XXIV. SPEECH COMMUNICATION*

Academic and Research Staff

Prof. K. N. Stevens	Prof. D. H. Klatt	Dr. Margaret Bullowa
Prof. M. Halle	Prof. A. V. Oppenheim	Dr. Paula Menyuk
Prof. W. L. Henke	Dr. A. W. F. Huggins	Dr. H. Suzuki†

Graduate Students

M. E. Barron	M. F. Medress	R. M. Sachs
A. J. Goldberg		J. J. Wolf

RESEARCH OBJECTIVES

The broad aim of our research in speech communication is to gain an understanding of the nature of the processes of human speech production and perception. A practical goal is to utilize knowledge gained through study of these processes to devise procedures that will permit limited communication between men and machines by means of speech. Several projects directed toward these goals are active at present. Studies of the relations between articulatory configurations and the speech sounds generated by these configurations, as well as experiments on the perception of speechlike sounds, are providing new evidence for the existence of discrete categories of speeech sounds and are leading to further examination and modification of the system of distinctive features underlying speech events. Studies of the production and perception of speech sounds by children, by using spectrographic techniques and tests involving the responses of children to natural and synthetic speech stimuli, are providing some insights into the process whereby language is acquired. Other projects are devoted to problems of automatic speech recognition, speaker recognition, and the development of visual displays of speech information. Much of our research is now being carried out with the aid of a new digital computer facility and associated peripheral equipment, including a graphical input, displays, a filter bank, a speech synthesizer, and other facilities. We are continuing to improve and develop this system, and we expect in the forthcoming year to use the computer facility for further studies of the simulation of articulatory processes in speech and for experiments on psychoacoustics and speech perception, as well as for projects of the kind outlined above.

K. N. Stevens, M. Halle

A. VISUAL PERCEPTION OF SPEECH STIMULI

An investigation of the perception of visible speech patterns is being conducted, at present, to determine what aspects of a visual display will prove most useful for visual recognition of auditory speech stimuli. Potter, Kopp, and Green[1] reported that subjects could be taught to distinguish approximately 300 word patterns in approximately 90 hours of training on a machine that produced a real-time spectrogram similar to that of Fig. XXIV-1. House[2] has reported, however, that on a similar device subjects that

*This work was supported principally by the U. S. Air Force (Electronics Systems Division) under Contract AF 19(628)-5661; and in part by the National Institutes of Health (Grant 2 RO1 NB-04332-06), the Joint Services Electronics Programs (U.S. Army, U. S. Navy, and U. S. Air Force) under Contract DA 28-043-AMC-02536(E).

†On leave from Tohoku University, Sendai, Japan.

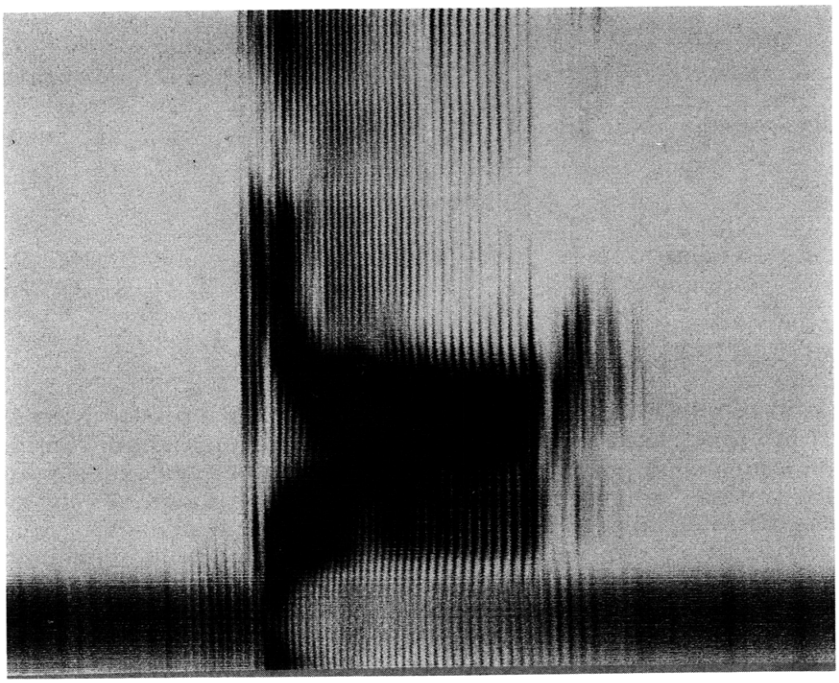


Fig. XXIV-1. Wideband spectrogram (0-3500 Hz) of /ga/. Time is plotted along the horizontal axis and frequency along the vertical axis.

were asked to distinguish between the eight stimuli /tʊ/, /tɛ/, /tʌ/, /tɪ/, /tæ/, /tu/, /ti/, /tɑ/ spoken by one male speaker could do so correctly only 75% of the time after being trained for several hours.

In the present study a real-time spectral input system[3] coupled with a PDP-9 computer is used to obtain the spectral information at 5-msec intervals that are necessary for the generation of the displays. This information, together with other relevant data, is stored on magnetic tape for later playback to the subjects.

Several types of displays are now being studied. In one display a single sampled spectrum is shown on a screen, as indicated in Fig. XXIV-2, and is updated at regular time intervals. The rate of presentation of the spectra can be varied by the subject. In another display, 9 spectra are displayed on a screen in the manner shown in Fig. XXIV-3 with each sampled spectrum being replaced at the end of each time interval by the next spectrum in the sequence. Consequently, the spectra move down in the picture with time and disappear off the screen, with new ones appearing at the top of the viewing area. The intensity of each point displayed is proportional to the spectral amplitude corresponding to that point. The effect of all of this movement is the illusion by the subject of complex wave motion.

Finally, a last display designed primarily for the perception of vowels and of consonant-vowel transitions (particularly stop consonants) is shown in Fig. XXIV-4.

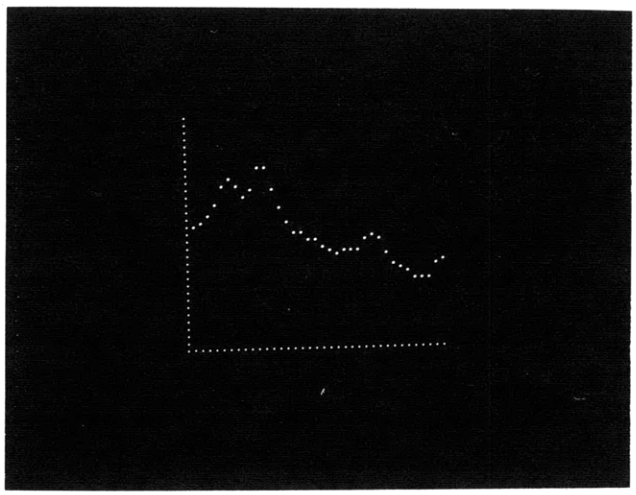


Fig. XXIV-2.

Sampled spectrum of /a/ in utterance /ga/. Frequency (0-7000 Hz) is plotted along the horizontal axis and amplitude in dB (0-64 dB) along the vertical axis.

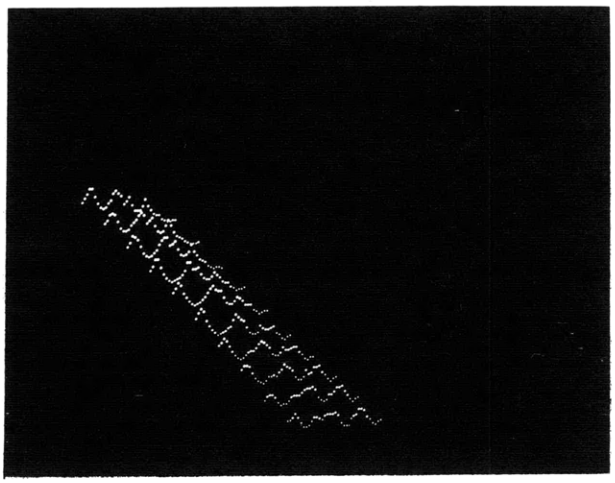


Fig. XXIV-3.

Several sampled spectra of /ga/ plotted simultaneously.

