X. SPEECH COMMUNICATION*

Academic and Research Staff

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

Graduate Students

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

A. PHARYNGEAL CONSONANTS

In English all of the consonants are produced with a constriction in the oral cavity between the velum and the lips. When the constriction is located at various places within this cavity, a series of consonants with well-defined acoustic attributes is produced.[1] There are some languages for which additional consonant categories are obtained by constricting the vocal tract in the pharyngeal and uvular regions between the velum and the glottis. The purpose of this study is to examine the acoustic properties of such pharyngeal and uvular consonants in one language (Arabic) and to show that there are well-defined quantal categories that exist when a vocal-tract constriction is made in this region.

We examine, first, the acoustic effects of creating a relatively narrow constriction at some point along the pharyngeal portion of the vocal tract. In order to obtain a rough idea of the natural frequencies of the vocal tract for this situation, let us assume the vocal-tract shape shown in Fig. X-1. The cross-sectional area in front of the constriction is uniform, and equal to 3 cm^2, and the area of the portion of the vocal tract behind the constriction is 1 cm^2. The constriction is 1 cm long, and the over-all length of the tube is 17 cm. The position and cross-sectional area of the constriction can be adjusted.

The frequencies of the four lowest resonances of the configuration on Fig. X-1 are shown in Fig. X-2. The abscissa is the constriction position and the parameter is the cross-sectional area of the constriction. These natural frequencies were calculated by using a computer program developed by W. L. Henke, whose cooperation in obtaining these data is gratefully acknowledged.

When the constriction is at the extreme glottal end of the tract, such that it constitutes the first centimeter of the length of the tube, then the resonances are approximately those of the front 16 cm of the tube, and correspond approximately to odd

*This work was supported principally by the U. S. Air Force Cambridge Research Laboratories, Office of Aerospace Research under Contract F19628-69-C-0044; and in part by the National Institutes of Health (Grant 2 RO1 NB-04332-06).

†On leave from Tohoku University, Sendai, Japan.

multiples of a quarter-wavelength. The resonant frequencies are modified only slightly by the presence of the constriction. As the constriction is moved to a more anterior position, the natural frequency that represents a resonance of the acoustic mass of the constriction and the acoustic compliance behind the constriction begins to play an important role. This resonant frequency depends very much on the size of the constriction, as well as on its position.

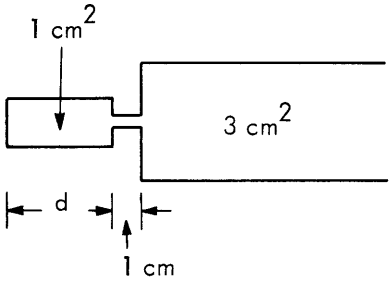


Fig. X-1

Idealized model of the vocal-tract area function used for study of pharyngeal consonants. The over-all length of the tube is 17 cm. The glottis is at the left-hand end.

Thus, for example, for a constriction size of 0.05 cm^2, this resonance is sufficiently low in frequency to be the first formant for d > 5 cm. In the limit, when the constriction size is small and d becomes large, this first-formant frequency approaches zero. This resonance is the second formant for d in the range 1-2 cm, and is probably the third formant when d is ~0.5 cm. Coupling between the various resonances makes it difficult to identify the source of any one of the resonances when two resonant frequencies are close together.

For d = 7 cm, the third and fourth resonances are approximately equal (the $\lambda/2$ resonance of the back cavity and the $3\lambda/4$ resonance of the front cavity). For this condition, F_3 is maximally high and F_4 is maximally low. In the range from d = 3 cm to d = 7 cm, F_1 and F_2 are close together, at least for constriction sizes that are not too small (0.05-0.2 cm^2). For this condition, the Helmholz resonance of the back cavity and constriction is approximately equal to the $\lambda/4$ resonance of the front section of the tube. When this condition occurs, F_1 is rather high, and F_2 is low, and for a given constriction size F_1 and F_2 are not very sensitive to changes in constriction position.

Although the curves shown in Fig. X-2 were obtained with a rather idealized vocal-tract shape, it is to be expected that the same general trends in the formant frequencies would be observed for more realistic shapes of the front and back cavities and for the constriction. In practice, of course, these area functions for the consonantal configuration are influenced by the vowel that is adjacent to the consonant.

The range of constriction positions between d = 3 cm and d = 7 cm is the region appropriate for the generation of pharyngeal consonants. One important feature of

such consonants would appear to be a high F_1 and a low F_2. The first formant is not, of course, high when there is complete closure at the constriction (as with a pharyngeal stop consonant), but except for a brief interval after release (which is presumably traversed very rapidly) the constriction size would be in a range that produces a high F_1. For a more constricted configuration, F_1 tends to be lower for d = 7 than for d = 3 cm.

