XIII. SPEECH COMMUNICATION

Academic and Research Staff

Prof. Kenneth N. Stevens
Prof. Morris Halle
Dr. Sheila Blumstein[*]
Dr. Margaret Bullowa
Dr. William L. Henke
Dr. A. W. F. Huggins
Dr. Allan R. Kessler
Dr. Dennis H. Klatt
Dr. Martha Laferriere[†]
Dr. Paula Menyuk[‡]
Dr. Colin Painter
Dr. Joseph S. Perkell
Dr. Jaqueline Vaissière[**]
Dr. Katherine Williams
Dr. Catherine D. Wolf[††]
Lise Menn

Graduate Students

Marcia A. Bush
William E. Cooper
Bertrand Delgutte
Gregory M. Doughty
William F. Ganong III
Ursula G. Goldstein
Stephen K. Holford
Shinji Maeda
Victor W. Zue

A. PERCEPTION OF SEGMENT DURATION IN SENTENCE CONTEXTS

National Institutes of Health (Grant 5 RO1 NS04332-12)

Dennis H. Klatt, William E. Cooper

1. Introduction

The temporal organization of speech is based on a complex system of rules involving some linguistic and nonlinguistic variables. The duration of a phonetic segment is influenced by linguistic factors that include aspects of the semantic, syntactic, and phonetic environments of the segment in a spoken utterance. For example, in order to predict the duration of a vowel in American English, one would have to include rules to account for the following observations.[1]

(i) Each vowel type has a different inherent duration.

(ii) Vowel duration depends on features of the postvocalic consonant (a large effect is seen only in phrase-final syllables).

(iii) Unstressed vowels are of shorter duration than stressed vowels.

(iv) Vowels in word-final syllables are slightly longer than in other positions.

(v) Vowels in phrase-final syllables are longer than in other positions within a phrase.

The perceptual relevance of these effects can be estimated by using any of several

[*] Assistant Professor, Department of Linguistics, Brown University.

[†] Assistant Professor of Linguistics, Southeastern Massachusetts University.

[‡] Professor of Special Education, Boston University.

[**] Instructor of Phonetics, Department of French, Wellesley College.

[††] Assistant Professor, Department of Psychology, Wellesley College.

psychophysical techniques. Nooteboom[2] employed synthetic speech and a method of adjustment to show that in Dutch the difference in duration between inherently long and short vowels is a part the listener's internal representation for these segments. Fry[3] has shown that changes in the relative durations of vowels in a two-syllable word can change the perceived stress pattern for the word. The changes in vowel duration induced by features of the following consonant have been shown to carry a functional load in speech perception in the sense that a change in the duration of the vowel will change the perception of the voicing feature of the following consonant.[4-6]

The fourth durational effect involving the position of a syllable within a word has been shown by Nooteboom[2] to influence vowel perception. He determined the preferred durations for vowels as a function of number of syllables and syllable position in a word, using synthetic speech and a method of adjustment. He showed that preferred vowel durations are shorter as a function of the number of syllables left to be produced in the word. The amount of shortening was similar to production data in Dutch, Swedish,[7] and English.[8,1] Nooteboom has also shown[6] that the perceptual boundary along a synthetic continuum between phonemically long and short vowels in Dutch decreases by approximately 15 ms if a second syllable is added to the word. He concludes that a categorical decision concerning vowel duration must be delayed until at least the end of a word.

The fifth durational effect involving the position of the syllable within a phrase has not been related directly to perceptual expectations of listeners, but Lehiste[9] showed that the last metric foot of a spoken sentence must be of longer duration than earlier metric feet or else listeners will perceive the final metric foot as too short.

Fluent pauses often occur at phrase boundaries. O'Malley et al.[10] have shown that the locations of pauses can be used by naive listeners to disambiguate spoken algebraic expressions. Segment lengthening occurs, however, at phrase boundaries whether or not a physical pause is present.[1] It is not known whether this lengthening is used by listeners to aid in the determination of the constituent structure of an utterance, nor is it known whether the lengthening is expected when listening to a fluently spoken sentence. The experiments described in this report were designed in order to examine the last problem.

A set of sentences was constructed in which the morpheme "deal" appears (a) in one- and two-syllable words, (b) as noun, verb, or adjective, and (c) in phrase-final and non phrase-final positions. Subjective estimates of the preferred duration for the vowel /i/ have been obtained from listeners by modifying the duration of /i/ in natural recordings of the sentences.

The duration of English consonants may also be influenced by the locations of word boundaries and phrase boundaries.[8,11] Accordingly, in a second experiment sentences containing the postvocalic fricative /$\check{s}$/ of the morpheme "fish" were used. A fricative was chosen because the duration of stops appears to be less influenced by the locations of word and phrase boundaries.

The experimental design permits calculation of a just noticeable difference (JND) for segmental duration as a function of position within a sentence. The JND data from these experiments are of interest because the experiments involve a randomized set of 6 or 7 sentences, which produces a more natural listening situation than in previous JND studies in which the same segment,[12-14] word,[15, 2] or sentence[16] was repeated over and over again.

2. Stimulus Preparation

The sentences listed in Table XIII-1 were recorded several times by one of the authors. The most natural-sounding recording of each sentence was selected for computer processing. The waveforms were digitized at 10, 000 samples per second, using a high-quality linear-phase 5000 Hz lowpass filter to remove alias components. The digitized waveforms were then digital-to-analog converted, again lowpass filtered, and rerecorded onto audio tape. Broadband spectrograms[17] were made to check for artifacts and to help in the location of good segmentation points. A spectrogram of one of the digitized utterances is shown in the lower half of Fig. XIII-1.

a. Experiment 1: Vowel Duration

A waveform editing program[18] was employed to define the beginning and end of each sentence and to place time markers at selected points in the sampled waveform. For the sentences involving the vowel /i/ of the word "deal", time markers were placed at nearly the same sample in each of 5 selected pitch periods in the approximately steady-state part of the vowel, as determined from the formant pattern seen in the spectrogram, so that any of 4 consecutive waveform segments could be played back zero, one, or two times. A durational increment of two pitch periods was selected for the sentences of the first experiment on the basis of an informal pretest.

Nine stimuli were constructed from each of 7 sentences involving the word "deal". In stimulus number 1, waveform segments numbered 1, 2, 3, and 4 were deleted from the digitized recording, thereby yielding the sentence with the shortest duration vowel. Stimulus number 5 was the original recording, and stimulus number 9 was generated by duplicating waveform segment numbers 1 through 4. Spectrograms of the vowels excised from stimuli constructed from one of the sentences are shown in the upper part of Fig. XIII-1.

The composition and measured duration of each syllable nucleus in "deal" for the sentence "Bill played dealer's choice" is shown in Table XIII-2. The duration measurement for the original recording (stimulus number 5) is taken from the instant of release of the /d/, as indicated by the onset of the plosive burst, to the instant of release of the tongue tip after the /l/, as indicated by the sudden shift of the first and second formants

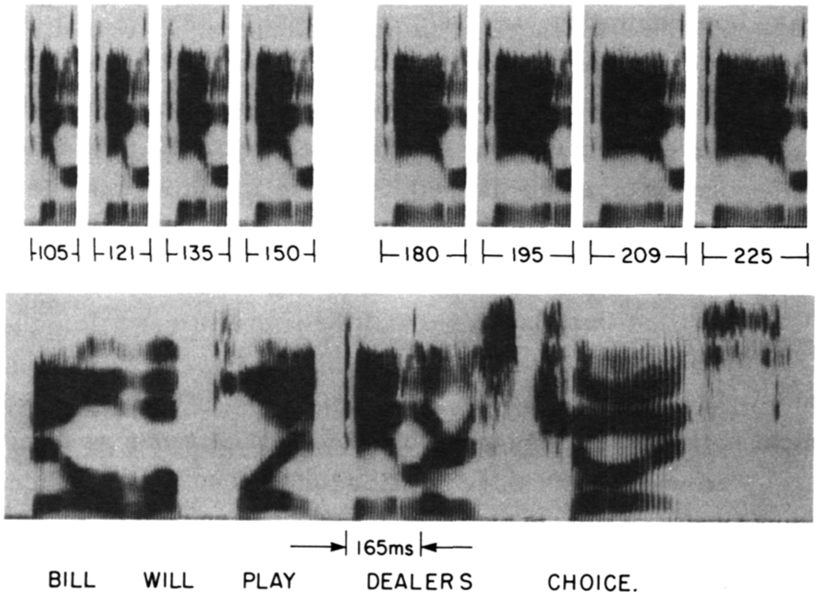


Fig. XIII-1. Broadband spectrogram of the original recording of the sentence "Bill will play dealer's choice." Upper: spectrograms of the word "deal", as excised from 8 other sentences that were constructed from this recording.

Table XIII-1. Sentences used in the experiments involving the words "deal" and "fish".

DURATION OF /il/ (ms)	SENTENCE
165	Bill will play dealer's choice.
195	The deal situation is bad.
210	Ken was the next dealer.
210	Henry will deal the next hand.
220	The deal rotates to the left.
310	He passed up the deal.
340	Deal.

DURATION OF /ɪ/ (ms)	DURATION OF /s/ (ms)	SENTENCE
55	95	Bill played near the fishing hole.
85	105	The fish market is open.
80	110	Ken wants to go fishing.
90	100	Henry will fish until dark.
135	160	The small fish were biting.
200	200	He cleaned the fish.

and the concomitant discontinuous increase in waveform amplitude. This measurement corresponds to the duration of the plosive burst, the /i/ and the postvocalic /l/. Measured durations for /il/ in each of the original "deal" sentences are presented in Table XIII-1. The /i/ duration appears to be about half of this value in each of the sentence contexts.

The fundamental frequency of the vowel in "Bill played dealer' s choice" was approximately 133 Hz. Therefore successive stimuli differ in duration by ~15 ms for this sentence. The durational increment was slightly different in other sentences because of changes in fundamental frequency.

A test tape was prepared in which 63 stimuli were randomized and recorded at 5-second intervals. A few extra stimuli were placed at the beginning and end of the randomized sequence.

b. Experiment 2: Postvocalic Fricative Duration

The six sentences involving /s/ in the word "fish" that are listed in Table XIII-1 were treated in an analogous fashion. Nine stimuli were prepared from each sentence by placing markers at 10, 15, or 20 ms intervals during the /š/ frication noise. The interval size was chosen on the basis of a pretest. Markers were always placed near an up-going zero crossing of the waveform in order to minimize the possibility of generating spurious clicks. Spectrograms of the original recording and the 8 modified versions of /š/ are shown in Fig. XIII-2 for the sentence containing the shortest original /š/ duration. Measured durations for the /š/ and the vowel /ɪ/ of "fish" are presented in Table XIII-1 for each of the sentences. A test tape was then prepared in which

Table XIII-2. Number of times each waveform segment was played to generate nine stimuli for the sentence "Bill will play dealer's choice." Also presented are the durations for the resulting syllable nuclei /il/ in each of the stimuli.