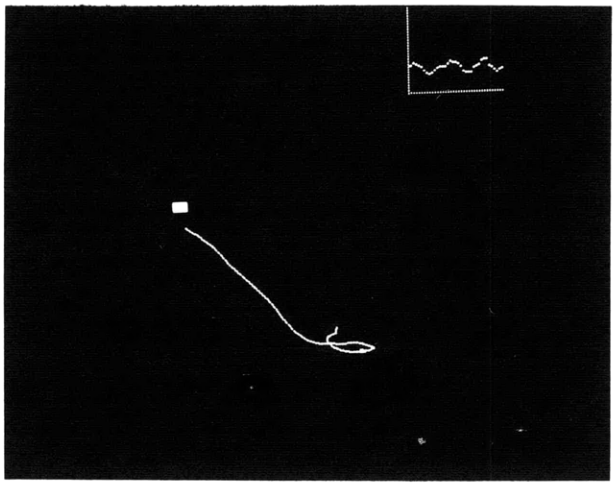


Fig. XXIV-4.

Formant 1 vs formant 2 of /ga/ plotted with time as a parameter. Solid bar indicates duration of stop consonant. The spectrum in insert is an average of all spectra from the beginning of the burst of the stop consonant to the start of voicing.

Here the length of the bar is proportional to the length of time from the beginning of the burst of the stop to the beginning of voicing. Formant 1 of the utterance is plotted along the horizontal axis, and formant 2 along the vertical, both as a function of time. The contour builds up in time, new segments being added as successive time intervals are sampled. Consequently, rapid formant motion at the beginning of the vowel is indicated by rapid motion of the trace and the relative steady-state condition of the mid portion of the vowel by a slowly moving trace. The spectrum appearing in the upper corner is an average of all spectra from the beginning of the burst to the start of voicing. The formant extraction for this display is done semiautomatically with errors edited out by the experimenter before presentation to the subject. These formant data are then stored on magnetic tape along with the other spectral information.

Very preliminary results indicate that subjects can easily obtain accuracies higher than 90% after only a few hours training with the displays described by Figs. XXIV-3 and XXIV-4 when asked to identify 8 types of stimuli similar to those used by House.[2]

Further work is planned to determine the discriminability of subjects to visual representations of simple sounds when spoken by one or more speakers using these and other displays, in order to determine which aspects of the displays are most useful for visual discrimination.

A. J. Goldberg

References

1. R. Potter, G. Kopp, and H. Green, Visible Speech (D. Van Nostrand Company, Inc., New York, 1947).
2. A. House, D. Goldstein, and G. Hughes, "Perception of Visual Transforms of Speech Stimuli and Learning Simple Syllables," Am. Ann. Deaf, March 1968, pp. 215-221.
3. N. Benhaim and Eleanor C. River, "Real-Time Spectral Input System for Computer Analysis of Speech," Quarterly Progress Report No. 84, Research Laboratory of Electronics, M.I.T., January 15, 1967, pp. 253-254.

B. COMPUTER RECOGNITION OF SINGLE-SYLLABLE WORDS SPOKEN IN ISOLATION

1. Introduction

A primary objective of verbal communication is the transmission of information from speaker to listener. This information consists of sequences of formations, or "words," which in turn can be abstractly represented as strings of phonemes. In contrast to its written form, the acoustic realization of a word is not a concatenation of basic waveforms corresponding to the string of phonemes in its presentation. Rather, it is a complex encoding in which the phonemic information is interwoven in time. Also, the

acoustic manifestation of a given phoneme depends on its phonemic environment. Consequently, it would seem that a successful scheme for the automatic recognition of words would have to take these two basic facts into account. In particular, such a scheme could neither be based on a segmentation of the acoustic waveform into portions corresponding to phonemes (or other basic elements), nor on measurements that are independent of the phonemic environment.[1,2]

2. Recognition Problem

To see how these facts can be practically incorporated into a speech-recognition scheme, a computer program designed to recognize words is being developed for the PDP-9 computer facility of the Speech Communication group, which has been described in a previous report.[3] The vocabulary list for this initial study consists of 80 single-syllable words as shown in Table XXIV-1. These words are of the form

$$\left\{\begin{matrix}\text{p}\\ \text{b}\end{matrix}\right\}\left\{\begin{matrix}\ell\\ \text{r}\\ \Phi\end{matrix}\right\}\left\{\begin{matrix}\text{vowel:}\\ [\text{i}],\ [\text{e}],\ [\text{u}],\ [\text{o}],\ [\text{a}]\end{matrix}\right\}\left\{\begin{matrix}\ell\\ \text{r}\\ \text{m}\\ \text{n}\\ \Phi\end{matrix}\right\}\left\{\begin{matrix}\text{d}\\ \Phi\end{matrix}\right\}, \tag{1}$$

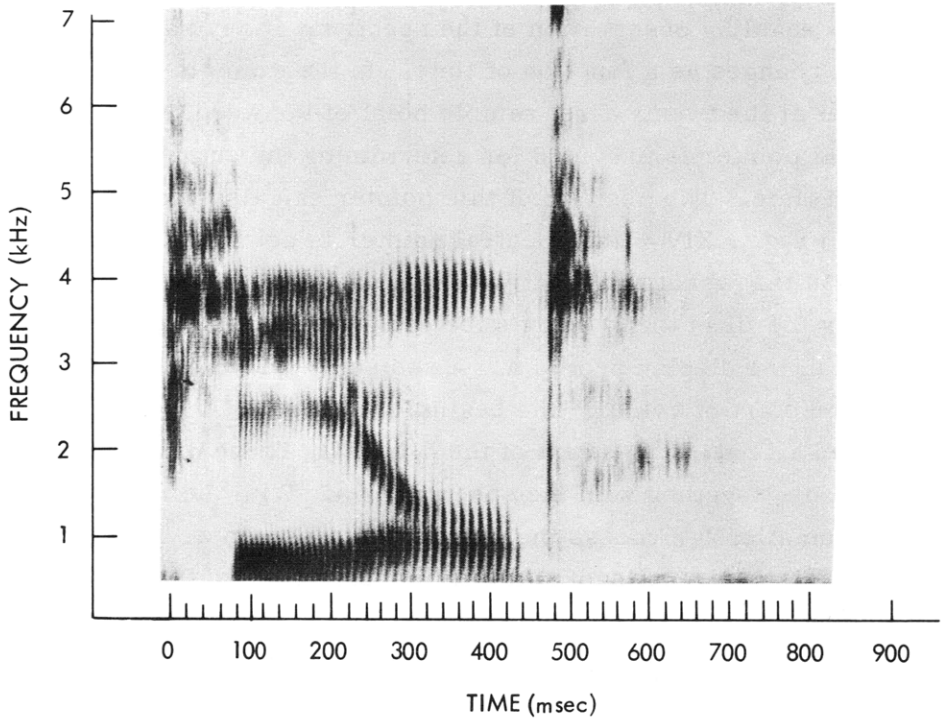
where Φ denotes the absence of a phoneme from a particular group. The list was chosen to focus on the problem of detecting liquids and nasals, since it seems to be important for these sounds to take into consideration the two facts just mentioned. The decision to use only single-syllable words was made to eliminate the need of finding stressed vowels and to provide words of nearly uniform duration.

Three male speakers recorded the word list in an anechoic chamber, and approximations to the short-time spectra were obtained by passing the analog recording through a 36-channel filter bank (see Flanagan[4]). The filter outputs were then sampled at 10-msec intervals and quantized on a logarithmic scale from 0 to 63 dB in increments of 1 dB. These digitized approximations to the short-time spectra, beginning with the sample preceding the release of the initial stops were stored on magnetic tape and provided the data base that was used in the recognition study. By means of the filing system under which the spectra were recorded, the data for any desired word spoken by any of the three speakers could be selected and read into the computer memory.