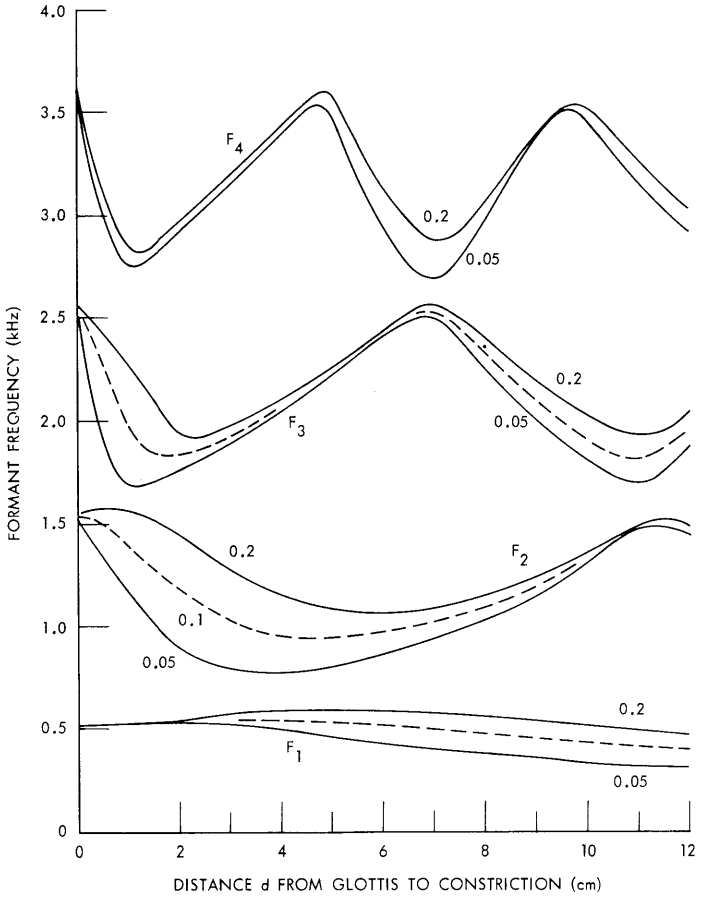


Fig. X-2. Four lowest natural frequencies of the vocal-tract shape of Fig. X-1 as a function of the length, d, of the back cavity. The parameter is the cross-sectional area of the constriction in cm^2. The dashed line corresponds to a constriction area of 0.1 cm^2.

Depending upon the specific location of the constriction in the range d = 3-7 cm, the properties of a pharyngeal consonant other than the F_1 and F_2 positions may differ appreciably. Consider, for example, the curves in Fig. X-2 corresponding to constriction sizes of 0.05 and 0.1 cm^2. For d = 3-4 cm, F_3 is considerably lower than it is for d = 7 cm, and also F_2 may be slightly lower. The third formant is a front-cavity

resonance for this more posterior constriction position, and F_2 tends to be a resonance of the back cavity. Thus, when frication noise is generated in the vicinity of the constriction, the front cavity and hence the third formant is strongly excited, but there is little excitation of the second formant. The opposite situation exists when the constriction is at the more anterior position, corresponding to d = 7 cm in Fig. X-2. Here F_2 is a front-cavity resonance, and this resonance is certainly excited by frication noise at the constriction. Furthermore, F_3 is relatively high for this constriction, and may be rather close to F_4. Thus there is a wide space between F_2 and F_3 for a consonant with a constriction at the more anterior position – at approximately d = 7 cm in Fig. X-2.

When the constriction is in a still more anterior location, around d = 11 cm in Fig. X-2, the second and third formants are close together, and F_1 is relatively low. This is the general region associated with velar consonants.[2]

In summary, then, a chart like that shown in Fig. X-2 predicts that there is a class of consonants, with a constriction well back in the vocal tract, having the distinctive property that F_1 is high and is relatively close to F_2. This property distinguishes the pharyngeal consonants from consonants with a more anterior constriction position, all of which are characterized by a low-frequency first formant. The chart further predicts that there are two subclasses within the class of pharyngeal consonants: a more posterior one with a low F_3, which is the lowest formant that is excited by noise in the case of a fricative, and a more anterior one with a high F_3, which is characterized by excitation of F_2 when there is frication noise at the constriction. The more posterior of these two constriction positions is usually called a pharyngeal consonant, and the more anterior one is a uvular consonant.

In order to verify the predictions derived from Fig. X-2, recordings of a number of consonant-vowel syllables produced by several speakers of Arabic were obtained. The syllables consisted of each of the pharyngeal, uvular and glottal stop, fricative and sonorant consonants followed by the vowels [i], [a], and [u]. The usual phonetic classification of the Arabic consonants produced in this region of the vocal tract is shown in Table X-1, and the phonetic symbols of the International Phonetic Association are shown in each case. Wide-band spectrograms

Table X-1. Phonetic classification of several Arabic consonants.