STIMULUS	SEGMENT NUMBER				DURATION OF /il/ (ms)
	1	2	3	4	
1	0	0	0	0	105
2	1	0	0	0	120
3	1	0	0	1	135
4	1	1	0	1	150
5	1	1	1	1	165
6	1	2	1	1	180
7	1	2	2	1	195
8	1	2	2	2	210
9	2	2	2	2	225

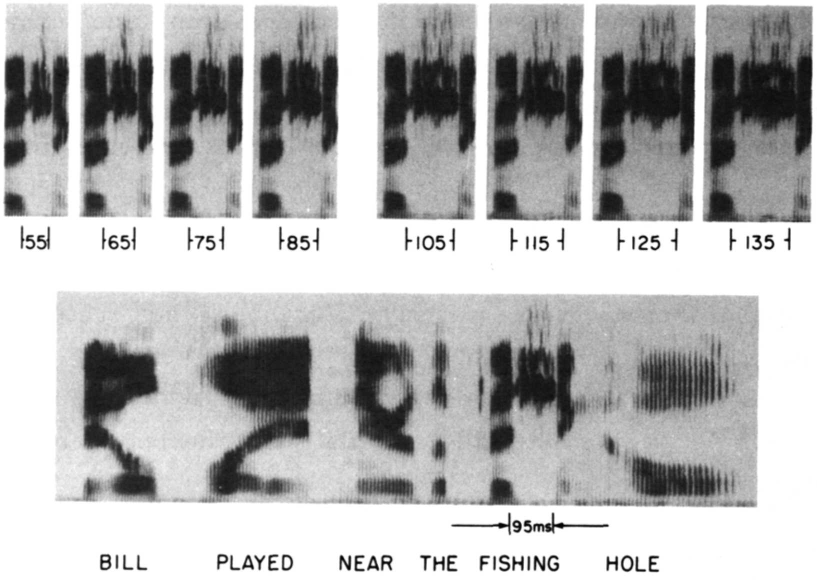


Fig. XIII-2. Broadband spectrogram of the digitized recording of the sentence "Bill played near the fishing hole." Upper: spectrograms of the fricative /š/, as excised from 8 other sentences that were constructed from this recording.

54 randomized sentences were recorded with a few extra stimuli at the beginning and end of the sequence.

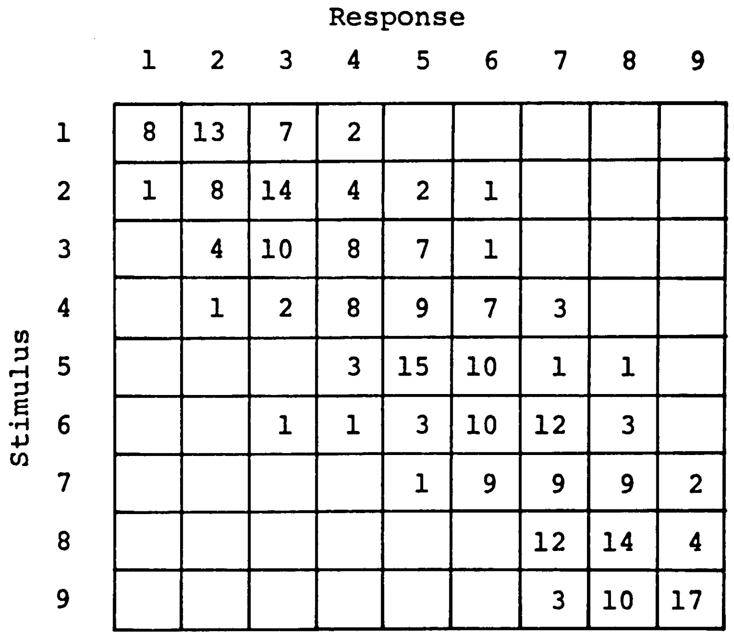
3. Test Procedures and Results

The experimental tapes were played to subjects binaurally at a comfortable listening level; an Ampex PR-10 tape recorder and matched Telephonics headphones were used. Ten subjects listened to each test tape a total of 3 times without feedback. Each tape was heard on a separate day, and one of the subjects participated in both experiments. The task of the subjects was to estimate the magnitude of the duration of the vowel /i/ in the word "deal" (Experiment 1) or the fricative /š/ in "fish" (Experiment 2), using the integers 1-9. The task represents a special case of the magnitude estimation task and is typically referred to as a category judgment task or an absolute identification task.[19] Subjects were played 3 stimuli before the test: the shortest, natural, and longest duration versions for the third sentence in each group, as listed in Table XIII-1, and they were told that these examples represented duration magnitudes of 1, 5, and 9. It was explained that number 5 should be used as a response for sentences in which the vowel had a natural duration relative to the particular sentence context. No more examples were given in order not to bias their judgments concerning what a natural duration might be in other sentence contexts.

Responses were organized in the form of a 9 × 9 confusion matrix for each sentence

Table XIII-3. Category judgment responses of one subject to 9 versions of the sentence "Bill will play dealer's choice" are shown in row (a) and his responses to the sentence "The deal rotates to the left" are shown in row (b).

	STIMULUS NUMBER								
	1	2	3	4	5	6	7	8	9
RESPONSE TO (a)	1	1	3	3	4	5	6	7	8
DURATION (ms)	105	120	135	150	165	180	195	210	225
RESPONSE TO (b)	2	4	4	4	5	3	5	5	6
DURATION (ms)	162	177	191	206	220	234	249	263	278



Stimulus	Response 1	2	3	4	5	6	7	8	9
1	8	13	7	2					
2	1	8	14	4	2	1			
3		4	10	8	7	1			
4		1	2	8	9	7	3		
5				3	15	10	1	1	
6			1	1	3	10	12	3	
7					1	9	9	9	2
8							12	14	4
9							3	10	17

Fig. XIII-3. Confusion matrix of the pooled responses of 10 subjects to 9 stimuli involving the sentence "Ken was the next dealer."

type. An example of a confusion matrix of pooled responses from 10 subjects is shown in Fig. XIII-3. Responses from one of the best subjects are shown in Table XIII-3. It can be seen that this subject is very adept at distinguishing among 9 different durations for the /i/ in "Bill will play dealer's choice," but he is unwilling or unable to distinguish similar durational changes in the sentence "The deal rotates to the left."

a. Original Segment Duration

The data in Table XIII-4, column 1 indicate that the durations of the /il/ of "deal" and the /š/ of "fish" in the original sentence recordings follow a pattern similar to that observed previously. Both segments are of longest duration in sentence-final position. Segments are longer in phrase-final positions. The difference between durations in the verb and in internal phrase-final positions is smaller than expected. The reason may be that each sentence was relatively short and was spoken in isolation; all segment durations in phrase-final syllables were longer in a connected discourse.[1]

The segments are shortest in the first syllable of a polysyllabic word. The effect of syllable position within a word is small in these data, presumably because of the relative incompressibility of vowels that have been shortened by the influences of other shortening factors.[20]

Table XIII-4. Summary of results.

ORIGINAL DURATION of /il/ (ms)	PREFERRED DURATION of /il/ (ms)	PERCEPTUAL MAGNITUDE of a 10 ms CHANGE	JND	SENTENCE
165	173	.40	26	Bill will play dealer's choice.
195	203	.25	51	The deal situation is bad.
210	201	.46	22	Ken was the next dealer.
210	213	.26	59	Henry will deal the next hand.
220	225	.23	50	The deal rotates to the left.
310	268	.29	36	He passed up the deal.
340	309	.33	40	Deal.

ORIGINAL DURATION of /š/ (ms)	PREFERRED DURATION of /š/ (ms)	PERCEPTUAL MAGNITUDE of a 10 ms CHANGE	JND	SENTENCE
95	114	.56	28	Bill played near the fishing hole.
105	111	.46	37	The fish market is open.
110	104	.60	25	Ken wants to go fishing.
100	98	.43	30	Henry will fish until dark.
160	98	.19	67	The small fish were biting.
200	142	.21	98	He cleaned the fish.

b. Preferred Duration

The pooled data were analyzed to determine the preferred (i. e., most natural sounding) duration for the syllable nucleus or fricative in each syntactic environment. The average category judgment was plotted as a function of stimulus number for each sentence type. Examples of several plots are shown in Fig. XIII-4. A straight line was fitted to the data points for each experiment, with the best least-squares fit used as the criterion for placement of the line. The stimulus duration corresponding to the place where this line crosses a response of 5 indicates the preferred stimulus duration. Preferred durations for each of the sentences are presented in Table XIII-4, column 2.

Agreement between the segment duration in the original sentence and the preferred duration is good over a duration range of almost an octave in the vowel data of Experiment 1. A rank-order comparison of the original and preferred durations produces a very high positive correlation with no reversals in rank order. There is a slight tendency for preferred durations to be more restricted in total range of variation with sentence context.

The fricative data of experiment 2 are similar, except that the /š/ of "The small fish were biting" was judged longer than normal for every stimulus, including stimulus 1 where 60 ms had been excised from the /š/. The average response is plotted in Fig. XIII-5 as a function of stimulus number for this sentence, and the best-fit straight line is also shown. A possible explanation for this result will be discussed later.

c. Comparative Perceptual Magnitude of a 10-ms Change in Duration

The slopes of the curves shown in Figs. XIII-4 and XIII-5 indicate the average change in response for a given change in segment duration. In order to make comparisons across all experimental conditions, the slope of each best-fit straight line has been converted into the change in response magnitude corresponding to a 10-ms change in segment duration. The results are presented in Table XIII-4, column 3. It can be seen that a 10-ms change induces the greatest subjective change in duration if the word contains a following second syllable. Generally, a 10-ms change in the /š/ is perceived to be slightly greater than a corresponding change in /i/, except in sentence-final position where a 10-ms change in /š/ is judged to be very small.

d. Just-Noticeable Difference in Segment Duration in Sentence Contexts

These perceptual magnitude values are related to the just-noticeable difference (JND) in segment duration. However, in order to compute a meaningful JND, it is necessary to know the variance in the category judgment data, which depends somewhat on the sentence

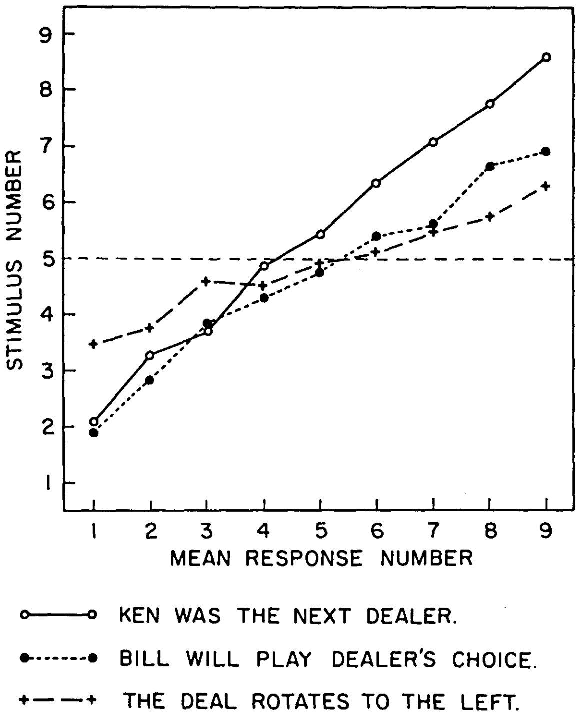


Fig. XIII-4. Mean category judgment response as a function of stimulus number for 3 sentences of the first experiment.