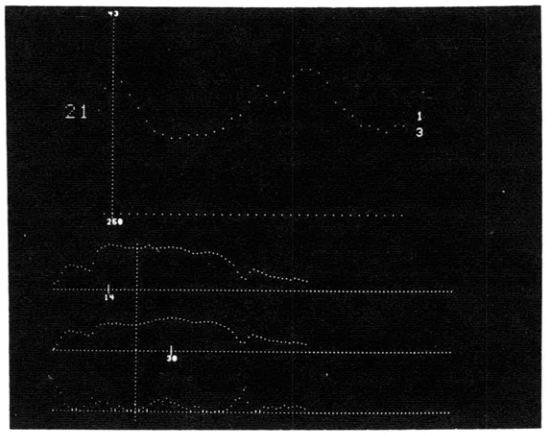
In order to examine the spectral data for the word list, a display program for a computer-controlled oscilloscope was developed. The program displays various time functions derived from the spectra, together with a graph of the filter outputs at a particular point in time. The computer display for one of the utterances is shown in Fig. XXIV-5. For comparison, the spectrogram of the word is shown also. The point in time at which the selected spectrum is observed is indicated by the frame number to the left of the spectral graph and the corresponding vertical line intersecting the time

Table XXIV-1. Single-syllable word list used in the initial recognition study.

i		u		a		e		o	
pea	bee	pooh	booh	pa	bah	pay	bay	Poe	bow
pead	bead	poohed	boohed	pod	bahd	paid	bade	pode	bode
peel		pool		pol		pale		pole	
peeled		pooled		pold		paled		poled	
plee	blead	plooh	blew	plod	blod	play	blade	ploe	blow
peer		poor		par		pair		pour	
peered		poored		parred		paired		poured	
preed	breed	prude	brew	prod	Brahm	pray	bray	pro	broe
	beam		boom		bomb		bame		bome
	beamed		boomed		bombed		bamed		bomed
	bean		boon		bon		bain		bone
	beaned		booned		bond		bained		boned



(a)



(b)

Fig. XXIV-5. (a) Spectrogram of "peeled" as spoken by K. N. S.
(b) Computer display of "peeled" as spoken by K.N.S. On the right of the spectral graph are identification numbers for speaker and word.

functions. This point can be varied by means of a potentiometer over the duration of the word, thereby enabling observation of the spectrum at any sample time and seeing how the spectrum changes as a function of time. In the example of Fig. XXIV-5, the time pointer is set at the twenty-first sample point of the word. In addition to the time pointer, a spectral pointer is provided for determining the output of a particular filter at a given point in time. The position of this pointer can also be varied by adjusting a potentiometer. In Fig. XXIV-5 the spectral pointer is set at the frequency of one of the spectral peaks (250 Hz) where the relative amplitude is 43 dB.

Before storing the filter-bank outputs for each utterance on magnetic tape, the data were examined with the display program. A subjective judgment was made of the positions in time of the start of voicing, the beginning and end of the vowel nucleus, and, if the word contained a final d, the start of the d-burst. These parameters were stored on the magnetic tape, together with the spectral data. With the recognition program it is possible to automatically process the 240 words in sequence, listing on the teletypewriter both decisions and results of measurements made by the program. The four subjectively determined time reference points can then be used in the automatic processing mode to check appropriate results of the recognition procedure and, for example, to type out results only for those words that contain errors, or to transfer control to the display program for such words, so that the causes of the errors can be determined by examining the spectral data and the derived time functions.

Some time functions computed from the digitized spectra are available for use in the recognition procedure. One of these is a sum of the outputs of an arbitrary set of filters. Another function (which is similar to one suggested by Klatt and Bobrow[5]) is calculated from the formula

$$\mathrm{LOUD}(t) = \sum_{i=2}^{6} f_i(t) + 2\max[f_2(t)+f_3(t)-f_4(t)-f_5(t), 0], \tag{2}$$

where $f_i(t)$ is the relative output in dB of the i^{th} filter at time t (i = 1-36). This function is designed to be greatest in the vowel portion of a word and includes a correction term to compensate for inherently weak vowels, that is, those with low first formants. A third function, a normalized spectral derivative, is used to indicate where in time spectral changes occur. It is given by

$$\mathrm{SPDR}(t) = \sum_{i=a}^{b} \left|[f_i(t)-\mathrm{DC}(t)]-[f_i(t-1)-\mathrm{DC}(t-1)]\right|, \tag{3a}$$

where

$$DC(t) = \frac{1}{b - a} \sum_{i=a}^{b} f_i(t). \tag{3b}$$

The term in (3b) is the normalization factor and helps to ensure that the spectral derivative is not large when the amplitude of the spectrum is changing, but the over-all spectral shape is constant. By varying the parameters a and b, a spectral derivative over any desired frequency region can be computed.

In Fig. XXIV-5 are shown examples of these functions for the word "peeled." The broad peaks in the spectral derivative correspond to the start of the p, the beginning of voicing, the transition to the l, and the end of the vowel and start of the d. The effect of the compensation term in the loudness function can be seen by comparing it with the sum of the outputs of filters 2 through 6. As indicated in Fig. XXIV-5 by the markers under these two waveforms, the filter sum attains its maximum during the l-transition, whereas the loudness function is greatest during the vowel.

3. Recognition Procedure

The first step in the recognition process is to determine voicing onset. A rough indication of voicing is given by the presence of sufficient low-frequency energy. By observing at what point in time the sum of the first two filter outputs crosses a threshold, a reasonable estimate of voicing onset can be obtained. This estimate is then used to decide whether the word begins with a p or a b, with small values indicating b and larger values p.

Next, an attempt is made to identify the vowel according to the Chomsky-Halle[6] system of distinctive features by examining the spectrum at the peak of a modified version of the loudness function. The modified loudness function is derived from the loudness waveform by setting it to zero whenever the spectral derivative is sufficiently large. This ensures that the spectral frame that is chosen for making the vowel identification is in a region where the spectrum is not changing rapidly with time. The spectrum at that point is first classified as being characteristic of either a front or a back vowel by looking for high- and low-frequency peaks with a sufficient minimum between them at ~1 KHz. If such a pattern is found, the vowel is classified as front; otherwise, it is back. The vowel is then characterized as either high, mid, or low by computing the center of mass of the spectral points around the first relative maximum and using the following decision function:

$$\begin{aligned} &\text{center of mass} \leq \alpha_1 && \text{high} \\ \alpha_1 < &\text{center of mass} \leq \alpha_2 && \text{mid} \\ \alpha_2 < &\text{center of mass} && \text{low,} \end{aligned} \tag{4}$$