	Stop		Fricative	
	voiced	voiceless	voiced	voiceless
glottal	ʔ			h
pharyngeal			ʕ	ħ
uvular		q	ʁ	x

of all of these syllables were made, using the expanded frequency scale that encompasses the range 0-3500 Hz. Examples of the spectrograms for one of the informants, who speaks a Lebanese dialect, are displayed in Fig. X-3. Other speakers show similar general characteristics, with some variation depending on the dialect.

The two voiceless fricative consonants, which are shown in the upper row of Fig. X-3a, display precisely the attributes predictable from Fig. X-2. The fricative [ħ] is the more posterior one, and shows strong excitation of F_3, and weak or nonexistent excitation of F_2. Furthermore, F_3 is relatively low for the consonant [ħ] (for example, F_3 = 2000 Hz in the fricative preceding [a]). In contrast, F_2 is strongly excited by noise for the more anterior fricative [x], and F_3 is somewhat higher (~2300 Hz for the fricative preceding /a/). For this utterance, F_3 and F_4 appear to be rather close together in the consonant.

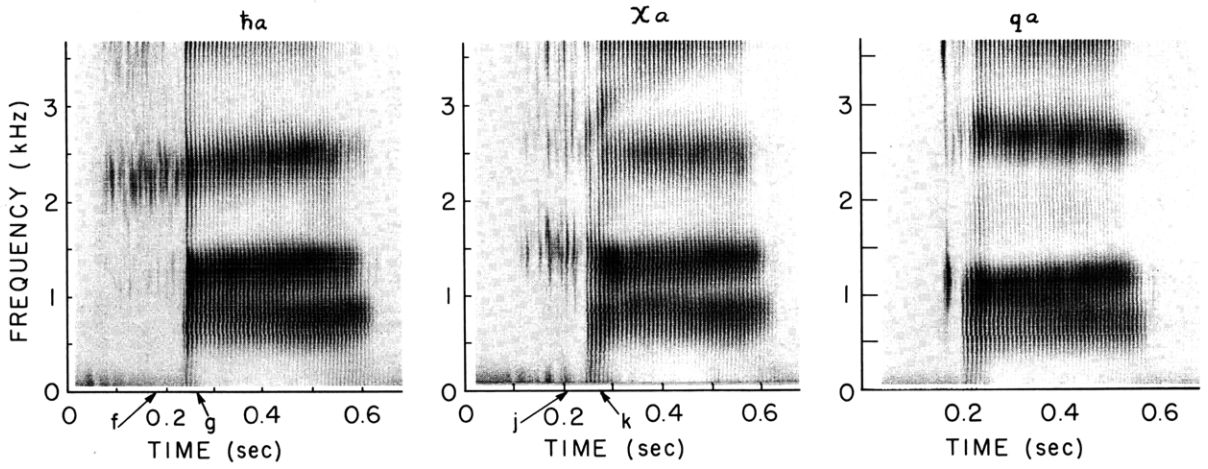
The voiced fricative consonants, shown if Fig. X-3b, demonstrate even more clearly the contrast between the lower F_3 for the posterior (pharyngeal) version and the higher F_3 for the anterior (uvular) one. Furthermore, the posterior consonant appears to have a higher F_1 – a fact that is evident for the simple model used to construct Fig. X-2 when the constriction is narrow (0.05 cm^2).

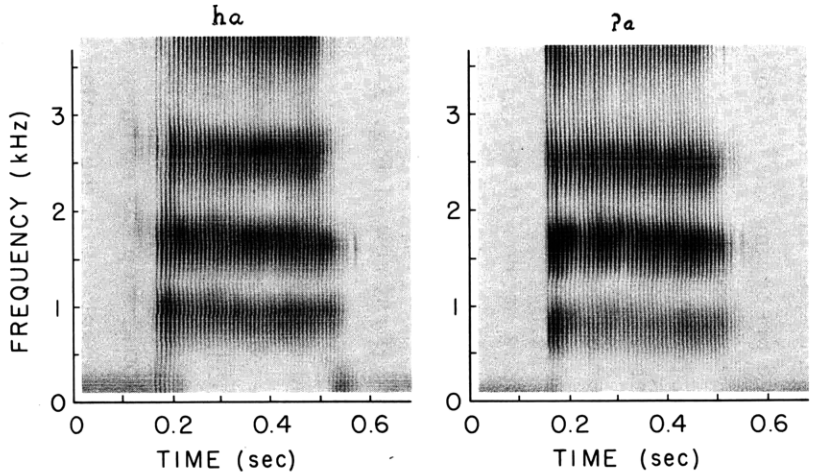
The stop consonant in this language is apparently produced with a constriction at the more anterior, or uvular, position, since the stop release is characterized by strong excitation of F_2 (first row of Fig. X-3a). Furthermore, F_3 remains relatively high at the onset of this consonant.