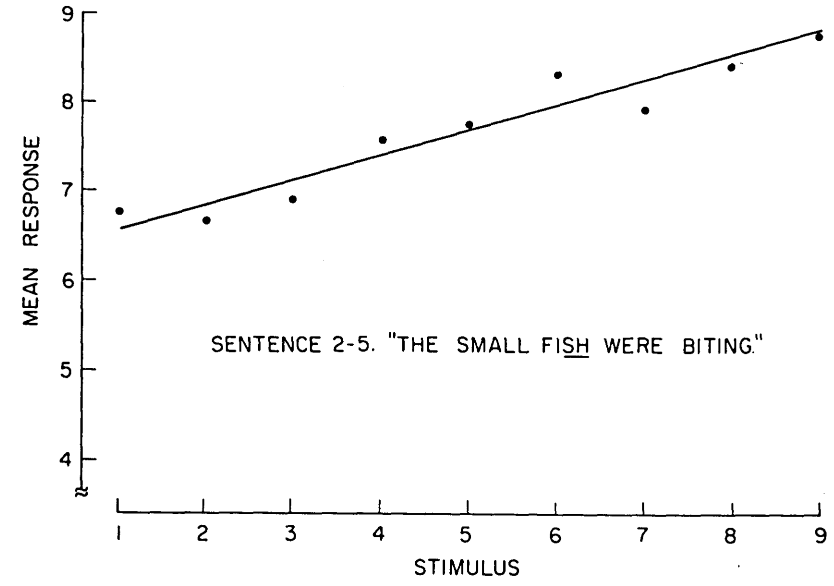


Fig. XIII-5. Mean category judgment as a function of stimulus number for the sentence "The small fish were biting." Note that the /š/ was judged longer than normal even in stimulus 1, which was obtained by excising 60 ms from the original /š/ duration.

position of the segment. Since data from individual subjects were insufficient to make quantitative claims, we used pooled data from all ten subjects. The pooled data allow comparisons of relative sensitivity across conditions, but it should be noted that JNDs obtained from pooled responses may underestimate individual discrimination abilities if subjects have different preferred durations. Aspects of the data from individual subjects will be discussed later.

It is assumed that the behavior of the subjects, as reflected by the pooled confusion matrix for each sentence, can be described by the following standard decision model.[21] The stimulus duration D of a segment results in a sensation magnitude X. Because of internal noise, external noise, and other factors, the distribution of sensation magnitude X resulting from the physical duration D is normally distributed with a mean $\mu(D)$ and a standard deviation σ. The stimuli $I = 1, 2, \ldots, 9$ have different durations D_I which result in different average sensation magnitudes $\mu(D_I)$, but it is assumed that the standard deviations of the X distributions are not a function of D_I.

The situation is illustrated in Fig. XIII-6, which is based on data obtained from the sentence "Bill will play dealer's choice." A subject behaves as if he compared the sensation magnitude X of an unknown stimulus with 8 response criteria thresholds that are specific to the sentence in question, and determined his response R on the basis of the number of thresholds that are exceeded plus one. The criteria threshold locations of the model are optimized to take into account response biases of the average subject.

From the confusion matrix data, it is possible to find the 8 criteria thresholds and the 8 differences between adjacent sensation magnitude means that best fit the data. An example of the results of these calculations is indicated in Fig. XIII-6. The most important characteristic of the model is that if σ is normalized to be 1.0, the difference between sensation magnitude means is equal to d', the standard psychophysical measure of stimulus discriminability. For example, d' for stimuli 7 and 8 in Fig. XIII-6 is approximately 1.0. Therefore these stimuli could be discriminated as different approximately 75% of the time in an appropriately designed paired discrimination test.

The cumulative d' values associated with going from stimulus 1 to stimulus 9 are a reasonable measure of the average discriminability of segment duration along the stimulus continuum for each sentence. These values have been converted to a just-noticeable difference (in ms) by computing the change in duration corresponding to a d' of 1.0. The JNDs are presented in Table XIII-4, column 4.

The data shown in Fig. XIII-6 do not appear to fit the model as well as we might hope. It is surprising and counterintuitive to discover that sensation magnitude is not a smooth function of stimulus duration. Intuitively, d' should be approximately equal for adjacent pairs of stimuli. There were only 30 responses in each confusion matrix row, whereas the model is usually applied in experiments involving considerably more

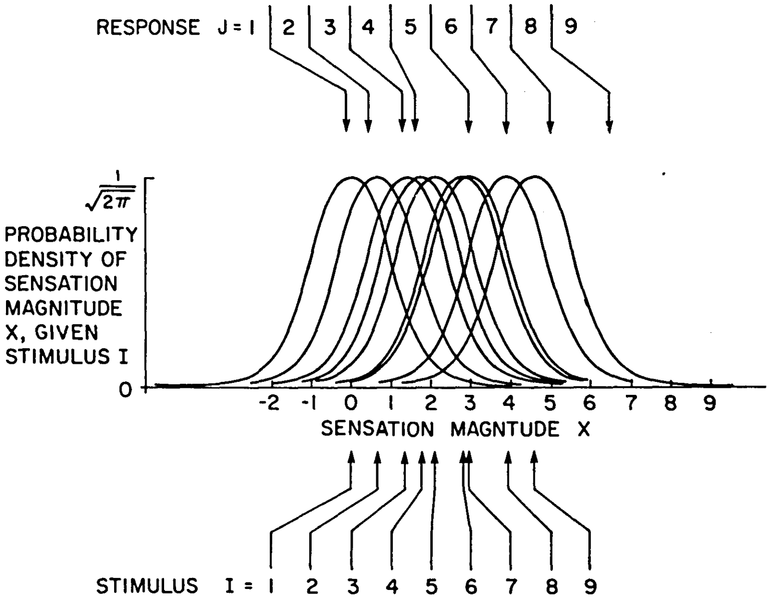


Fig. XIII-6. Distributions of sensation magnitude X for stimuli 1-9 plotted in units of d'. The differences between distribution means and the response criteria thresholds are based on the best-fit decision model to the pooled data from the sentence "Bill will play dealer's choice."

data from only one subject. Thus some caution must be used in drawing conclusions from the JND values presented in Table XIII-4.

Data from individual subjects were compared by making the following approximation. All trials for a given sentence type can be combined to yield 27 responses (9 category judgments three times). The responses to adjacent pairs of stimuli can be compared to see if the rank order of the responses is such that the stimulus of longer duration received a larger response number than the stimulus of the pair of smaller duration, whether both received the same response number, or the response numbers were reversed. There are 24 pairs of stimuli for which this comparison can be made for each sentence for each subject. Assuming that these comparisons were the responses of a single two-category judgment experiment, we computed the d' for each subject. Results did not fit the model very well in some cases, but general statements can be made. The JNDs thus obtained were in agreement with those presented in Table XIII-4, and the best subject is nearly twice as sensitive to changes in stimulus duration as the worst subject.

e. Production Data from Subjects

A few weeks after the perception tests, 4 subjects returned to read the sentences in random order. Spectrograms were made of the recordings and segmental durations were measured. The /il/ durations in the "deal" sentences were found to be somewhat longer than those of the original speaker. But the rank ordering of /il/ durations across sentences is nearly perfect when each subject is paired against the original speaker. Only two of 24 rank-order comparisons failed to agree.

In the sentences of the second experiment, the duration of the vowel in "fish" was found to be short, except in sentence-final position. There is some variability among speakers, but a perfect rank-order correlation exists between the median data from the four subjects and the data from the original speaker. The durations of the fricative /š/ display more variability and disagreement among all speakers. The median data from the four subjects indicates that an utterance-final or nearly final fricative is longer, and that fricatives are shortened when adjacent to other consonants.[11] The overall picture suggests that phrase-final lengthening within a sentence is optional in a short phrase, depending on whether the reader wishes to call attention to the syntactic boundary.

4. Discussion

The preferred durations in the perception of /il/ in sentences from Experiment 1 agree well with the measured durations from the original recordings and with theoretical predictions.[1] As we have noted, the differences in /il/ duration across sentence environments seen in both the original recordings and in the preferred durations are somewhat smaller than those observed previously for the same speaker reading a meaningful connected discourse.[1] It may be that internal phrase boundaries are marked by duration increases to a greater extent in longer, more complex sentences or when sentences form a meaningful part of a connected discourse. Alternatively, subjects in the present experiments may have been making some absolute duration magnitude estimates, even though the instructions were to rate the duration relative to an expected natural duration for a particular sentence. This possibility is suggested because the responses clearly indicate a tendency to judge stimuli from sentences with a normally short duration segment as too short, and segments in utterance-final positions as too long. An experiment is planned to test this hypothesis.

The preferred durations for the fricative /š/ in sentences from Experiment 2 also agree with measured durations and theoretical predictions, except in phrase-final positions. The two sentences involving a noun-phrase final use and an utterance-final use of the word "fish" produced data indicating poor sensitivity to durational changes, and shorter preferred durations than expected. These data are intriguing, indicating either an error in stimulus preparation or a significant deviation from the expected pattern. Having checked the stimuli carefully, we offer a tentative explanation.

Consider the hypothesis that the end of a fricative in utterance-final position is difficult to perceive because of the gradual decrease in sound intensity with falling subglottal pressure. In sentence-medial positions, a phonetic segment with a rapid voicing onset would normally follow the /š/. Voicing onset is a well-defined acoustic event that can be interpreted as the end of the fricative. Therefore the fact that the JND is almost 4 times smaller in the utterance-final word "fishing" than in the utterance-final word "fish" can be explained in part by the absence of a clear acoustic cue to the termination of the frication energy in the latter word. It is not surprising that perceptual strategies would evolve to make use of event onset times to determine fricative terminations, since fricative duration is an important cue to the voicing feature for fricatives in English,[1] and room noises and reverberation often mask the true offsets of weak speech sounds.

Given this hypothesis, an explanation can also be offered for the unusual rating behavior of subjects to the sentence "The small fish were biting." Perhaps this /š/ was judged long even when 60 ms was excised from the /š/ duration of the original recording because subjects were not responding to /š/ duration directly, but rather to sentence rhythm, as reflected in the times between stressed vowel onsets.[22, 23] There was no sudden voicing onset following the /š/ in this sentence because the next segment, /w/, was partially aspirated and had a gradual onset of voicing. Thus changes to the duration of the /š/ in this sentence could only be perceived insofar as they produced a perceptible change in the rhythm of the sentence, as reflected in the time of release of the /b/ of "biting".