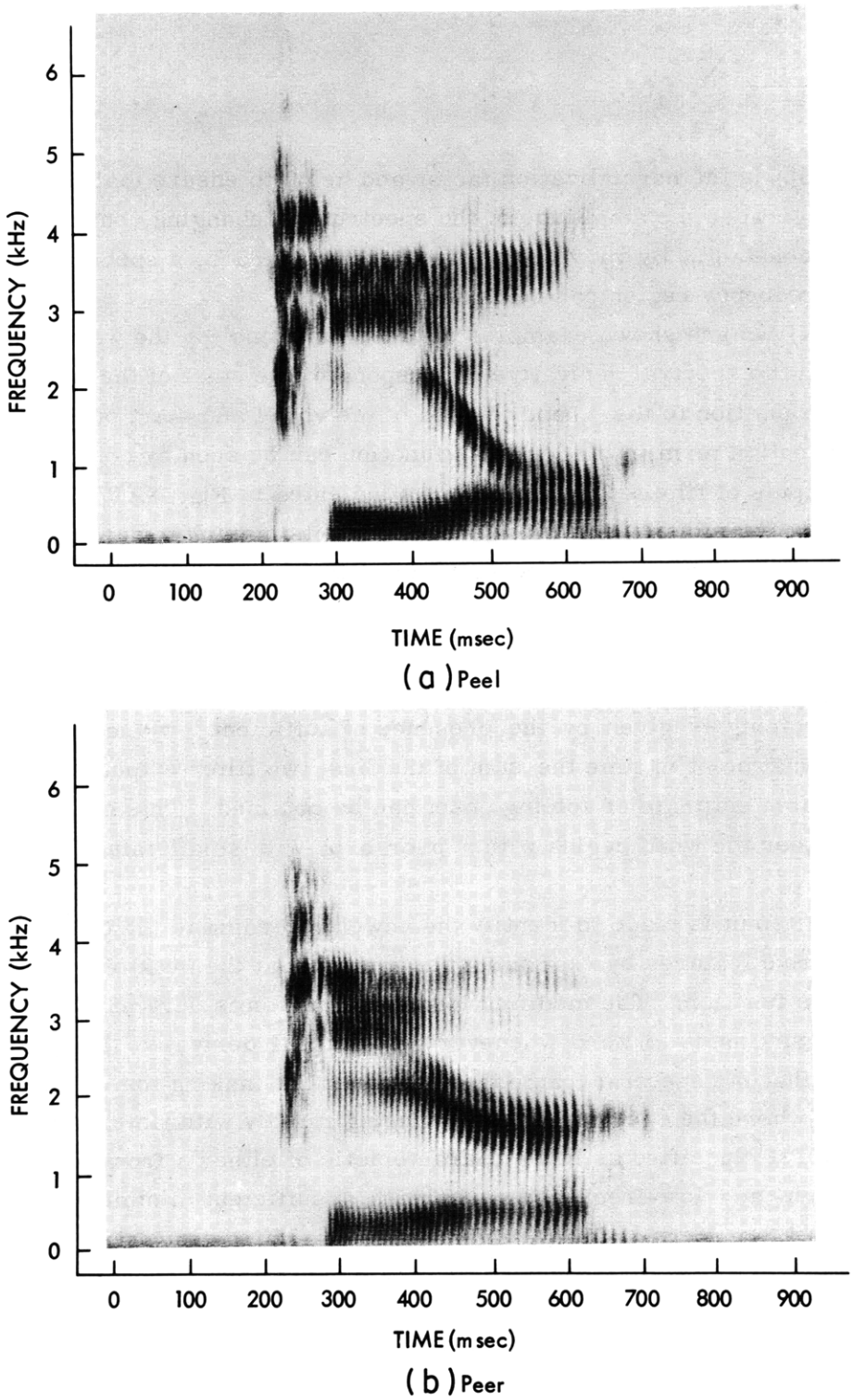


Fig. XXIV-6. Spectrograms of some words with final l's and r's as spoken by K. N. S.

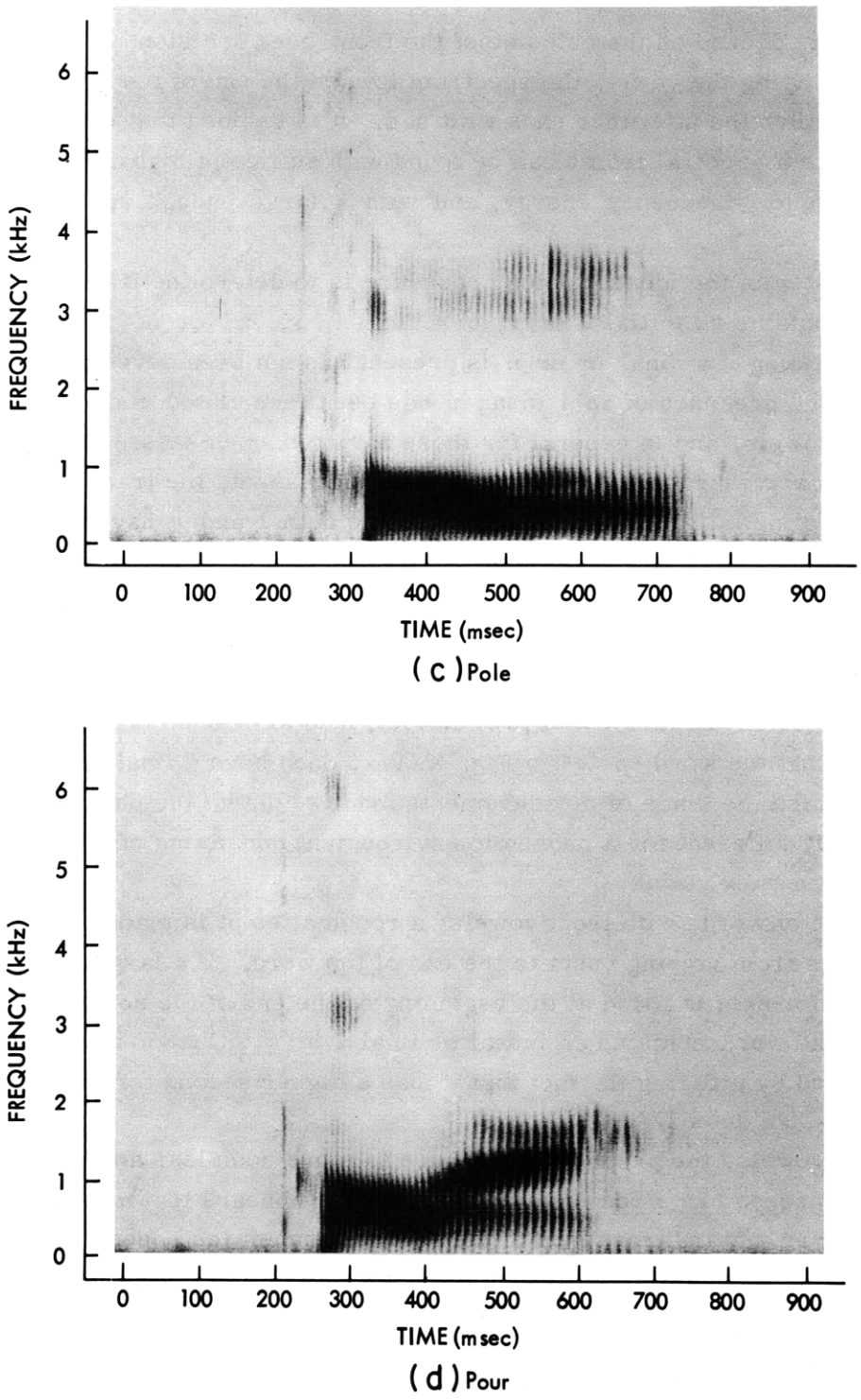


Fig. XXIV-6. Spectrograms of some words with final l's and r's as spoken by K. N. S.

where a_1 and a_2 depend on the outcome of the front-back decision.

After identifying the vowel, the spectrum toward the end of the word is examined to determine whether the utterance ends with a d. It is decided that the word does contain a final d if a spectral frame can be found with sufficient high-frequency energy as compared with low-frequency energy, and with a large enough value for its spectral derivative.

The final step in the identification procedure is to determine if the sonorant portion of the word contains an initial l or r, or a final l, r, m, or n. At present, the procedure for deciding if a final m or n is present has not been developed, but the method for detecting the presence of an l or an r has been formulated in detail for words containing front vowels, and in general for those having back vowels. For words with front vowels, which are characterized by a high second formant, the transition between the vowel and the glide is fairly easy to detect, since both l and r have a low second formant in the steady state. In the case of back vowels, which, like l and r, also have a low second formant, the transitions are much more difficult to find. This is shown in Fig. XXIV-6 by spectrograms of some words containing final l's and r's. For the front vowel [i] in this figure, the transitions are very clear, whereas they are much less dramatic in the case of the back vowel [o]. In fact, it is almost impossible to tell from its spectrogram that the word "pole" in Fig. XXIV-6 does have a final l. Here, then, is a situation in which the kinds of measurements made to detect the phonemes l and r would have to be very different for a phonemic environment consisting of a front vowel than for one containing a back vowel.

In the case of words with front vowels, a rough attempt is made to track the first three formants from voicing onset to the end of the word. If a large enough transition of the second formant is found at the beginning or the end of the sonorant portion, it is decided that the word contains an initial or final l or r. One of the two possibilities is then selected by utilizing the fact that r has a higher second formant and much lower third formant than does l.

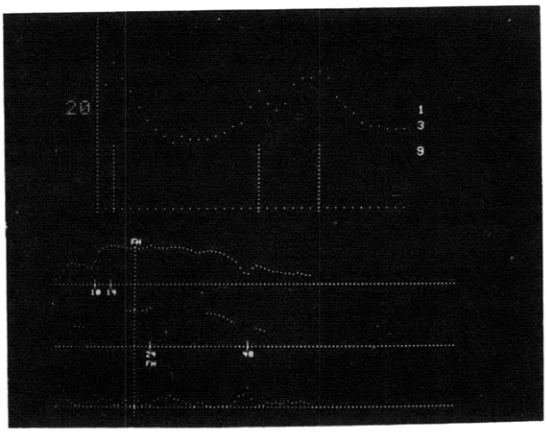
For back vowels, the proposed procedure is more complex, and in all likelihood, will be less successful. From studying the data, it appears that a final r can be found by looking for an upward transition of the second formant (as opposed to a downward transition in the case of an r following a front vowel). This transition is not found for an initial r, but the third formant transition may be able to be reliably detected in that case. Another cue for an initial r may be found by examining the energy concentration in the aspiration of the p, if that is the way the word begins. This aspiration energy is concentrated at a slightly higher frequency if the word contains an initial r than if the sonorant portion begins immediately with the vowel. The same is true for an initial l, with the energy being concentrated at an even higher frequency. This may be the only cue for an initial l, since its third formant is at approximately the same

frequency as that of the vowel. For an initial l following a b, or for an l in final position, it is not clear that any reliable cues can be found in the filter-bank outputs. The vowel [a] may be an exception, however, because its first and second formants are somewhat higher than those of an l, and, therefore, it may be possible to find their transitions.