The /h/ and the glottal stop /ʔ/, which are shown in the second row of Fig. X-3a, differ from the pharyngeals in that there are no formant transitions at the vowel onsets. In the case of [h], there is weak noise excitation of the formants, but this is much weaker than the frication noise excitation of F_2 or F_3 for the pharyngeal and uvular fricatives. The glottal stop has no burst of energy corresponding to onset of excitation for a particular formant, as in the uvular stop.

In the spectrograms of the voiced fricative consonants in Fig. X-3b it is difficult to see good evidence of frication noise. The principal vocal-tract excitation is clearly at the glottis, and it might be argued that these consonants are sonorants rather than fricatives. Further examination of the acoustic data reveals, however, that the state of the larynx during the production of the voiced uvular and pharyngeal fricatives is not the same as in the following vowel, thereby suggesting that these consonants do not have the normal laryngeal excitation that is characteristic of sonorants.

The extent of the change in the glottal source for these consonants is shown by the sampled spectra displayed in Fig. X-4. These spectra were obtained from a 36-channel bank of simple-tuned filters (see Flanagan). The spectra were sampled at instants of time marked by the arrows on the spectrograms in Fig. X-3b.


ħa
ʕa
qa
FREQUENCY (kHz)
0
1
2
3
f
g
j
k
0.2
0.4
0.6
TIME (sec)


ha
ʔa
FREQUENCY (kHz)
0
1
2
3
0.2
0.4
0.6
TIME (sec)

(a)