The JND data from the present experiments are important because they were obtained from fairly natural sentence materials. The JND results indicate that sensitivity to changes in the duration of a vowel or postvocalic fricative is a function of at least three variables: (i) the absolute duration of the segment, i. e., there was a general tendency for the JND data to conform to Weber's law, (ii) syllable position within the word, i. e., the JND was smaller for segments in the first syllable of a two-syllable word than for segments of comparable duration in word-final syllables, and (iii) word position within the sentence.

An unexpected result of Experiment 1 suggests that backward masking may play a role in determining JND values. The JND for /i/ in sentences involving the single-syllable word "deal" is smaller in utterance-final position even though the absolute duration of the /i/ is longest. There must be no difficulty in determining vowel duration when using either the timing of the formant transition cues in /il/ or the offset of voicing at the end of the utterance. This result suggests that some sort of backward masking effect is present; i.e., if other words follow "deal", the JND increases. It is surprising to find indications of backward masking on a variable such as segment duration. The effect spans too long a time period to be related to peripheral masking,[24] but it may bear some relation to backward masking that has been observed when subjects have to make categorical decisions based on information contained in precategorical storage.[25]

As we have mentioned, the situation for utterance-final /š/ is just the opposite. The JND is very large, thereby indicating the inability of subjects to determine frication offset in an utterance-final fricative.

The present data may be used to estimate the accuracy with which segmental durations must be computed in, for example, a speech synthesis by rule program. If all other segment durations are produced with good accuracy, an error can be tolerated in the duration of a segment such as a vowel or postvocalic fricative consonant of one or more JNDs. Thus the minimum accuracy that would be required in this situation is of the order of 25-100 ms. This is a surprisingly large tolerance limit, given some previous experimental results,[26, 2] but our results probably reasonably reflect the difference between laboratory measurements made on isolated sounds and human performance in more natural speech context. (The Ruhm study[26] also employed an unorthodox definition of JND but the remarkable performance of subjects in the Nooteboom experiment is most likely due to the simplified experimental task.)

Huggins[16] measured the JND for different types of phonetic segments embedded in a single sentence. He found that a 40-ms or greater range of durational changes was accepted as normal (not long or short) by his subjects. He also investigated the matter of whether subjects would change their judgments of normality of a given segment duration if an adjacent segment were lengthened or shortened.[22] He found a small negative correlation between the changes to two adjacent segments if and only if the change to the second segment would reestablish isochrony, i. e., restore the onset time of the next stressed vowel to its proper relationship with the previous stressed-vowel onset. For example, there was no interaction between members of a stressed consonant-vowel sequence, but interactions did occur across word boundaries and within unstressed syllables.

Fujisaki et al.[15] studied the discrimination of changes in segmental duration in Japanese vowels and consonants. Their data are of interest because in Japanese there exist minimal word pairs differentiated primarily on the basis of the duration of a vowel or a consonant. They found that the JND near a phoneme boundary for a two-syllable word spoken in isolation or placed in a carrier sentence was ~10 ms (for a segment duration of ~100 ms) for vowels, fricatives, plosives, and nasals. The somewhat better performance that they report may be due to improved discrimination at a phoneme boundary, or to the fact that each experiment involved only a single word or carrier sentence.

In addition to implications for speech synthesis by rule, results of the present study open up the possibility that the perception of segment duration is used actively in the decoding of constituent structure during normal listening. This possibility has received little attention in previous studies of sentence perception.[27]

The experimental paradigm used in the present study can be applied to this problem in several ways. As a first step, a more detailed comparison must be provided[15] between the perception of segment durations in different syntactic environments in which the possible influence of factors other than syntactic environment per se (factors such as the phonetic environment and the number of words or syllables in the phrase or sentence) is neutralized. The sentence materials and data base in the present study were not entirely appropriate for such comparisons. Possibilities for future work include comparing the perception of segment durations in the following environments: (a) other head noun vs verb environments, as in "The deal rotates to the left" vs "Henry will deal the next hand," and (b) head noun immediately dominated by sentence node vs head noun immediately dominated by verb-phrase node, as in "The deal rotates to the left" vs "Tom understands the deal thoroughly." It is hoped that a study of systematic comparisons and more work along similar lines (e.g., perception of "garden path" sentences) will allow us to determine the extent to which the perception of phonetic segments is used in the decoding of constituent structure.

We thank Richard P. Lippmann for providing a computer program for the estimation of d' and Steven LaPointe for assistance in running the subjects.

References

1. D. H. Klatt, "Vowel Lengthening Is Syntactically Determined in a Connected Discourse," J. Phonetics 3, 161-172 (1975).
2. S. G. Nooteboom, "The Perceptual Reality of Some Prosodic Durations," J. Phonetics 1, 25-45 (1973).
3. D. B. Fry, "Experiments in the Perception of Stress," Language and Speech 1, 126-152 (1958).
4. P. Denes, "Effect of Duration on the Perception of Voicing," J. Acoust. Soc. Am. 27, 761-764 (1955).
5. L. J. Raphael, "Preceding Vowel Duration as a Cue to the Voicing Characteristics of Word-Final Consonants in English," J. Acoust. Soc. Am. 51, 1296-1303 (1972).
6. S. G. Nooteboom, "Contextual Variation and the Perception of Phonemic Vowel Length," Proc. Speech Communication Seminar, Stockholm, Vol. 3, pp. 149-154 (Almqvist and Wiksell, Uppsala, 1974).
7. B. Lindblom and K. Rapp, "Some Temporal Regularities of Spoken Swedish," Papers from the Institute of Linguistics, University of Stockholm, Publication 21, 1973.
8. D. K. Oller, "The Duration of Speech Segments: The Effect of Position in Utterance and Word Length," J. Acoust. Soc. Am. 54, 1235-1247 (1973).
9. I. Lehiste, "Rhythmic Units and Syntactic Units in Production and Perception," J. Acoust. Soc. Am. 54, 1228-1234 (1973).
10. M. H. O'Malley, D. R. Kloker, and D. Dara-Abrams, "Recovering Parentheses from Spoken Algebraic Expressions," IEEE Trans., Vol. AU-21, No. 2, pp. 217-220, 1973.
11. D. H. Klatt, "The Duration of [s] in English Words," J. Speech Hear. Res. 17, 51-63 (1974).
12. S. M. Abel, "Duration Discrimination of Noise and Tone Bursts," J. Acoust. Soc. Am. 51, 1219-1223 (1972).

13. C. D. Creelman, "Human Discrimination of Auditory Duration," J. Acoust. Soc. Am. 34, 582-593 (1962).

14. A. M. Small and R. A. Campbell, "Temporal Differential Sensitivity for Auditory Stimuli," Am. J. Psychol. 53, 329-353 (1962).

15. H. Fujisaki, K. Nakamura, and T. Imoto, "Auditory Perception of Duration of Speech and Non-Speech Stimuli," Proc. Symposium on Auditory Analysis and Perception of Speech, Leningrad, August 1973.

16. A. W. F. Huggins, "Just-Noticeable Difference for Segment Duration in Natural Speech," J. Acoust. Soc. Am. 51, 1270-1278 (1972).

17. A. J. Presti, "High-Speed Sound Spectrograph," J. Acoust. Soc. Am. 40, 628-634 (1966).

18. A. W. F. Huggins, "A Facility for Studying Perception of Timing in Natural Speech," Quarterly Progress Report No. 95, Research Laboratory of Electronics, M.I.T., October 15, 1969, pp. 81-83.

19. L. D. Braida and N. I. Durlach, "Intensity Perception II. Resolution in One-Interval Paradigms," J. Acoust. Soc. Am. 51, 483-502 (1972).

20. D. H. Klatt, "Interaction between Two Factors That Influence Vowel Duration," J. Acoust. Soc. Am. 54, 1102-1104 (1973).

21. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (John Wiley and Sons, Inc., New York, 1960).

22. A. W. F. Huggins, "On the Perception of Temporal Phenomena in Speech," J. Acoust. Soc. Am. 51, 1279-1290 (1972).

23. G. D. Allen, "The Location of Rhythmic Stress Beats in English: An Experimental Study I," Language and Speech 15, 72-100 (1972).

24. L. L. Elliott, "Backward Masking: Monodic and Dichotic Conditions," J. Acoust. Soc. Am. 34, 1108-1115 (1962).

25. D. W. Massaro, "Perceptual Auditory Storage in Speech Recognition," Proc. Symposium on Dynamic Aspects of Speech Perception, Institute for Perception Research, Eindhoven, Holland, August 6-8, 1975 (to be published by Springer Verlag).

26. H. B. Ruhm, E. O. Mencke, B. Milburn, W. A. Cooper, Jr., and D. E. Rose, "Differential Sensitivity to Duration of Acoustic Signals," J. Speech Hear. Res. 9, 371-384 (1966).

27. J. A. Fodor, T. G. Bever, and M. F. Garrett, *The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar* (McGraw-Hill Book Company, New York, 1974), Chaps. 5 and 6.

B. AN EXPERIMENT ON "PHONETIC ADAPTATION"

National Institutes of Health (Grants HD05168-04 and 5 RO1 NS04332-12)

William F. Ganong III

1. Introduction

Several studies have attempted to determine exactly what sort of feature detectors are fatigued during "phonetic adaptation."[1-4] Many of these studies have addressed the question of whether the observed changes in perception are due to the fatigue of detectors sensitive to the acoustic details of the experimental stimuli or of detectors sensitive to the linguistic descriptions of these sounds. This question is hard to answer because, generally, sounds with similar acoustic manifestation are similar linguistically, and vice versa.

The experiment reported here attempts to answer this question for the case of adaptation measured on a /bae-dae/ test continuum. Repetition of the most /dae/-like test stimulus causes subjects to hear more of the test stimuli as /bae/.[3] Is this change in perception due to the fatigue of detectors for falling formant transitions (the only acoustic difference between /bae/ and /dae/ test stimuli) or to the fatigue of detectors sensitive to the linguistic feature "place of articulation" (the only linguistic difference between b and d)? Dr. A. M. Liberman of Haskins Laboratories has suggested that the syllable /sae/ could be synthesized to as to distinguish between these two hypotheses. The consonants in /sae/ and /dae/ share the same place of articulation and, therefore, by the "linguistic" hypothesis, ought to have about the same adaptation effect. A "normal" /sae/ also has formant transitions similar to those of /dae/ and, according to the acoustic hypothesis, should have an effect similar to that of /dae/. We can, however, synthesize /sae/ by joining the fricative noise characteristic of /s/ to the steady-state vowel /ae/. The result, "/sae/ without formant transitions" (SWT), sounds to listeners like /sae/, and therefore should have the same adaptation effect as /dae/, if the linguistic hypothesis is correct. If the acoustic hypothesis is correct, however, it should have no effect, since it lacks formant transitions. Comparison of the adaptation effects of /dae/, /sae/, and SWT should then enable distinguishing between the two theories.