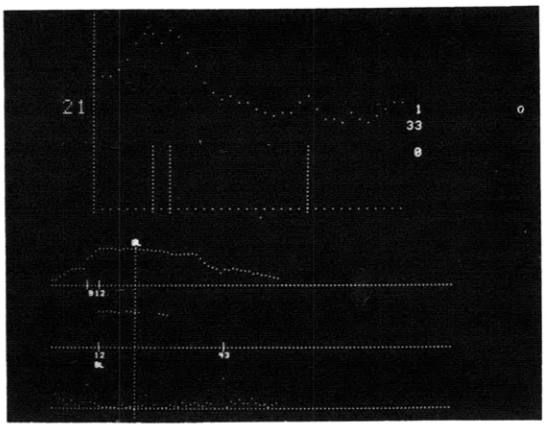
Figure XXIV-7 shows three examples of how the results of some of these decisions are displayed on the oscilloscope. As in Fig. XXIV-5, the top half of each display is a graph of the spectral frame corresponding to the current position of the time pointer. Below it are plotted the loudness function, the modified loudness function, and a spectral derivative computed over the first 20 filters. The first marker under the loudness waveform denotes the program's determination of voicing onset, and the second marker corresponds to the peak of the loudness function. Under the modified loudness waveform, the first marker denotes its maximum, and the second, if present, marks the program's determination of the beginning of the d-burst. The outcome of the vowel identification made at the peak of the modified loudness function is shown under the appropriate marker for that waveform. Moreover, the vowel identification results for the spectral frame corresponding to the current position of the time pointer are displayed just above that pointer, thereby enabling one to see how the vowel identification procedure works at frames other than the one picked by the program. In the spectral graph, the vertical lines indicate the positions of the first three formants, as determined by the formant tracking algorithm. Finally, the numeral below the two identification numbers to the right of the spectral graph gives the outcome of the l-r decision, according to the following convention: 0 denotes neither l nor r, 1 and 2 correspond to initial l and r, respectively, and 9 and 10 indicate final l and r. For the word "peeled," the first example in Fig. XXIV-7, the display indicates that voicing onset occurs at frame number 10, that by examining frame 24 the vowel is found to be front and high ([i]), a d-burst begins at frame 48, and the word contains a final l (compare with the spectrogram in Fig. XXIV-5). For a value of voicing onset equal to 10, it is decided that the word begins with a p, and, therefore, the program correctly identifies this word. The other two examples in Fig. XXIV-7 are interpreted in the same way. The third example, a display for the word "bray," shows how using the peak of the modified loudness function, rather than that of the loudness function itself, yields a frame for vowel identification that is further into the vowel nucleus.

4. Results and Discussion

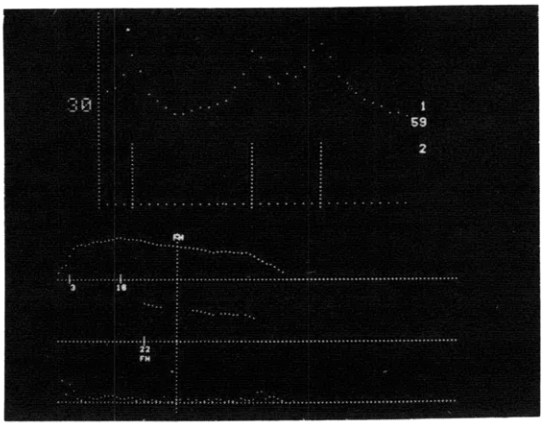
The results obtained so far for the 240 words being tested are presented in Table XXIV-2. These results are fairly encouraging, at the present time, and indicate, perhaps, that the approach that is being developed will be reasonably



(a) Peeled



(b) Pod



(c) Bray

Fig. XXIV-7. Displays of recognition results for three words by speaker K. N. S.

Table XXIV-2. Results of the initial recognition study.

Test	Number of Errors (240 words)
p – b	none
vowel	3
final d	3
r – l: front vowels	none
r – l: back vowels	incomplete
m – n	incomplete

successful. One would want to know, however, how sensitive the routines developed for this set of data are, and it is planned to test the final and complete recognition program with another recording of the entire list for each of the three speakers. Presumably some errors will be made, since the words in this list are in some sense very close to one another. On the other hand, human subjects would for the same reason probably make errors, too. It is hoped that the errors made by the recognition program will be similar to those made by humans and of the same order of magnitude (or less). This is something which will be checked by giving an absolute identification test to a group of subjects and comparing the results with those of the computer program.

5. Future Research

An attempt will be made to extend the procedure described here so that it is capable of recognizing the single-syllable words shown in Table XXIV-3, as spoken by the same three persons used for the initial study. This group of 105 words was selected to contain examples of the twelve vowels and three diphthongs used in English, and all possible initial and final consonants, in addition to many initial and final consonant clusters. After the recognition program for the expanded list has been completed, it will be tested on another recording of the list for each of these three speakers, along with recordings made by seven other men.

In developing the recognition program for the new word list, no attempt will be made initially to utilize the structure of the lexicon. Rather, each word will be examined to determine its vowel and its initial and final consonant clusters, similar to the way in which the words of the initial study are recognized. If, however, the identification does not match one of the lexical entries, it can then be changed to the word of the list which it is most likely that it should be. With such an approach, items from any arbitrary list of single-syllable English words can be chosen as the inputs to the recognition program, the only modifications required being those made in the final correction stage at which point referral is made to the lexicon. Presumably in constructing a list of words

Table XXIV-3. Single-syllable word list to be used in the expanded study.

a	ʌ	ɔ	o	ʊ
palm	dusk	false	though	crook
top	cuff	chalk	soaks	stood
cocks	young	quart	drove	full
flocked	trudge	spawn	clothe	bush
hot	shuns	raw	won't	good
czar	much	gauze	old	nook
throb	pluck	small	notes	wolf

u	ɝ	i	I	e
you	urge	lead	sift	may
gloom	thirst	wreaths	build	Jane
chewed	dearth	chief	give	wake
roost	learn	please	this	brave
shoots	twirl	ear	smith	vague
loop	skirt	fee	kiss	ape
juice	nerve	zeal	till	paid

ɛ	æ	ɔI	aI	au
health	slab	soy	sly	sprout
dense	thatch	toyed	grind	lounge
them	draft	joint	scribe	vowed
blend	clash	moist	bike	blouse
fresh	past	Floyd	types	mouth
sketch	knack	coin	fife	prowl
threat	shall	voice	ninth	ounce

to be used with the recognition program, one would want to keep it from containing entries that are acoustically very similar (such as "Poe" and "pole," for example) so that errors made in the recognition stage can be corrected by referring to the lexicon.

M. F. Medress

References

1. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psychol. Rev. 6, 431-461 (1967).
2. K. N. Stevens, and A. S. House, "Speech Perception," from Foundations of Modern Auditory Theory, edited by H. Tobias and M. Schubert (in press).
3. W. F. Henke, "Speech Computer Facility," Quarterly Pregress Report No. 90, Research Laboratory of Electronics, M. I. T., July 15, 1968, pp. 217-219.
4. J. L. Flanagan, "A Speech Analyzer for a Formant-Coding Compression System," Sc. D. Thesis, Department of Electrical Engineering, M. I. T., May 1955, pp. 5-4 to 5-9.
5. D. H. Klatt, and D. G. Bobrow, "A Limited Speech Recognition System," Report No. 1667, Bolt, Beranek and Newman, Cambridge, Massachusetts, May 1968.
6. N. A. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row, New York, 1968), Chap. 7.

C. ENGLISH AS A TONE LANGUAGE: THE ACOUSTICS OF PRIMARY STRESS

Although stress, as a linguistic feature, has been fairly extensively studied, there has been relatively little work done to determine the nature of its acoustic correlates in the speech signal. We shall show that primary stress in English is correlated with large drops in the fundamental frequency of the signal.

Narrow-band spectrograms were made of a list of English phrases, as recited by three speakers. The fifth harmonic was traced in each case and divided into segments. An example of a spectrogram of one of the utterances is shown in Fig. XXIV-8, and the fundamental-frequency contour traced from this utterance is displayed in Fig. XXIV-9. The frequency at the midpoint of each stressed vowel was measured and the differences in these frequencies for successive stressed vowels was recorded. Each of Tables XXIV-4 through XXIV-6 gives the results for one speaker.