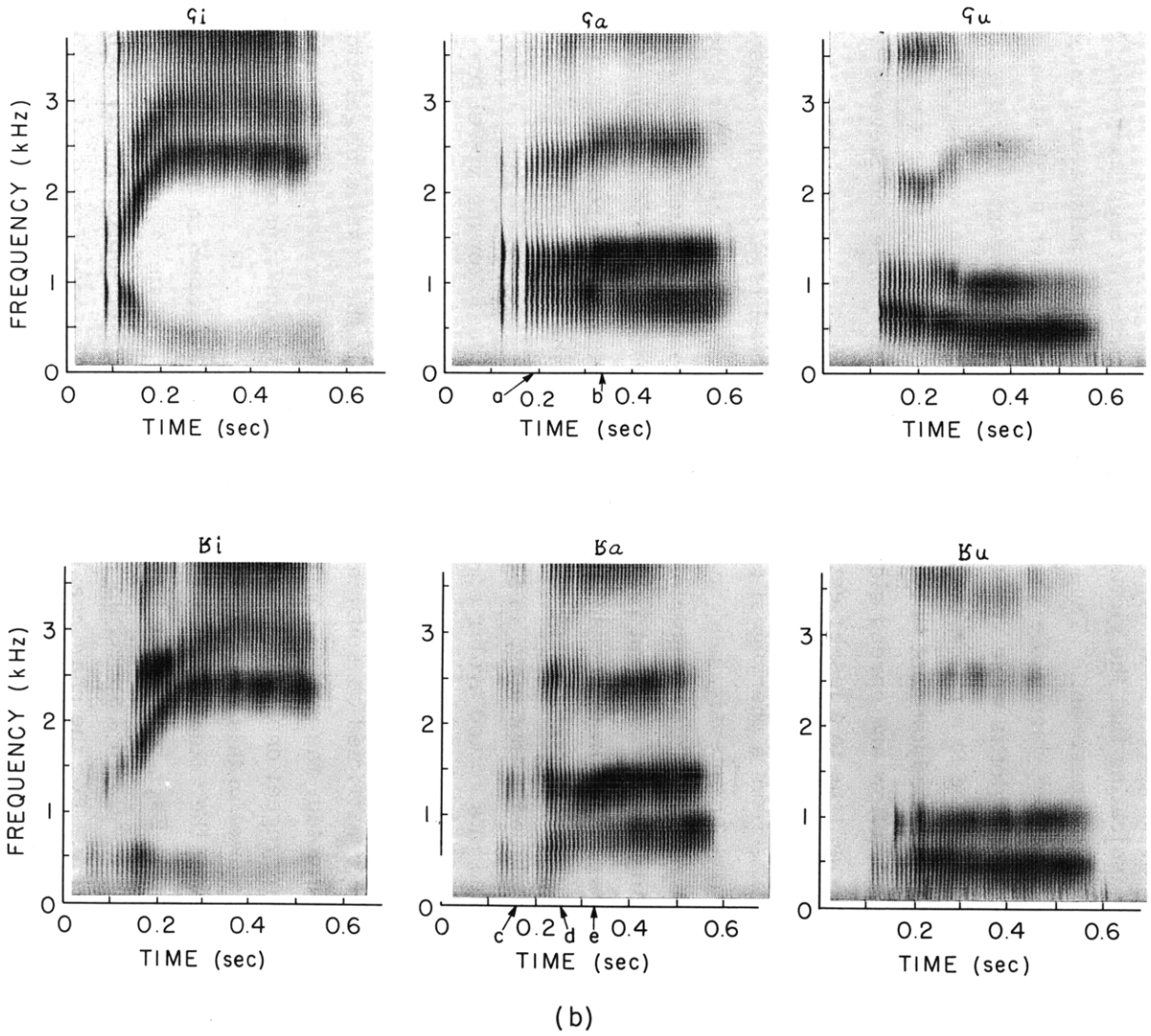


Fig. X-3. Spectrograms of several Arabic consonants in a consonant-vowel frame, produced by a speaker of Lebanese Arabic. The phonetic symbols are identified in Table X-1. The arrows on some of the spectrograms identify instants of time at which the spectra in Figs. X-4 and X-5 were sampled.

In the case of the pharyngeal consonant, the sampled spectra again show the high first formant in the consonant interval. During the transition into the vowel there is a slight downward shift in F_2, and an increase in over-all intensity which amounts to approximately 6 dB at low frequencies from spectrum sample a to sample b. The increase in spectrum amplitude is greater at high frequencies (~12 dB) than at low frequencies, thereby indicating that the glottal spectrum is richer in high frequencies for the vowel than for the consonant. Thus the individual glottal pulses are probably smoother and less peaked during the consonantal interval. Such a vibration pattern would be expected if the glottis were spread slightly for the consonant, which would give rise to a glottal resistance to air flow that is lower than in a normal vowel. There is evidence that a laryngeal adjustment is typical of voiced fricatives, i.e., a voiced fricative has increased airflow and spread glottal vibrations (see Halle and Stevens[4]). The same situation may be true to a lesser extent in sonorants.

It may be argued that the increase in amplitude of the glottal output that occurs as the articulation shifts from a pharyngeal consonant to a vowel is an essential gesture if the syllable is to have an intensity peak in the vowel. For sonorant and other voiced consonants generated with a more anterior constriction position, (such as [w] and [y]), the reduced intensity of the consonant relative to an adjacent vowel is an automatic consequence of the lowered F_1 in the consonant. It is not necessary to postulate a reduced glottal output for these consonants in order to explain the reduced acoustic intensity (although such a reduced output may, in fact, occur).

Spectra sampled in a voiced uvular consonant and in the following vowel are also shown in Fig. X-4. The instants at which spectra are sampled are indicated in the spectrogram of [ʁa] in Fig. X-3b. Again there is a substantial increase in intensity as the transition from the constricted consonantal configuration to the vowel configuration is made. As we have noted, the uvular consonant tends to have a lower first-formant frequency than the pharyngeal one. There is a slight upward transition of F_1 between the consonant and the vowel in this example.

The properties of the voiceless fricative consonants are shown in the form of sampled spectra in Fig. X-5. Each part of the figure displays a spectrum sampled in the voiceless consonantal interval and a spectrum sampled in the vowel approximately 30 msec after onset of voicing. For the pharyngeal consonant, the predominance of the third formant in the voiceless interval is quite evident. The major spectral peak in the uvular consonant corresponds to the second formant. It should be noted that all of the spectra shown in Figs. X-4 and X-5 were obtained with a rising frequency characteristic of 6 dB/octave in the analyzer preceding the filters. If this characteristic is taken into account, the data of Fig. X-5 suggest that the intensity of the uvular [x] is somewhat greater than that of its pharyngeal counterpart, at least in these examples.

These observations with regard to the pharyngeal and uvular consonants provide

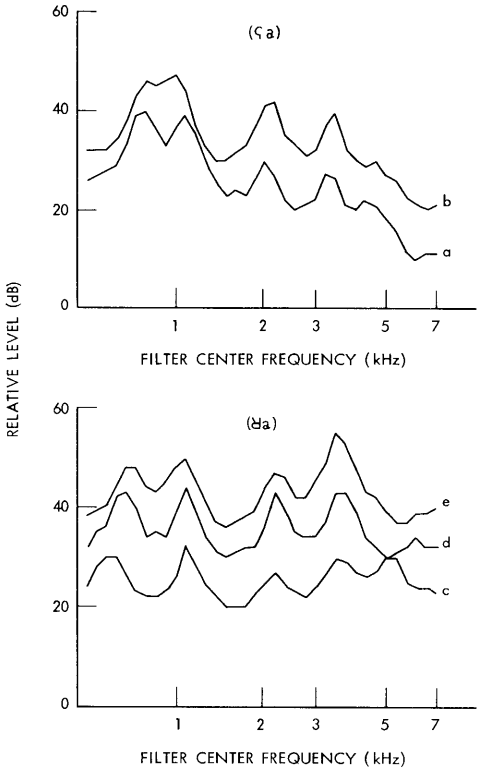


Fig. X-4.