R. Diehl[4] reported a significant adaptation effect (measured with a b-d continuum) by using an adapting stimulus that was a short noise burst followed by a steady-state vowel. Subjects who thought this stimulus sounded like /t/ reported hearing fewer "d"s when listening to the test continuum after repetition of this stimulus. I decided to add a "/tae/ without formant transitions" (TWT) which was a short piece of fricative noise of the /s/ followed by the steady-state vowel /ae/, to see if such a stimulus would also have an adaptation effect. Again the two theories make quite different predictions. Since

/t/ shares the same place of articulation with /d/ and /s/, TWT, according to the linguistic hypothesis, should have an adaptation effect similar to that of /sae/ and /dae/; in contrast, the acoustic theory predicts that TWT, lacking formant transitions, will have no effect.

2. Stimuli

All stimuli used in this experiment were synthesized with a digitally simulated (10 kHz sampling rate) 5-pole series-resonance terminal analog synthesizer,[5] and recorded on audio tape. The stimuli all shared the same steady-state vowel (formants at 800, 1700, 2400, 3500, and 4500 Hz), and the same fundamental frequency contour (a linear decline from 120 Hz to 90 Hz during the 270 ms of voicing).

The /bae/-/dae/ test series: Good exemplars of /bae/ and /dae/ were synthesized which differed only in their second- and third-formant transitions during the first 40 ms of voicing. The first formant rose linearly from 400 to 800 Hz in both stimuli, while the second and third formants moved linearly from their starting frequencies to their steady-state frequencies. For F2 and F3 the starting frequencies for /bae/ were 1100 Hz and 1600 Hz, for /dae/, 1900 Hz and 3000 Hz. The differences in starting frequencies in F2 and F3 were broken into 12 equal steps (one step = a change of 67 Hz in F2, 117 Hz in F3). Then 9 stimuli, corresponding to 0, 2, 4, 5, 6, 7, 8, 10, and 12 steps were synthesized (stimulus 0 = /bae/, stimulus 12 = /dae/). After 40 ms of formant transition, each test stimulus had 230 ms of a steady-state vowel.

Adapting stimuli: The /dae/ adapting stimulus was exactly the same as test stimulus No. 12. It had 40 ms of formant transitions (F1, F2, and F3 moved linearly from 400, 1900, and 3000 Hz to their steady-state values), and a 230-ms vowel. The /sae/

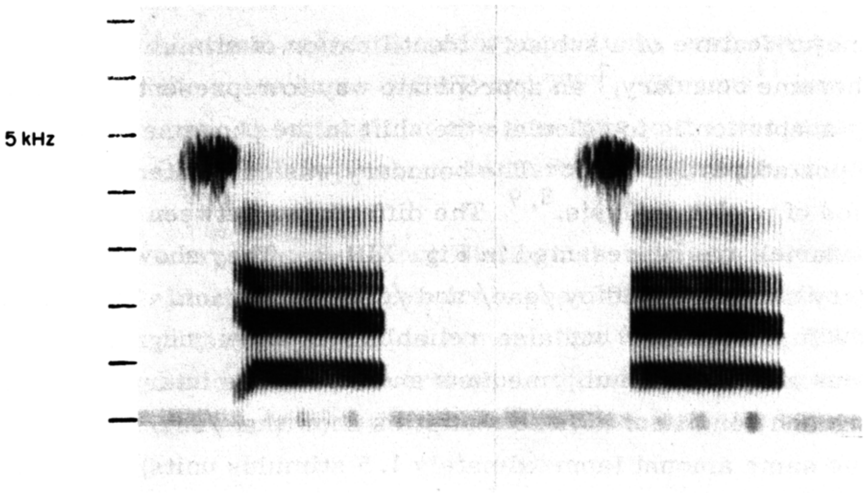


Fig. XIII-7. Spectrograms of /sae/ and SWT adapting stimuli.

adapting stimulus had 100 ms of /s/ noise followed by an acoustic pattern similar to that of /dae/. The fricative noise for /s/ was synthesized by filtering broadband noise through the fifth-formant filter (frequency 4500 Hz, bandwidth 250 Hz), and had peak amplitude 15 dB above the fifth-formant amplitude during the vowel. The amplitude rose linearly (in dB) 15 dB to the peak amplitude 20 ms before voicing onset, then fell linearly 10 dB. The only difference between the voiced portion of/sae/ and the /dae/ adapting stimulus was that F3 started at 2700 Hz rather than at 3000 Hz at the beginning of the formant transitions in /sae/. The SWT adapting stimulus used the same frication source as /sae/, and followed it with a steady-state vowel 270 ms long. Spectrograms of /sae/ and SWT are given in Fig. XIII-7. The TWT adapting stimulus used the fifth-formant filter to produce a 15-ms noise burst 9 dB above the amplitude of the fifth formant in the vowel. After the noise burst and a 15-ms pause, this stimulus had a 270-ms steady-state vowel.

3. Procedure

Twenty subjects from the subject pool of the Department of Psychology, M. I. T., agreed to participate in the experiment, which had 4 one-hour sessions, one session for each adapting stimulus. Each session included some practice at identifying the test stimuli as "b" or "d", a preadaptation identification test containing 16 presentations of each of the 9 stimuli in random order, and a post-adaptation identification test. This included 16 repetitions of each of the 9 stimuli, with each test stimulus preceded by 8 repetitions of that day's adapting sound.[6] Stimuli were played on an Ampex PR-10 tape recorder, amplified, and delivered to subjects at constant amplitude through headphones.

4. Results

Since the major feature of a subject's identification of stimuli on a test continuum is the sharp phoneme boundary,[7] an appropriate way to represent the change in perception induced by adaptation is to calculate the shift in the phoneme boundary between preadaptation and postadaptation tests. The boundary was calculated in each condition by using the method of probit analysis.[8, 9] The differences between preadaptation and postadaptation boundaries are represented in Fig. XIII-8. They show clearly the strong and reliable boundary shifts induced by /sae/ and /dae/ adaptation. The adaptation effects of SWT and TWT are weaker but also reliable. To allow judgments of the relative effects of various adapting stimuli, medians and confidence intervals for the median are also plotted for each condition. These statistics show that /dae/ and /sae/ both moved the boundary the same amount (approximately 1.5 stimulus units). SWT and TWT had smaller effects, each moving the boundary approximately .5 stimulus unit.

How do these results answer the question that the experiment was supposed to

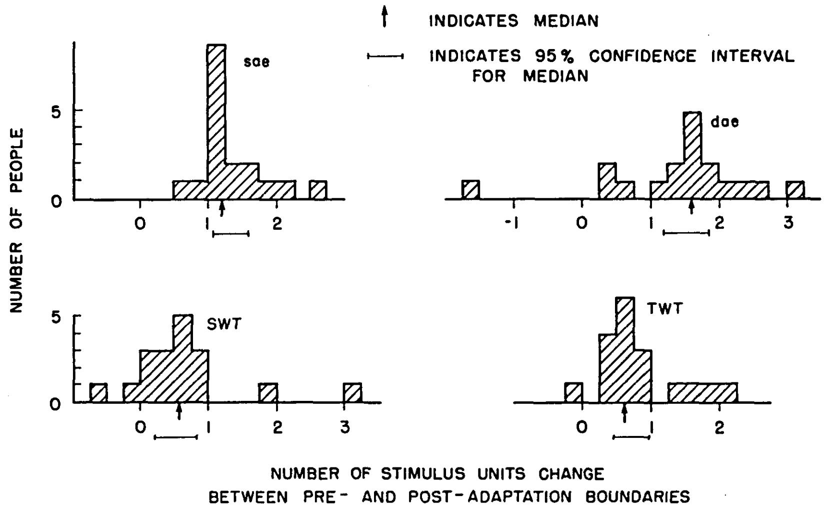


Fig. XIII-8. Adaptation-induced boundary shifts.

answer? Not very well; there seem to be both large acoustic effects (/sae/ and /dae/, with formant transitions, moved the boundary substantially), and weak linguistic effects not mediated by formant transitions (SWT and TWT). One could then answer the question by saying that elements of both acoustic and linguistic fatigue are involved in the phonetic adaptation effect, but this is not the only possible explanation for these results. With a few post hoc modifications, either acoustic or linguistic models can account for these results. The linguistic model can explain the weak effect of SWT and TWT by positing that these stimuli are "weak exemplars" of their linguistic classes; in fact, subjects reported that these stimuli did "fly apart" (in an effect rather like "auditory streaming"[10]) much more easily than did either /sae/ or /dae/. The acoustic theory can explain these results by denying the proposition that the acoustic detectors mediating the discrimination between /b/ and /d/ attend only to formant transitions, and instead claim that they would be influenced by the noise burst in SWT and TWT. Stevens[11] has proposed, in fact, that there are acoustic detectors that are sensitive to the whole pattern of bursts and formant transitions, to explain how a child learns to discriminate place of articulation.

The results of this experiment do not distinguish between linguistic and acoustic models of adaptation, although they do constrain possible models of the adaptation effect. They also suggest that it may be invalid to assume that the fatigued detectors are all either acoustic or linguistic. A better strategy is to ask "Can we experimentally isolate acoustic and linguistic components of the adaptation process?"

Footnotes and References

1. P. D. Eimas and J. D. Corbit, "Selective Adaptation of Linguistic Feature Detectors," Cognitive Psychol. 4, 99-109 (1973).

2. W. E. Cooper, "Selective Adaptation for Acoustic Cues of Voicing in Initial Stops," J. Phonetics 2, 303-313 (1974).

3. A. A. Ades, "How Phonetic Is Selective Adaptation? Experiments on Syllable Position and Vowel Environment," Percep. Psychophys. 16, 61-66 (1974).

4. R. L. Diehl, "The Effect of Selective Adaptation on the Identification of Speech Sounds," Percep. Psychophys. 17, 45-52 (1975).

5. The synthesizer was implemented by D. H. Klatt on the computer system of the Speech Communication Group of the Research Laboratory of Electronics, and has been described by D. H. Klatt, "Acoustic Theory of Terminal Analog Speech Synthesis," Proc. 1972 Conference on Speech Communication and Processing, April 24-26, 1972, Boston, Mass., IEEE Catalog No. 72 CHO 596-7 AE, pp. 131-134 (1972).

6. While most phonetic adaptation studies have used more repetitions of the adapting stimuli, P. Bailey, in "Procedural Variables in Speech Adaptation," Speech Percep., Vol. 2, No. 3, pp. 27-34, 1974 (Queen's University of Belfast, Northern Ireland), found no significant difference in the adaptation effects of 8, 16, 24, and 32 repetitions of the adapting stimulus.

7. I. G. Mattingly, A. M. Liberman, A. K. Syrdal, and T. Halwes, "Discrimination in Speech and Non-speech Modes, "Cognitive Psychol. 2, 131-157 (1971).

