particularly in speech sounds occurring in the vicinity of unstressed syllables. The kinds of changes and variability observed serve to emphasize the limitations of speech recognition systems based on template matching of patterns that are stored in the system during a training phase. There is need for a better understanding of these phonetic modifications and for developing ways of incorporating knowledge about these changes within a speech recognition system.