Spectra sampled in syllables containing voiced pharyngeal fricatives (upper curves) and voiced uvular fricatives (lower curves). The letters labeling the spectra identify instants of time (in Fig. X-3) at which spectra were sampled. Spectra were obtained with a 36-channel bank of filters.

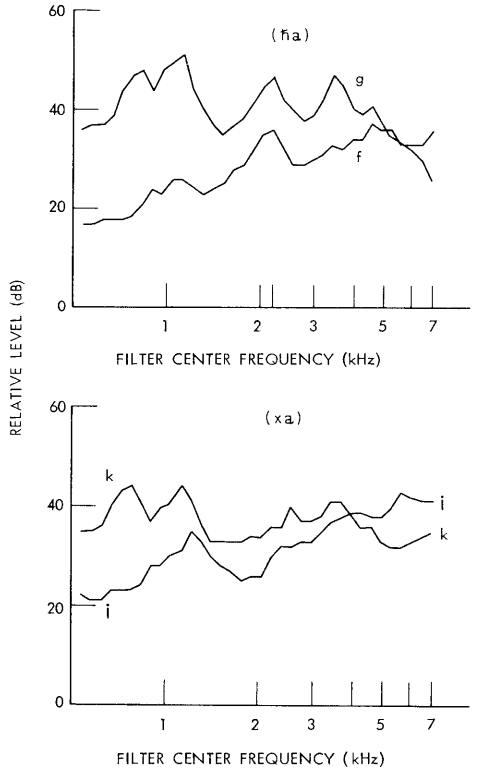


Fig. X-5.

Same as Fig. X-4, but for voiceless pharyngeal and uvular fricatives.

support for a view that the acoustic attributes associated with the various articulatory configurations and maneuvers that are found in speech are basically discrete or quantal.[2] That is, there are particular ranges of articulatory activity that give rise to acoustic outputs with well-defined attributes. These attributes are relatively insensitive to perturbations of the articulation within these ranges, thereby suggesting that the selection of these particular articulations for use in language imposes less severe restrictions on the precision with which the articulations must be effected.

In the case of the pharyngeal region of the vocal tract, it would appear that there are two well-defined places of articulation with distinctive acoustic properties. While these acoustic properties appear to be relatively insensitive to at least small perturbations in place of articulation, they may be somewhat sensitive to the degree of vocal-tract constriction, but this point needs further study.

D. H. Klatt, K. N. Stevens

References

1. K. N. Stevens, "Acoustic Correlates of Place of Articulation for Stop and Fricative Consonants," Quarterly Progress Report No. 89, Research Laboratory of Electronics, M.I.T., April 15, 1968, pp. 119-205.
2. K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in E. E. David, Jr. and P. B. Denes (eds.), Human Communication: A Unified View (McGraw-Hill Publishing Company, New York, in press).
3. J. L. Flanagan, "A Speech Analyzer for a Formant-Coding Compression System," Sc.D. Thesis, Department of Electrical Engineering, M.I.T., 1955 (unpublished).
4. M. Halle and K. N. Stevens, "On the Mechanism of Glottal Vibration for Vowels and Consonants," Quarterly Progress Report No. 85, Research Laboratory of Electronics, M.I.T., April 15, 1967, pp. 267-271.

B. PROSODIC FEATURES AND CHILDREN'S LANGUAGE PRODUCTION

1. Problem

It has been observed that there is a period in the child's acquisition of language during which he primarily produces one-word utterances, although, at this same stage, he may also be producing babbled utterances exceeding 3 or 4 syllables in length and, on occasion, 2-word utterances. The communicative function of this period of language acquisition can be and has been viewed as (i) to name objects and events, and (ii) to express a whole sentence with a single word. Those holding the latter view have called this period the holophrastic stage.[1] There is very little evidence to corroborate either view. An utterance such as "moo" could be interpreted as the child's name for milk, since he uses it consistently in the presence of milk or as a sentence either demanding

milk or making a statement about the fact that it is milk or asking a question about the milk.