For each speaker, there appears to be a reasonably well-defined cutoff point, such that the frequency difference in the transition from a primary-stressed syllable to a tertiary-stressed syllable (English has no 1 → 2 transition) is almost always greater than the cutoff, while all other transitions display frequency drops smaller than the cutoff. We may conclude from these results that primary stress in English is heard when there is a large drop in fundamental frequency on the following stressed syllable. In those cases in which primary stress appears on the last stressed syllable (stress

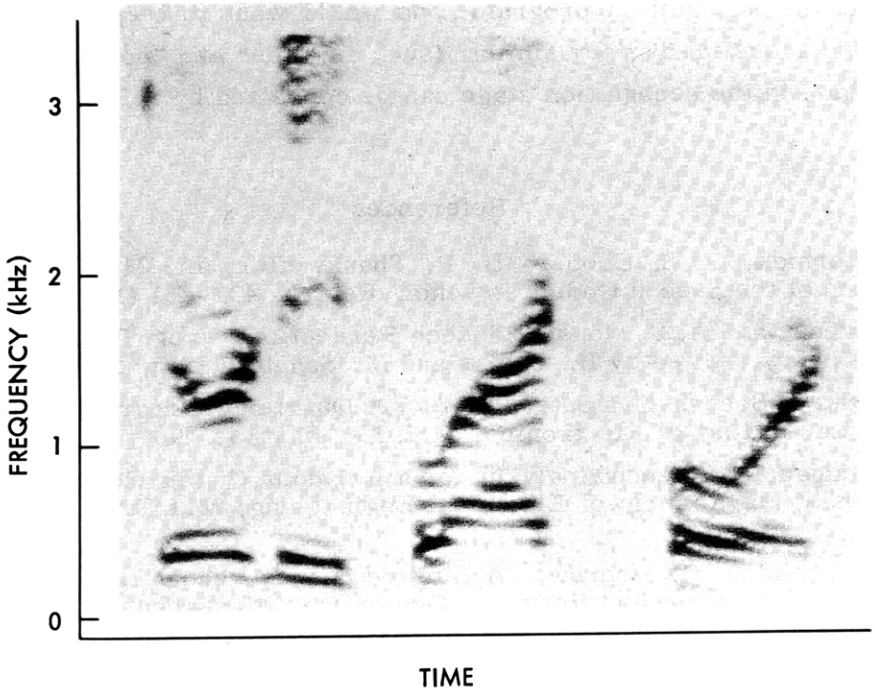


Fig. XXIV-8. Narrow-band spectrogram of a typical phrase used in this study. The phrase is "dirty blackboard" produced by Speaker 1. The stress assignment of the vowels for this utterance is 2 1 3.

patterns 231 and 31), we observe that all frequency differences are small, that is, below the cutoff. Apparently, primary stress is heard on the last syllable of a phrase when the expected large pitch drop has not yet occurred, so that the drop from the last stressed syllable to null is heard as the largest difference and thus taken as the signal for primary stress.

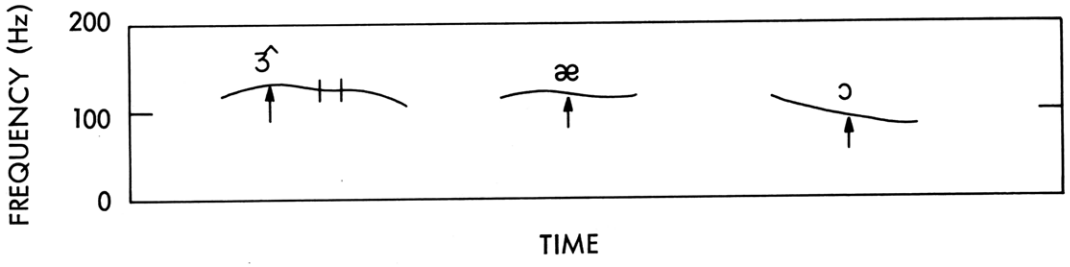


Fig. XXIV-9. Contour obtained by tracing the fifth harmonic during voiced portions of utterance in Fig. XXIV-8. The frequency scale is adjusted to read absolute fundamental frequency. Arrows indicate times in the vowels at which fundamental-frequency measurements were made.

Table XXIV-4. Fundamental frequency drop (in Hz) associated with transition from one stressed syllable to the next for the phrases listed at the left. The stresses assigned to the two adjacent vowels are indicated at the top of each column. Data are for Speaker 1. If cutoff is set at 16 Hz for this speaker, then frequency shifts in the left column are greater than values in all other columns.

Phrases	Frequency drop for each transition				
	1 → 3	2 → 1	2 → 3	3 → 1	3 → 2
2 1 3					
dirty blackboard	27	8			
broad sidewalk	26	7			
heavy tabletop	21	2			
ancient motorcar	43	1			
tall school teacher	39	4			
Boston Red Sox	58	1			
radio telegraphy	35	14			
British Columbia	44	5			
antisemitism	40	-21			
2 3 1					
Bowling Green Avenue			9	7	
Telegraph Hill			13	16	
1 3 2					
sailboat builder	41				7
high school teacher	33				-1
air raid warden	26				8
churchyard fence	46				4
blackboard brush	23				14
Red Sox game	24				10
3 2 1					
Columbia Point Housing Project		1			10
Royal College of Surgeons		2			9
Boston Symphony Orchestra		-1			8
3 1					
Minnesota				-2	
metamorphosis				12	
Ticonderoga				13	
1 3					
boatyard	36				
photograph					
generate	31				
telegraph	49				
kaleidoscope	34				
motorcar	45				
car barn	49				
hurricane	40				
avenue	34				
anticipate	51				

Table XXIV-5. Same as Table XXIV-4, except that data are for Speaker 2, and cutoff is set at 14 Hz. Asterisks indicate exceptions, that is, less than 14 Hz in the left column, or greater than 14 Hz in other columns.

Phrases	Frequency drop for each transition				
	1 → 3	2 → 1	2 → 3	3 → 1	3 → 2
2 1 3					
dirty blackboard	12*	10			
broad sidewalk	25	4			
heavy tabletop	26	8			
ancient motorcar	13*	3			
tall school teacher	31	-7			
Boston Red Sox	22	-3			
radio telegraphy	15	7			
British Columbia	16	8			
antisemitism	14	0			
2 3 1					
Bowling Green Avenue			21*	16*	
Telegraph Hill			4	-2	
1 3 2					
sailboat builder	19				1
high school teacher	33				1
air raid warden	15				13
churchyard fence	39				10
blackboard brush	24				2
Red Sox game	24				0
3 2 1					
Columbia Point Housing Project		0			5
Royal College of Surgeons		4			13
Boston Symphony Orchestra		1			11
3 1					
Minnesota				1	
metamorphosis				8	
Ticonderoga				6	
1 3					
boatyard	25				
photograph	40				
generate	26				
telegraph	27				
kaleidoscope	21				
motorcar	19				
car barn	27				
hurricane	30				
avenue	32				
anticipate	36				

Table XXIV-6. Same as Table XXIV-5, except that data are for Speaker 3, and cutoff is set at 7 Hz.

Phrases	Frequency drop for each transition				
	1 → 3	2 → 1	2 → 3	3 → 1	3 → 2
2 1 3					
dirty blackboard	7	1			
broad sidewalk	6*	6			
heavy tabletop	8	14*			
ancient motorcar	19	-1			
tall school teacher	8	-1			
Boston Red Sox	21	-4			
radio telegraphy	29	0			
British Columbia	22	7			
antisemitism	17	-10			
2 3 1					
Bowling Green Avenue			7	4	
Telegraph Hill			4	5	
1 3 2					
sailboat builder	20				-3
high school teacher	2*				6
air raid warden	12				4
churchyard fence	19				4
blackboard brush	19				6
Red Sox game	7				7
3 2 1					
Columbia Point Housing Project		-1			3
Royal College of Surgeons		-4			5
Boston Symphony Orchestra		13*			-7
3 1					
Minnesota				-1	
metamorphosis				4	
Ticonderoga				-1	
1 3					
boatyard	16				
photograph	9				
generate	19				
telegraph	10				
kaleidoscope	11				
motorcar	11				
car barn	9				
hurricane	15				
avenue	22				
anticipate	19				

Our conclusion that primary stress in English is heard on a syllable that is followed by a large frequency drop may seem, at first, to conflict with the description of stress given by Lieberman.[1] Lieberman found that the frequency of a stressed syllable itself was larger than it would be if the same word were said without stressing that syllable. Thus according to his findings, a model for the speaker would have the speaker raise the frequency of the syllable he wanted to stress, instead of leaving it alone and lowering the frequency of the following syllable, as our results would imply. A careful look at Lieberman's examples reveals, however, that he was investigating emphatic stress, rather than normal linguistic stress, which is what our work is concerned with. For example, Lieberman compared the frequency contours of such sentences as "Joe ate his soup," "Joe ate his soup," and "Joe ate his soup," while we would be interested only in the place of primary stress in the first sentence. There is thus no incompatibility between the two sets of results. Primary and emphatic stress are both signaled by large differences in the frequency of the sound signal. The frequency contrast needed to trigger primary linguistic stress is produced by passively allowing the frequency to fall after uttering the syllable on which stress is desired, while the contrast that triggers emphatic stress is produced by actively raising the frequency of the syllable to be stressed. This result seems quite in accord with the fact that linguistic stress is placed mechanically on a certain syllable of a word, while emphatic stress can occur on any syllable and must thus be actively placed there by the speaker.