8. D. J. Finney, Probit Analysis (Cambridge University Press, London, 1962).

9. Two subjects produced data unsuitable for probit analysis. These subjects called stimulus 0 (the most b-like stimulus, for most subjects) "d" more than half the time. Like other subjects, they called stimulus 4 "b" almost all the time, and stimulus 12, "d". Thus these subjects had two b-d boundaries, so probit analysis, which expects only one, is inappropriate. One subject produced this sort of anomalous data in only one condition, but another consistently labeled stimulus 0 "d", before and after adaptation. I asked him to return for an extra session, and this time told him to write down whatever consonant he thought he heard while listening to the preadaptation test tape. All of his responses were "b", "d", or "r". He called the low-numbered stimuli "r", the middle stimuli "b", and the high-numbered stimuli "d". The r-b phoneme boundary coincided with his second (anomalous) b-d boundary when he was asked to respond only "b" or "d". Stimulus 0 shares some acoustic properties with "r", a low F3 and an onset of F1 at a higher frequency than is common for stops, so these "r" responses make some sense. I conclude that in the original experiment this subject heard the low-numbered stimuli as "r", and when forced to choose between "b" and "d", chose "d", which is linguistically more like "r".

10. R. A. Cole and B. Scott, "Perception of Temporal Order in Speech: The Role of Vowel Transitions," Can. J. Psychol. 27, 441-449 (1973).

11. K. N. Stevens, "The Potential Role of Property Detectors in the Perception of Consonants" (to appear in J. Phonetics).

C. SOME SPEAKER-IDENTIFYING FEATURES BASED ON FORMANT TRACKS

U.S. Navy Office of Naval Research (Contract N00014-67-A-0204-0069)

Ursula G. Goldstein

1. Introduction

Methods of achieving automatic speaker recognition generally fall into two categories: template matching and feature extraction. The template matching method makes a decision about the identity of the speaker on the basis of the mathematical proximity of the sample utterance to a reference, but does not make a detailed comparison of certain acoustic events arising from individual speech sounds. This approach might be applicable for speaker verification, where the speaker is cooperative and would not purposely introduce large variations in the speech sample. Its advantage lies in its simplicity. Most template matching schemes perform some form of time and amplitude normalization on the unknown and then calculate the distance of this unknown from the reference that it is supposed to represent. If the distance is larger than a certain threshold, an answer of no-match is given.[1-6]

The speech-theoretic approach examines linguistic units and tries to extract an optimum set of features, thereby eliminating from further consideration some of the information in the speech signal that does not pertain to the speaker's identity. Several criteria have been suggested for selecting these features.[7] They should occur frequently in normal speech, vary widely among speakers but not for a given speaker, not change over time, not be affected by background noise or poor transmission, not be affected by conscious efforts to disguise the voice, and be easily measurable.

The list of possible identifying features is virtually endless, and has been only partially examined. A data base that has given promise of providing some useful features is the set of formant tracks obtained from diphthongs, tense vowels, and r-colored sounds. These sounds have evidenced a large amount of variability from one speaker to another, especially across different dialects.[8-11]

The use of formant information in speaker identification systems has been limited almost exclusively to the measurement of formant frequencies inside a single window at the center of a vowel, leaving much of the formant structure unexplored.[7, 12] One measurement that did include a larger amount of formant-structure information is the slope estimate of the second formant of [aI], which Sambur[12] ranked as the tenth best of a large number of attributes that he examined. The success of this measurement gave further evidence that a closer examination of formant tracks might reveal some useful speaker-identifying features.

The purpose of our study was to investigate the formant trajectories of the diphthongs [ɔɪ], [aɪ], and [aʊ], the tense vowels [o], [e], [i], and [u], and retroflex sounds in 3 phonetic environments [rɛ], [ɝ], and [ar], in order to find and evaluate statistically features that could be relevant to speaker identification.

2. Procedure

a. Data Base

Ten adult male speakers of American English with no noticeable foreign accents, strong regional dialects, or speech defects were chosen to make recordings of the sounds to be studied. In order to facilitate comparison of one sound with another, all sounds were recorded in the context, "Say b__d again." Each person repeated each of the 10 sentences 5 times in one recording session. Six of the original 10 speakers returned at least 3 weeks after the first recording session to make another set of recordings, again with 5 repetitions of each of the 10 sentences. This second set of recordings was made in order to check on changes in the speakers' voices over time.

b. Formant Tracking Procedure

Recordings were processed semiautomatically using linear prediction analysis on a PDP-9 computer specially set up for speech analysis. The software written for this purpose seeks to minimize the amount of computer and operator time needed to compute a highly accurate set of formant tracks. For several reasons, no attempt was made to automate this procedure fully. It is not possible to devise an automatic formant-tracking program that gives 4 formants with 100% reliability. The best systems giving 3 formants generally have very complicated decision algorithms, and even then cannot absolutely guarantee 100% reliability.[13] Since an error in formant identification can produce serious errors in a speaker identification scheme based on specific formant frequencies, and the major emphasis of the study is on the use of formant tracks rather than on their computation, we decided that an interactive system should be constructed. With such a system, the operator can observe the results of a first automatic stage of tracking and can intervene to correct errors manually. Informal observations by the program user during the process of trying to identify missing formants or eliminate extraneous ones could be useful in the construction of a more automatic system at a later time.

Audio input was preemphasized 6 dB/octave, bandlimited to 5 kHz, and sampled at 10 kHz. The sampled signal was displayed on a cathode-ray tube, and then marked by hand to indicate the beginning and end of processing. The beginning was defined at 20 ms after the noise burst indicating the release of [b]. The end was marked when a sudden drop in amplitude and an obvious loss of high frequencies indicated the closure for [d].

The first main processing loop of the formant tracking program calculates 12 predictor coefficients for each 10-ms frame, using the covariance method pitch asynchronously, and stores these values on a disk.[14,15] The second loop calculates pole locations of the transfer function for each frame and then transforms them to formants and bandwidths.[16] Formants with bandwidths greater than 700 Hz are removed from the general formant array and placed in a temporary location for extraneous formants. The last phase of the program displays a set of formant tracks, as shown in Fig. XIII-9, and allows corrections to be entered manually by the operator, according to continuity considerations and his knowledge of acoustic theory. Formants can be restored from the temporary locations if they are judged to be not extraneous. Formants can also be removed if they seem extraneous but are not eliminated automatically. On very rare occasions, the root-finding subroutine could not find the roots in 10 iterations. In this case, the formant frequencies were set to zero by the program and were later filled in by averaging the formants of the previous and the next frame. During the course of the measurements, this situation occurred twice. Corrected formant tracks were appropriately labeled and stored on DEC-tape.

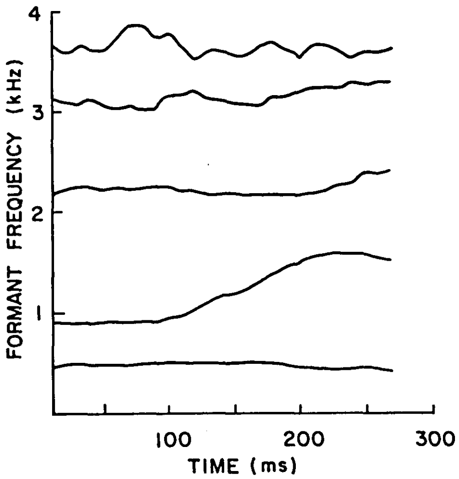


Fig. XIII-9. Example of a formant track display for the vowel [aɪ]. The lines represent time plots of formant frequencies, the lowest being the first formant and the highest line being the fifth formant. The line representing a particular formant is formed by connecting the estimates of that formant frequency from one frame to the next with straight-line segments.

Occasionally, a situation arose in which the user had trouble deciding which poles to choose as formants. To help in this decision, the program calculated a log-magnitude plot of the spectrum and a pole plot in the z-plane for any frame indicated by the user. In actual practice, however, these plots rarely provided more insight than the actual formant frequencies and bandwidths.

For approximately 5 sentences, it was not possible to mark the beginning and end of processing from a simple visual examination of the time waveform, because of imprecise pronunciation by the subject. In these cases analysis was started well before the expected beginning and stopped after the expected end. The two end markers were then readjusted to the two points where gross discontinuities in the formants indicated the presence of a consonant.

3. Speaker-Identifying Features

The formant tracks, as determined by these methods, contain a mixture of noise, linguistic information, and personal information. The next phase of a speech-theoretic speaker identification system would be to extract the personal information pertaining only to the identity of the speaker. In this study, a total of 199 features were measured and evaluated in terms of effectiveness in speaker identification.

In the measurement process, similar phones were grouped together so that approximately the same set of features was tested for each phone of a given class. The three major classes that were used were (i) diphthongs [ɔɪ], [aɪ], and [aʊ]; (ii) tense vowels [e], [o], [i], and [u]; and (iii) retroflex sounds [rɛ], [ar], and [ɝ].

The total data collected can also be divided into three groups according to when they were recorded. Recordings from the first day for all ten speakers form one group, recordings from the second day for six of the ten speakers form the second group, and recordings from both days for the six speakers form the third group. The statistical evaluation process computed the mean and standard deviation of each feature for each speaker. For days 1 and 2, these calculations were based on 5 repetitions, and for the combined group on 10 repetitions. An F ratio was calculated for each feature for each of the three time groupings.

A higher F ratio indicates a feature that exhibits larger interspeaker variation in relation to the intraspeaker variation, and is therefore usually more useful for speaker identification than a feature with a lower F ratio. Although not as sophisticated as the probability-of-error method used by Sambur,[12] the calculation of F ratios has the advantage of being fast and easy to implement, which is very desirable when dealing with a large number of features. The problems of possible dependence between features and artificially high F ratios caused by one very different speaker must still be addressed. Therefore, the distribution of speaker means was checked visually for all features with high F ratios. Possible dependence between features was tested by computing correlation coefficients.

4. Finding the Most Effective Features

An efficient speaker-identifying system makes use of a small set of highly efficient features. To facilitate the selection of these features, we prepared Table XIII-5, listing

Table XIII-5. Features having average F ratios greater than 60.

Feature	Sound	F Ratio	Possible Disadvantage
MAXF1	ar	119.3	
MAXF1	e	107.4	1 speaker very different
MAXF1	o	103.6	
MINF2	ar	97.2	
AVEF4	aʊ	97.1	
MIN3AV	ɝ	92.2	
FINAL2	aɪ	90.8	
AVEF4	aɪ	88.1	correlated with AVEF4 for aʊ
MAXF1	u	86.2	1 speaker very different
MAXF2	aɪ	86.0	same as FINAL2
MID3AV	ɝ	80.7	same as MIN3AV
MINF3	ɝ	77.3	same as MIN3AV
AVE3M2	rɛ	74.0	
AVEF4	ar	74.0	correlated with AVEF4 for a ʊ
INITL2	e	73.4	
MIDF3	ɝ	73.4	same as MIN3AV
MINF2	e	73.4	same as INITL2
F1MAX2	o	72.5	1 speaker very different
MAXF2	o	71.8	
AVEF3	u	71.0	1 speaker very different
MAXF1	ɝ	70.6	1 speaker very different
MAXF2	aʊ	69.0	
MAXF2	ɔɪ	65.8	correlated with FINAL2 a ɪ
F1MAX2	u	65.4	1 speaker very different
MID2AV	i	63.7	correlated with FINAL2 aɪ
MID1AV	o	63.2	correlated with MAXF1 o
FINAL2	e	62.5	correlated with FINAL2 aɪ
INITL4	ar	61.5	correlated with AVEF4 aʊ
INITL2	i	60.9	correlated with FINAL2 aɪ

Table XIII-6. Description of features.