It has been hypothesized that the prosodic features of the utterance (intonation and stress) can be an indication of whether these single words are names or sentences.[2] If all such utterances terminate in the falling fundamental frequency contour of a declarative statement, then there is evidence that such utterances are names. If prosodic features are used generatively with these words to indicate different meanings, then there is some evidence that such utterances are used as sentences. Thus, depending on the prosodic features of the utterance, the utterance "moo.." could be the declarative statement "(That's) milk."; the utterance "moo?" could be the question "(Is that) milk?" or "(May I have) milk?"; and the utterance "moo!" could be the imperative "(I want) milk!" or "(Give me) milk!"

There are alternative speculations about the structure of the language used during this period. Prosodic features might be used imitatively, that is, without any underlying syntactic structure, but merely as a repetition of adult utterances. It has been stated that infants imitate the intonation of their mother's utterances during the babbling period[3] and, therefore, this is certainly a possibility. It is also possible that certain words are always produced with the same intonational and stress contour (for example, "No!" always being produced as an emphatic). This would be an indication that prosodic features are not being used generatively to indicate differing syntactic structure, but are only some other phonological aspects of a particular word.

The study undertaken here was a preliminary attempt to resolve some of the questions concerning the structure of the language used during this period of language development by an analysis of the prosodic features of a child's utterances during this period.

2. Procedure

The recorded utterances of a child at the stage at which he was producing primarily one-word utterances (age 18-20 months) were examined, and a series of utterances that were repetitions of the same word were isolated and re-recorded. These series consisted of the words "no," "door," "touch," and "up," and two series of names, one with and one without possessive markers (for example, "daddy" and "daddy's" and "Jeanie" and "Jeanie's"). None of these utterances were immediate repetitions of the mother's utterances but, rather, were introduced by the child into the conversation.

Two listeners attempted to classify the isolated utterances as declaratives, questions, and emphatics. According to the listeners, each series contained the three types of classification, except for the name series without possessive markers which had no emphatics. There was 81% agreement between the two listeners on the categorizations of the utterances. Spectrograms were made of the utterances and it was observed that,

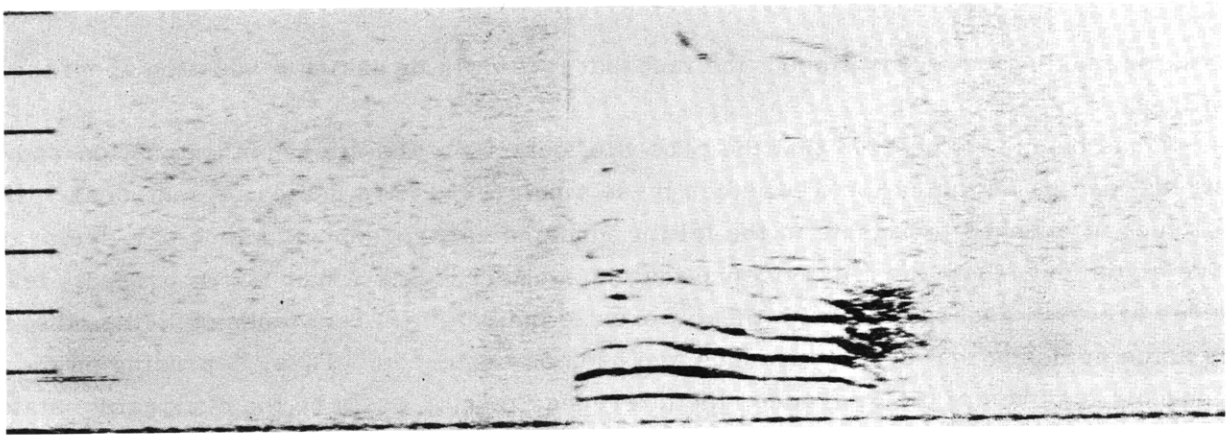


Fig. X-6a. Statement "door."