A small number of our measurements are exceptional; that is, they fall on the wrong side of the cutoff frequency. These are starred in Tables XXIV-4 through XXIV-6. Such exceptions are caused presumably by idiosyncratic features of the pronunciation of individual speakers. But they suggest something quite fundamental in the nature of stress and of linguistic features in general. If we assume that the speaker has internalized a grammar that contains rules of stress assignment (such as those presented in Chomsky and Halle[2]), then there is no reason to expect stress to be invariably accompanied, in all of its occurrences, by its usual acoustic correlate. As long as there are sufficiently many nonexceptional cases for the hearer not to have to re-evaluate the possible grammars for the primary linguistic data that he receives, he will still hear primary stress in the exceptional cases, since it is assigned to them by the rules that are supported by the much more numerous nonexceptional cases. The total number of exceptional measurements for our three speakers is only 5.6% of all our data, a figure which, presumably, is too low to motivate a hearer to change his rules. Similar reasoning can be applied in order to try to understand the wide range of values which occurs, both above and below the cutoff points.

One of the goals of research in speech production and perception is the development of mechanical devices for the synthesis and recognition of speech. The results presented here should have important applications in the mechanical synthesis of speech.

If our discussion of exceptions is accurate, however, they may constitute a major obstacle to work in automatic speech recognition, since the perception of at least one important linguistic feature would seem to depend not only on the acoustic cues in the speech signal, but also on a system of rules that has been internalized in the speaker's mind. Strong evidence supporting our account of exceptions is provided by an experiment of Selezneva (as reported by Slobin[3]). A Russian-speaking subject is asked to compare a string of four equally stressed sounds with a similar sequence in which the second sound is lower in "intensity"[4] than its corresponding sound in the first string. If the four sounds are arbitrary and do not constitute a Russian word, then the first sound of the second string is heard as stressed. If the four sounds are the syllables of a Russian word, the subject hears the second string as he would normally hear the word, that is, with stress on whatever syllable normally takes it, rather than on the first syllable, as one might expect from analogy with the nonword case. Stress thus seems to be heard where one's rules place it, rather than as a direct consequence of acoustic cues. Once the rules are learned, such cues serve to supplement them, telling the hearer whether they are correct or should be changed. An interesting sequel to Selezneva's and our work would be to devise an experiment to determine what proportion of heard utterances must be exceptional in order for a hearer to be motivated to change his rules of stress placement.

Our work can be extended in several interesting directions, one of these being the determination of the acoustic correlates of secondary, tertiary, and weaker stresses, since there appear to be no regularities among these features in our frequency measurements. It may be that stresses weaker than primary are signaled by completely different acoustic properties, such as differences in time duration. Similar work should also be done on the acoustics of stress in languages with stress patterns considerably different from those of English, for example, Russian, which distinguishes only between stressed and unstressed syllables, or French, which always assigns primary stress to the last syllable of a word.

Our results have significant implications for both psychology and linguistics. We have been careful to present these results as a correlation between primary stress and differences in fundamental frequency, rather than differences in pitch. It is important to keep separate the purely perceptual quality of pitch and the purely acoustic property of fundamental frequency, which ordinarily plays a major role in determining what pitch is heard.

By keeping this distinction in mind, we can see that our basic result constitutes significant evidence for the hypothesis of Chomsky and others that the language capacity is innate in man. We have seen that primary stress in English is determined by a large drop in fundamental frequency. However, it is clear that a speaker of English, when he hears a phrase such as dirty blackboard or Red Sox game in ordinary conversation, is

not normally aware of any differences in pitch, the "usual" perceptual correlate of frequency, among the syllables of these phrases, despite the large variation in frequency that we have found in them. Whereas, when listening to pure tones, he hears frequency differences as differences in pitch, here he hears the major frequency drops as primary stress instead. Thus the same acoustic property is heard as distinct perceptual properties (or produces distinct perceptual experiences) according to whether it appears in an arbitrary sound or in a piece of language. This suggests that the sounds heard by an English speaker in his ordinary life are sharply divided into two classes. The sounds of the first class are marked, in some vague way, as "linguistic," while all other sounds are not so marked. Large frequency drops in sounds of the first class are heard as primary stress, while similar drops in sounds of the second class are heard as differences in pitch. It is difficult to see how one might account for this phenomenon without assuming some sort of innate structure in the mind that classifies all heard sounds as either "linguistic" or "nonlinguistic." Without such a structure, there is no reason why a child learning English should ever start hearing the same frequency differences in different ways.

The preceding discussion suggests at least two sorts of experiment that might profitably be carried out. First it would be instructive to determine whether or not a native speaker of a "tone" language such as Chinese actually perceives pitch differences in conversation.[5] If not, then the "tones" that are heard by a nonspeaker of that language can be seen simply as a consequence of the fact that he hears the sounds of the language as "nonlinguistic." The very notion of "tone language" would turn out to be an artificial construct, invented by linguists because of their failure to take seriously the perceptual realities of the native speaker/hearer.

The term "tone language," if it were to have any meaning at all, would then have to apply to any language that uses frequency differences to mark linguistic features, and English would be one of these, because of our results on stress. A further experiment could be devised to determine whether a speaker of a language quite different from English[6] would, in fact, hear the frequency drops that an English-speaker hears as primary stress as a "falling tone," since English phrases would be "nonlinguistic" for such a hearer. He might very well conclude that there are phrases with different meanings in English, such as lighthouse keeper and light housekeeper, which are identical in all respects but pitch, since he would hear our stress as pitch, and that English and Chinese are no different in this respect.

I would like to express my appreciation for the guidance and encouragement of Professor Kenneth N. Stevens and Professor Morris Halle, without whom this work would never have been done.

S. Cushing

Footnotes and References

1. P. Lieberman, Intonation, Perception, and Language (The M. I. T. Press, Cambridge, Mass., 1967).
2. See N. A. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row, New York, 1968) for a proposed account of the generative rules of English stress placement; also, see J. W. Jeaffreson, "Stress and Rhythm in Speech," Trans. Philolog. Soc. (London), pp. 73-95, 1938 for a discussion of some articulatory correlates of stress.
3. D. Slobin, "Soviet Psycholinguistics," in N. O'Connor (ed.), Present-Day Russian Psychology (Pergamon Press, Oxford, 1966).
4. It is not clear what Slobin means by the "intensity" of a sound and I have been unable to check Selezneva's own account of the experiment. Because of the common terminological confusion between such acoustic and perceptual features as amplitude and loudness or frequency and pitch, he probably means either that the second sound of the string was reduced in perceived loudness or that the amplitude of its sound wave was decreased. These are not quite the same, as was once erroneously thought, and as is still often assumed. In fact, in view of our findings, one suspects that what actually took place was a lowering of the frequency of the second sound, since the result was the perception of stress on the first sound. The experiment should be repeated with more attention paid to the acoustic properties of the sounds used.
5. M. Pei, The Story of Language (J. B. Lippincott Company, New York, 1965), especially see p. 378. Pei tells of a legend that, for centuries, the Chinese were unaware that they were using tones in their language. Even when this fact was first pointed out by a scholar in 500 A. D., the people were very reluctant to accept it. Whatever the truth of the legend may be, the fact that it exists at all suggests that the use of tones is at least not taken for granted even by those who allegedly use them most.
6. For example, a language that does not use frequency differences systematically to signal linguistic features would be ideal for this purpose, if such a language exists.