AVEF3	average third formant not including last 20 ms of formant track
AVEF4	average fourth formant not including last 20 ms of formant track
AVE3M2	average third minus second formant omitting last 20 ms of formant track
FINAL2	F2 measured 20 ms before the end of formant track
F1MAX2	F1 at point in time when F2 reaches a maximum
INITL2	initial F2, measured at the beginning of formant track
INITL4	initial F4, measured 20 ms after the beginning of formant track
MAXF1	maximum first formant
MAXF2	maximum second formant
MIDF2	second formant at midpoint
MIDF3	third formant at midpoint
MID1AV	MIDF1 averaged with 2 surrounding F1 values
MID2AV	MIDF2 averaged with 2 surrounding F2 values
MID3AV	MIDF3 averaged with 2 surrounding F3 values
MINF2	minimum second formant
MINF3	minimum third formant
MIN3AV	MINF3 averaged with 2 surrounding F3 values
TOTALS	total slope, measured by fitting a straight line to the second formant over its total duration

Table XIII-7. Correlation coefficients relating features from Table XIII-5 that did not have immediate disadvantages. Taken on day 1.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
OI	MAXF2	1.00															
AI	AVEF4	0.74	1.00														
AI	FINAL2	0.95	0.76	1.00													
AU	MAXF2	0.73	0.51	0.79	1.00												
AU	AVEF4	0.68	0.80	0.73	0.75	1.00											
ER	MIN3AV	0.30	0.25	0.27	0.40	0.37	1.00										
AR	INITL4	0.46	0.74	0.50	0.42	0.88	0.14	1.00									
AR	MINF2	-0.06	0.21	0.05	-0.15	0.17	0.15	0.18	1.00								
AR	MAXF1	0.47	0.22	0.45	0.22	0.24	0.63	-0.01	0.50	1.00							
AR	AVEF4	0.52	0.76	0.59	0.51	0.91	0.09	0.98	0.21	0.03	1.00						
RE	AVE3M2	0.14	0.26	-0.05	-0.19	-0.08	0.54	-0.12	0.15	0.42	-0.21	1.00					
EI	FINAL2	0.85	0.50	0.84	0.64	0.43	0.41	0.23	-0.14	0.52	0.30	0.11	1.00				
EI	INITL2	0.70	0.45	0.79	0.55	0.35	-0.00	0.24	0.09	0.30	0.35	-0.23	0.83	1.00			
OU	MAXF2	0.48	0.54	0.52	0.65	0.71	0.77	0.58	0.32	0.43	0.58	0.20	0.50	0.36	1.00		
OU	MAXF1	0.10	-0.04	-0.04	-0.03	0.03	0.74	-0.15	0.30	0.77	-0.20	0.70	0.12	-0.28	0.33	1.00	
OU	MID1AV	-0.13	-0.29	-0.22	-0.28	-0.27	0.56	-0.44	0.36	0.74	-0.48	0.57	-0.04	-0.35	0.04	0.91	1.00
I	MID2AV	0.88	0.56	0.90	0.78	0.55	0.43	0.30	-0.14	0.48	0.38	0.02	0.97	0.82	0.56	0.06	-0.12
I	INITL2	0.82	0.51	0.86	0.67	0.47	0.27	0.31	-0.11	0.39	0.39	-0.12	0.96	0.92	0.49	-0.10	-0.24

29 features with average F ratios greater than 60. An explanation of the feature names in Table XIII-5 is given in Table XIII-6. These F ratios were obtained by averaging the three F ratios associated with the three recording times: day 1, day 2, and both days.

An inspection of the individual speaker averages for each feature showed that some of these F ratios were unduly high because of a large variation in one speaker. For this reason, MAXF1 for [e], [u], and [ɝ], F1MAX2 for [o] and [u], and AVEF3 for [u] were eliminated from further consideration. Several sets of features are obviously redundant, since they represented essentially the same acoustical property but with slightly different measurement procedures. Therefore it is best to keep in each set only the one feature with the highest F ratio. This procedure eliminates MAXF2 for [aɪ], MID3AV for [ɝ], MINF3 for [ɝ], MIDF3 for [ɝ], and MINF2 for [e]. In order to further eliminate redundant features, correlation coefficients were calculated for the remaining 18 features, as shown in Table XIII-7. As might be expected, several groups of features are quite dependent. The following groups of features all had relative correlation coefficients greater than .8 within the group:

1. MAXF2 for [ɔɪ], FINAL2 for [aɪ] and [e], MID2AV for [i], and INITL2 for [i]
2. MAXF1 for [o] and MID1Av for [o]
3. AVEF4 for [aʊ], [aɪ], and [ar], and INITL4 for [ar].

Keeping only the feature with the highest F ratio in each group of dependent features leaves 10 features that are listed in Table XIII-8.

One of the original motivations for studying diphthongs and r-colored sounds was the large dialect variation shown by these sounds.[9,11] It can be seen by examining Table XIII-8 that features derived from these sounds, which presumably depend more upon speaker habits than on vocal-tract anatomy, in fact showed large individual differences. Also, the first-formant measures were generally uncorrelated with the second-formant measures, which indicates that one or both of these measures contains more information than just the overall length of the vocal tract.

One feature that was particularly uncorrelated with any of the others was MINF2 for [ar]. This feature gives an indication of how strongly a speaker's [a] is affected by the adjacent [r] and may also depend on the way he shapes his tongue blade for the retroflex [r]. This notion is supported by the low correlation coefficients between MINF2 for [aɪ] and MINF2 for [ar]: .56 for day 1 and .26 for day 2. Three other features in Table XIII-8 are related to retroflex sounds. MIN3AV for [ɝ] is a direct indication of the degree of retroflexion, while AVE3M2 for [rɛ] takes into account both the degree of retroflexion and the duration of [r] relative to [ɛ]. MAXF1 for [ar] shows the influence of the retroflex on F1 of the vowel.

Three features are measurements made on mid vowels. Since mid vowels are not produced with an extreme high or low tongue position, they may be subject to more

Table XIII-8. Ten features of this study that are most likely to be useful for speaker identification.

Feature	Sound	F ratio
MAXF1	ar	119.3
MAXF1	o	103.6
MINF2	ar	97.2
AVEF4	aʊ	97.1
MIN3AV	ɝ	92.2
FINAL2	aɪ	90.8
AVE3M2	rɛ	74.0
INITL2	e	73.4
MAXF2	o	71.8
MAXF2	aʊ	69.0

individual variation than [i] or [u]. Their acoustic characteristics might also reflect the shape of the palate, since these sounds are produced with the sides of the tongue in contact with the lower edges of the palate.[17] Two features are taken from vowel targets of diphthongs. Since the main acoustic cue identifying a diphthong is the rate of formant-frequency change,[18] we might expect the target frequencies of the first and second formants to reflect a person's individual speaking habits.

The feature AVEF4 of [aʊ] is also a measurement taken from a diphthong, but probably reflects mainly vocal-tract shape and size. Since all of the fourth-formant measures in Table XIII-6 were highly correlated, we chose only this one, which has the highest F ratio, for the best-feature list.

The list of "best" features includes two measures involving the third formant and one involving the fourth. Of these three features, the only one that appeared completely reliable was MIN3AV for [ɝ], as evidenced by the consistently narrow bandwidths and good continuity from one frame to the next for F3 of [ɝ]. This was not always the case with AVEF4. For example, the F ratio of AVEF4 for [o] was extremely low because of measurement difficulties encountered with speaker 10. On 2 repetitions a resonance that had been identified as the fourth formant during the other 3 repetitions was too weak to be detected, possibly because of a zero in the spectrum of either the glottal sources or the vocal-tract filter. Spectrographic analysis confirmed the problem. This finding might indicate that the fourth formant makes a rather unreliable feature, and that the high F ratio associated with AVEF4 for [aʊ] was coincidental. Similar measurement problems were encountered with the third formant of speaker 3.

Considering the trouble given by the higher formants during the formant-tracking procedure, it might be more reasonable not even to try to measure them. A system where the speech waveform was lowpass filtered to 2500 Hz and then sampled at 5000 Hz would allow at most 3 formants. The linear-prediction program could then be run with 8 predictor coefficients, and the root finder would have to solve only an eighth-order polynomial. With very little loss of information, such a system could be expected to run approximately twice as fast as the one used in this study.

5. Comparison of Best Features with Previous Work

Since no actual identification experiment was performed in this study, a direct comparison between this study and the results of previous work is rather difficult, but a few simple estimates concerning relative effectiveness of different features can be made.

The comparison with the work of Sambur[12] is facilitated by the fact that our study duplicated 5 of his measurements and very nearly duplicated 4 others, as shown in Table XIII-9. The slope of the second formant of [aɪ], which was ranked as tenth best in Sambur's work, showed an intermediate degree of effectiveness in this work. Sambur's third best measurement, the third formant of [u], was measured at a single point in time in the middle of the vowel. At the outset of this study, it was decided that third and fourth formant measures for tense vowels should be the average frequency of these formants taken over the duration of the vowel because there was very little higher formant

Table XIII-9. Feature performance comparison of this study and that of Sambur.

Sambur Study		This Study			Exact
Name	Ranking	Name		F Ratio	Duplicate ?
UF3	2	AVEF3	u	71.0	no
AI	10	TOTALS	aɪ	39.6	yes
UF2	14	MIDF2	u	36.5	yes
EEF2	15	MIDF2	i	48.4	yes
UF1	18	MIDF1	u	42.5	yes
EEF1	23	MIDF1	i	6.2	yes
EEF4	24	AVEF4	i	44.3	no
EEF3	25	AVEF3	i	39.0	no
UF4	31	AVEF4	u	14.3	no

movement and because the averaging helped reduce some of the noise of the measurement process. Therefore the closest measure available to match with Sambur's UF3 was AVEF3 for [u], which probably had a slightly higher F ratio than a single measurement of F3 would have given. Nevertheless, the average F ratio of 71.0 for AVEF3 of [u] was still considerably lower than the F ratio of 119.3 for MAXF1 of [ar], which was judged to be the most effective feature found in this study. Therefore, the better features of this study would probably be somewhat more effective for speaker identification in comparison with those found by Sambur.

Of interest is the fact that Sambur's work seems to show that first-formant measurements are less important for speaker identification than second or third-formant measurements, whereas in this study the two most effective features were first-formant measurements. A closer examination of the data from F1 measurements shows a very large variation of F ratios according to exactly where the measurement was taken. For example, F = 103.6 for MAXF1 of [o], but F = 12.1 for F1MIN2 of [o]. The maximum value of F1 generally yielded the highest F ratios and was characterized by the lowest intraspeaker standard deviations.

Six of the best measurements found in this study had higher F ratios than the best feature found by Wolf,[7] which had an F ratio of 84.9. Wolf's nine best features were fundamental frequency measures, which Sambur downrated because of their variability from one recording session to another. The feature with the tenth largest F ratio in Wolf's study was the second-formant frequency of [æ], having an F ratio of 46.6. This value is considerably lower than the F ratios of the better features of this study.