D. PERCEPTUAL INTEGRATION OF DICHOTIC CLICK TRAINS

In a recent paper, Axelrod and his co-workers[1] reported that when subjects are asked to select a monaural train of pulses, whose repetition rate matches the total pulse rate of a second train, in which successive pulses go to alternate ears, they select a train whose rate is approximately 60% of the actual total rate, at least when the total rate is between ~7 and 40 pulses per second. At very slow rates such as 1 pulse per second subjects can match the total rate. It is hard to think of any mechanism that could explain why the matching rate should be a constant proportion of the standard rate, over a two-octave range of the standard rate. The inhibitory model suggested by Axelrod as one possible explanation would surely produce effects that varied with rate, unless the inhibition of the response to a click in one ear by a click in the other ear did not depend on the interval between the two clicks, which seems unlikely.

The authors went on to relate their results to earlier experiments on the perception of alternated speech, that is, speech that is switched alternately to the left and right

ears of a listener, so that all of the speech "enters the head," but never through both ears simultaneously. Unfortunately, the pulse rates used by Axelrod were too widely spaced for any to fall in the range of rates at which the perception of alternated speech is most disturbed. It is at least possible, however, that the two effects have a common cause, so a pilot study has been performed to try to replicate Axelrod's results, and to fill in their data in the critical range between approximately 2 pulses per second and 10 pulses per second.

Sixteen alternation pulse rates between 1 and 50 pps were selected, to cover the same range as those used by Axelrod. The rates were logarithmically spaced, and the rates between ~4 pps and 12 pps (corresponding to 2 and 6 complete cycles per second, respectively, where 1 cycle consists of 1 pulse to each ear) were rather finely sampled. The method of adjustment was chosen instead of the method of limits used by Axelrod, and the train of pulses whose rate the subject adjusted was presented binaurally instead of monaurally. It may be necessary to repeat the experiment later with monaural pulses, to make sure that this is not the cause of the failure to replicate their results.

The rates were presented in irregular order, and subjects were run individually. For each rate, the subject first heard the train of alternated pulses at the selected rate, and then could select whichever train he wished to hear by pressing one of two buttons in front of him. The repetition rate of the variable train was controlled by a knob, which had an exponential law. When the subject had adjusted the variable rate until the repetition rate sounded like the total pulse rate of the alternated pulse train, he pressed a third button, and the next alternated train was presented for him to match.

The whole experiment was programmed on our group's PDP-9 computer, which automatically randomized the presentation order, and tabulated and plotted the results at the end of a session. Every rate was matched once in each session. All events were timed by a clock running at 10 kHz, and pulse duration was set at one clock cycle. Pulse amplitude was set to a comfortable listening level, and was approximately matched for the two trains. (Pulse amplitude was approximately 0.35 Volt into 500-Ω Sharpe HA-10 Mk 2 headphones. A sine wave with the same rms voltage would produce ~88 dB SPL.)

Preliminary results for one subject are presented in Fig. XXIV-10. The ratio of the rate of the adjustable train divided by the total rate of the alternated train is plotted as a function of the total rate of the alternated train. Each data point represents the median of 8 matches, each carried out in a different session. For comparison, Axelrod's data are also plotted as crosses. Clearly, the two sets of data are not in agreement. The subject in the present produced a veridical match when the total pulse rate was less than ~8 pps, but switched over very sharply to matching the pulse rate in one ear when the total rate was faster than ~15 pps.

The scatter of the data points about the plotted points was also interesting. There

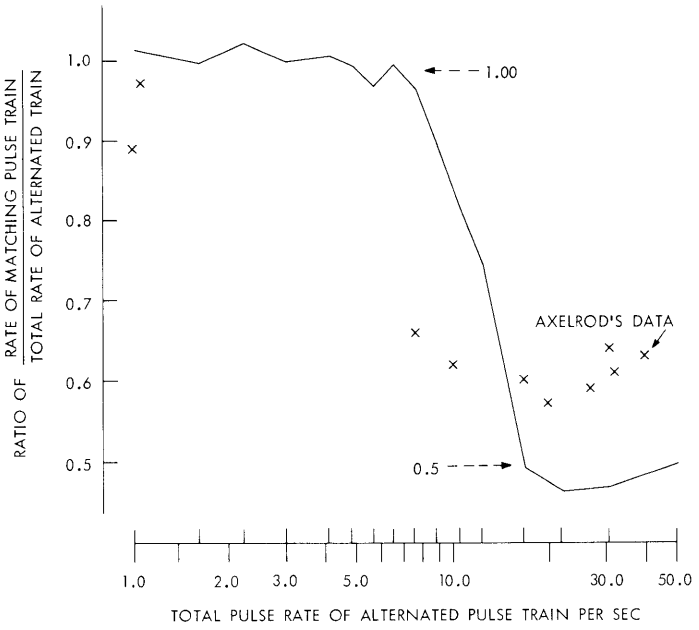


Fig. XXIV-10. Preliminary results.

was very little scatter about the points for the five slowest and four fastest rates, all of the data points falling within a range of approximately plus or minus 3 per cent. But between the rates of 5 pps and 12 pps there was substantially more scatter, and the task was subjectively much harder. These rates correspond very closely to those at which alternated speech is most disturbed (Cherry,[2,3] Schubert and Parker,[4] Huggins[5]).

A possible explanation for both the present result and those on alternated speech is that when successive events in one ear are separated by more than ~250 msec of silence, the successive events form separate percepts. When the separation is less than ~150 msec, successive events in one ear form a single percept, given certain constraints on the similarity of the two events. Thus when pulses are alternated between the ears at slow rates, each pulse forms a separate percept, which the subject can then combine according to his instructions. But when the separation between pulses in one ear drops below ~150 msec, a percept of a pulse train is formed for each ear, and there is no way of combining the two trains to form a third percept of a single train with twice the rate.

It should be noted that the duration of 150-250 msec agrees well with Guttman and Julesz's finding that homogeneity of quality in a repeated sample of random noise persists down to a repetition rate of 4 Hz (Guttman and Julesz). It also ties in with some of Zwislocki's[7] results on temporal integration, for which he showed that "... (the threshold shift for detection of either of a pair of pulses) is mathematically analogous to that of an electronic integrating network with a time constant of 200 msec ...".[8]

The hypothesis stated above, that separation of events in one ear by more than 250 msec leads to separate percepts, is very similar to the explanation put forward by Cherry to explain the loss of intelligibility of alternated speech. Cherry suggested[3] that the critical rate of alternation is a measure of the time taken by the subject to switch attention between his ears, so that at low rates the subject succeeds in stringing together the separate percepts into the correct sequence, whereas above the critical rate, he attends to one ear. Huggins[5] showed that, in fact, the subject uses the interrupted speech in both ears above the critical rate, and that the intelligibility of alternated speech can be described roughly as the logical sum of the intelligibilities of the two interrupted messages, one of which reaches each ear. The results are not compatible with Huggins' other finding, however, that when the playback speed of the speech is changed, the critical alternation rate changes by the same factor. Further experiments will consolidate the results presented in this report, and will explore the foregoing discrepancy.

A. W. F. Huggins

References

1. S. Axelrod, L. T. Guzy, and I. T. Diamond, "Perceived Rate of Monotic and Dichotically Alternating Clicks," J. Acoust. Soc. Am. 43, 51-55 (1968).
2. E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," J. Acoust. Soc. Am. 25, 975 (1953).
3. E. C. Cherry and W. K. Taylor, "Further Experiments on the Recognition of Speech, with One and with Two Ears," J. Acoust. Soc. Am. 26, 554 (1954).
4. E. D. Schubert and C. D. Parker, "Addition to Cherry's Findings on Switching Speech between the Two Ears," J. Acoust. Soc. Am. 27, 792(L) (1955).
5. A. W. F. Huggins, "Distortion of the Temporal Pattern of Speech: Interruption and Alternation," J. Acoust. Soc. Am. 36, 1055 (1964).
6. N. Guttman and B. Julesz, "Lower Limits of Auditory Periodicity Analysis," J. Acoust. Soc. Am. 35, 610(L) (1963).
7. J. Zwislocki, "Theory of Temporal Auditory Summation," J. Acoust. Soc. Am. 32, 1046 (1960).
8. Ibid., p. 1047.