6. Limitations of this Study

The most obvious limitation of this study is the small data base that was used and, in particular, the very limited amount of data concerning speaker variation over time. Another problem is the somewhat artificial nature of the recordings. The recordings were made under ideal laboratory conditions with the subjects reading a prearranged set of sentences. In practice, one will probably have to cope with background noise, distortion, or bandlimiting of transmission equipment, and different sentence contexts for the sounds under investigation. Other complications include the possibility of an alteration in a person's voice because of emotional or physical stress, or an uncooperative speaker, i.e., a person who is trying to disguise his voice or mimic another person. All of the speakers in the present study were cooperative and under no particular stress.

On the other hand, the speakers in this study did not represent a cross section of all dialects of American English. The speakers that were chosen originally for the study had no noticeable foreign accents and no strong regional dialects. Therefore the interspeaker variations of some of the features are probably not as high as they might have been with a more varied group of subjects.

Besides extending the work of this study to include some of the additional variables mentioned above, it would be useful to test some of the more effective features in a more rigorous manner. First, we might perform a probability-of-error analysis[12] of these features, together with some of the more successful features of other studies, such as the second formant of [n], the voice-onset time of [k], the third and fourth formants of [m], and an FO measurement. Next, we might run an identification experiment with this combined feature set, using speakers who had not been involved in the original feature evaluation.

Another area of interest might be the study of the higher formants; why they appear and disappear unexpectedly, and how to compensate for this problem. If the higher-formant-measuring problems with speakers 3 and 10 were caused by zeros in the spectrum, perhaps we could devise a system to indicate the presence of a zero, and then determine its frequency.[19]

References

1. S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Am. 35, 354-358 (1963).
2. K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive System," J. Acoust. Soc. Am. 40, 966-978 (1966).
3. J. E. Luck, "Automatic Speaker Verification Using Cepstral Measurements," J. Acoust. Soc. Am. 46, 1026-1032 (1969).
4. S. K. Das, and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans., Vol. AU-19, No. 1, pp. 32-43, March 1971.
5. R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans., Vol. AU-21, No. 2, pp. 80-89, April 1973.
6. A. E. Rosenberg and M. R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans., Vol. ASSP-23, No. 2, pp. 169-176, April 1975.
7. J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition," J. Acoust. Soc. Am. 51, 2044-2056 (1972).
8. A. Holbrook and G. Fairbanks, "Diphthong Formants and Their Movements," J. Speech Hear. Res. 5, 38-58 (1962).
9. W. Labov, M. Yaeger, and R. Steiner, "A Quantative Study of Sound Change in Progress," Report on National Science Foundation Contract NSF-GS-3287, University of Pennsylvania, Philadelphia, 1972.
10. D. Klatt, "Acoustic Characteristics of /w, r, l, y/ in Sentence Contexts," J. Acoust. Soc. Am. 55, 397 (1974).
11. H. Kurath, Handbook of the Linguistic Geography of New England (The American Council of Learned Societies, Providence, R.I., 1939).
12. M. R. Sambur, "Selection of Acoustic Features for Speaker Identification," IEEE Trans., Vol. ASSP-23, No. 2, pp. 176-182, April 1975.
13. Stephanie S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans., Vol. ASSP-22, No. 2, pp. 135-141, April 1974.

14. J. I. Makhoul and J. J. Wolf, "Linear Prediction and the Spectral Analysis of Speech," Bolt, Beranek, and Newman, Inc., Cambridge, Mass., Report 2304, August 1972.

15. J. I. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE 63, 561-580 (1975).

16. M. A. Jenkins and J. F. Traub, "A Three-Stage Algorithm for Real Polynomials Using Quadratic Iteration," SIAM J. Numer. Anal., Vol. 7, No. 4, pp. 545-566, December 1970.

17. K. N. Stevens "Quantal Configurations for Vowels," J. Acoust. Soc. Am., Vol. 57, Supplement No. 1, p. 570, April 1975.

18. T. Gay, "A Perceptual Study of American English Diphthongs," Language and Speech 13, 65-88 (1970).

19. J. M. Tribolet, "Identification of Linear Discrete Systems with Applications to Speech Processing," S.M. Thesis, Department of Electrical Engineering, M.I.T., January 1974.

JSEP

D. COMPUTER-AIDED SIGNAL PROCESSING: HIGHER LEVEL DIALOGUES FOR PROCESSING SCHEMATA SPECIFICATION

Joint Services Electronics Program (Contract DAAB07-74-C-0630)

William L. Henke

During this period progress continued toward the realization of a transparent "higher level" dialogue/notation to be used in the interactive development of signal processing and feature extraction schemata for application to given problem situations. As previously reported, we are using a graphical block diagram type of representation for the sequential uniform-rate signal-processing component of the notation. Recent work has been concerned with extending the block diagram representation in ways that are applicable to uniform-rate processing, and also to the additional and different anticipated needs of more "event" or "feature" oriented types of processing. These extensions are the resolution of issues concerned with blocks that have multiple outputs (multiple inputs are usual and raise no major issues), and the addition of a macrocircuit or a subcircuit definitional facility. This facility, as well as being of great assistance in the conceptualization and notation of more complex processing schemata, will also introduce another "level" into the hierarchy of schema specification. Such different levels of "subschema binding" are needed when users have different levels of sophistication, as is almost always the case. In this way more sophisticated analysts can bind together processing primitives into schemes that subsequently may be selected easily by nonspecialists who will then only have to adjust parameter values. An interesting feature of the design, which should be an important aid to less sophisticated users, is that both the interconnection topology and the formal or dummy parameter requirements of subcircuits are known when instances of application or invocation are specified, so that conformal requirements are enforced in a way that becomes a powerful aid rather than an undesirable but necessary hindrance. For example, when actual parameter values are needed they are requested one at a time with prompts indicating their function. Much of the design of the subcircuit facility has been completed, and implementation has been started so as to be able to prove the viability and usefulness of these concepts.

JSEP

E. PRELIMINARY REPORT ON LONGITUDINAL INTENSIVE OBSERVATION OF THE ACQUISITION OF ENGLISH AS A FIRST LANGUAGE

National Science Foundation (Grant SOC74-22167)

Lise Menn

A child of English-speaking parents was studied intensively (20 hours per week) over a basic period of 8 1/2 months observation. His age at the beginning of the study was 12 1/2 months, at which time he had acquired one 'word.' Natural interaction patterns were approximated during the observation by the investigator having care of the child. Six to eight hours per week were audiotaped in the child's home, with supplemental periodic tapings in a sound studio at the Research Laboratory of Electronics, M. I. T. and several hours of videotaping in the child's home each month. Periodic assessments of the child's cognitive development have been made by S. Haselkorn, of the Harvard Graduate School of Education.

Extensive instrumental analysis of the corpus is planned; meanwhile, preliminary study of phonetic transcription of the tapes, supplemented by some spectrographic analysis and by use of the voice pitch extraction program developed by Douglas O'Shaughnessey of the Cognitive Information Processing Group support the following claims.

1. The earliest learning of segmental patterns by this child cannot be described felicitously in terms of phonemes or phones, but in terms of holistic lexical items, each of which is represented by a segmental target and a specification of the degree of variation from that target. Such variation was by no means constant from one lexical item to the next. One of the child's early words was not specified more closely than $\begin{bmatrix} {}^{\text{C}}\text{-labial} & {}^{\text{V}}\text{-back} & {}^{\text{C}}\text{-labial} & {}^{\text{V}}\text{-front} \\ \text{-nasal} & & \text{-nasal} & \end{bmatrix}$. A sequence of tokens of this word, glossed as 'thankyou,' showed the sounds which, in adults, would be ascribed to dental, alveolar, palatal, and velar articulations. This sort of variation was not found in the few other words then in the child's vocabulary. Furthermore, two other words, do 'toast' and do 'don't, no' were distinguished solely by nasalization of the vowel; such a fine distinction did not function elsewhere in the child's lexicon. The anomalous words 'thankyou' and 'don't' both disappeared and were replaced by other forms with the same meaning as the child developed an output that was more amenable to description in the familiar terms of phonological theory.

2. The learning of phonemic contrasts and of phonetic targets are distinct but related tasks, the first organizational, the second motoric. The following sequence of events may be interpreted as phonetic target learning followed by mastery of phonemic contrast:

Stage 1. The child possesses output words beginning with dentals or the anomalous nonlabial target of his 'thankyou.'

Stage 2. In addition to the items of stage 1, the child produces several words with velar consonants (these are never subject to replacement by dentals or palatals).

Stage 3. 'Thankyou,' the word that ignores the separation between dental and velar stops, is replaced by a strictly dental form. Here it seems that the learning of separated targets, dental and velar, preceded the organization of the lexicon by the phonemic contrast dental/velar; the rejection of the one form, 'thankyou' which was not categorized by that split, followed.

In contrast, another sequence of events may be interpreted as the manifestation of phonemic (organizational) learning followed by phonetic (motoric) learning.

Stage 4. The child avoids most English words containing labial stops, although he possesses a good number of words containing dentals and velars. When he attempts words with labial stops, he usually deletes or replaces the labial; he rarely has a labial output.

Stage 5. During a certain session, he is observed to close his lips silently at the end of two English words containing 'p' ('up,' 'apple'); at the next session, some tokens of 'p' in these words are released.

Stage 6. Words using initial and final labial stops appear in the child's vocabulary.

In this case, it seems as though the child, in order to avoid labials consistently with the exceptions noted, must have distinguished them as a class distinct from the other stop positions at a time when he found them difficult to say.[1]

It should be emphasized that only a diary study or a large corpus such as the present one can sustain claims that the child avoids certain items. In the present case, there is additional evidence to show that the child recognized such words as 'block,' 'box,' and 'ball' which were in daily use but which he never attempted to say.

3. This child, like the subject of Menyuk and Bernholtz,[2] can be shown to control several distinct intonation patterns which may be manifested on words, and also on babble sequences. Study of sequences of behavior on videotape make it clear that these sequences are not only controlled by the child, but have semantic content. In the broadest terms, rising contours characterize requests for action by adults; falling contours narrate ongoing events. These contrasts are clear at 16 months and probably exist before then. Considerable detailed instrumental study of the phonetic character of these contours will be undertaken, and independent observers will score the correlated videotaped behavior.

4. Adult speech to the child seems to show phonological foregrounding and backgrounding of parts of sentences addressed to him. The portion of a sentence which is foregrounded (pronounced very distinctly) increases in length as the child's ability to speak becomes greater. Instrumental characterization of the subjective notion 'foregrounding' will be attempted.

This child will be followed at the reduced rate of 2 hrs/wk during the next 18 months.

References

1. C. A. Ferguson and C. B. Farwell, "Words and Sounds in Early Language Acquisition," Language 51, 419-439 (1975).

2. Paula Menyuk and Nancy Bernholtz, "Prosodic Features and Children's Language Production," Quarterly Progress Report No. 93, Research Laboratory of Electronics, M. I. T., April 15, 1969, pp. 216-219.