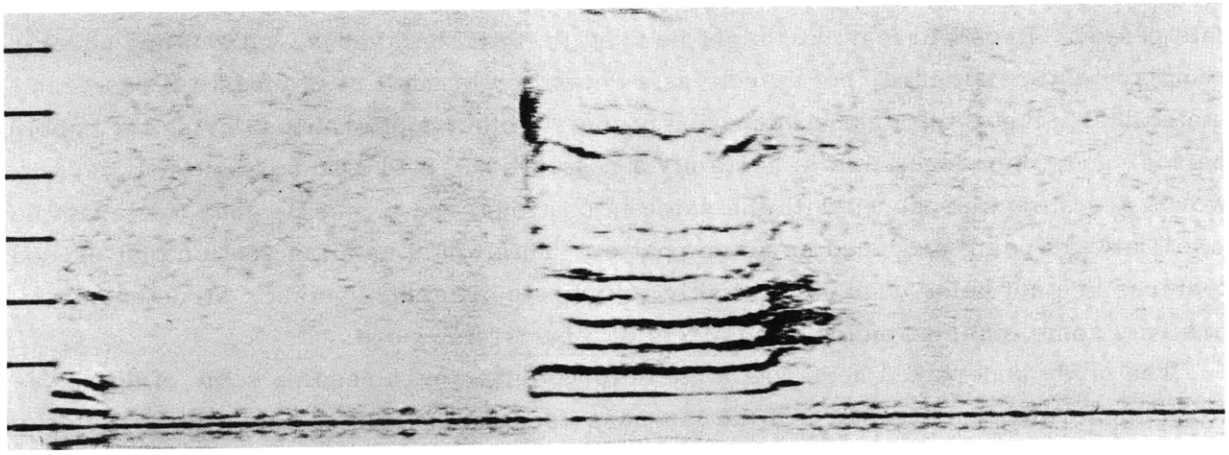


Fig. X-6b. Question "door."

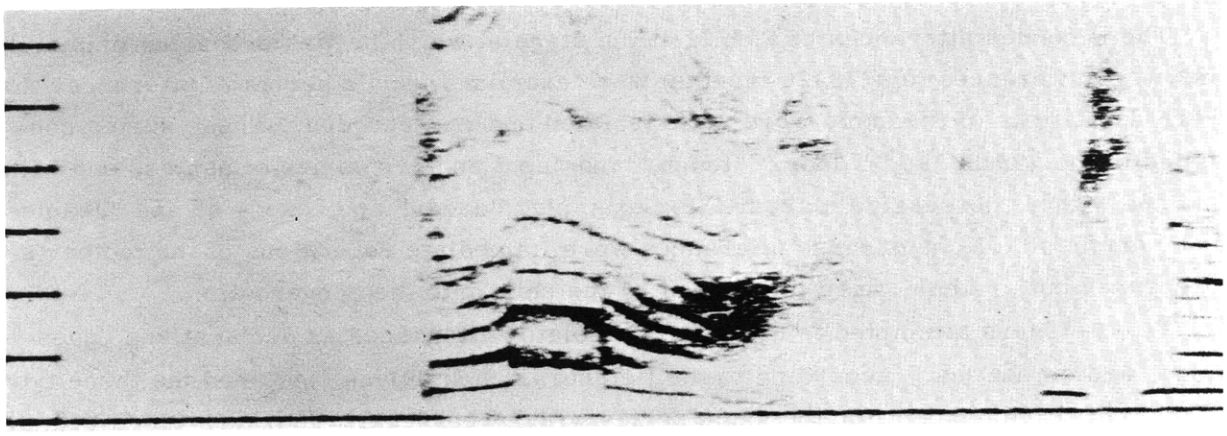


Fig. X-6c. Emphatic "door."

although there were individual variations within each category, a general characteristic of each type of utterance could be found. Declarative utterances terminated with a falling fundamental frequency contour, questions terminated with a rising fundamental frequency contour, and emphatics had a sharp rise and then fell in fundamental frequency contour during the utterance. Figure X-6 contains spectrograms of the utterance "door" (/dɔə/) as (a) a statement, (b) a question, and (c) an emphatic according to the classification of both listeners. Table X-2 presents measurements of the spectrograms.

Table X-2. Measurements of spectrograms of the utterance "door."

		Fundamental Frequency		
Type	Total Length	Beginning	Peak	End
(a) Statement	.60 sec	350	426	213
(b) Question	.53 sec	382	560	560
(c) Emphatic	.72 sec	455	1010	255

Measurements of the total set of spectrograms are now being carried out to determine if the pattern observed for each type of utterance is indeed consistent and to determine how well these patterns correlate with perceptual categorizations. Analyses of the utterances produced by two other children during this period of development is planned to determine how consistent this performance is across children. In addition, analyses of mothers' responses to these types of utterances will be carried out to see if these responses contain evidence that some differentiation of the utterances, in accordance with their acoustic and perceptual classifications, is being made by mothers.

Although the data are extremely limited there appear to be indications that the child's single word utterances are not simply names of objects and events and that the child uses prosodic features generatively to create sentence types rather than merely imitating prosodic features or including these features as part of the speech sound composition of a particular word.

Paula Menyuk, Nancy Bernholtz

References

1. D. McCarthy, "Language Development in Children," in L. Carmichael (ed.), Manual of Child Phychology (John Wiley and Sons, Inc., New York, 1954), pp. 492-630.

2. Paula Menyuk, Sentences Children Use (The M.I.T. Press, Cambridge, Mass., 1969), Chap. 2.

3. S. Nakazima, "A Comparative Study of the Speech Developments of Japanese and American Children," Studia Phonologica II, pp. 27-39, 1962.

C. VOWEL IDENTIFICATION AND DISCRIMINATION IN ISOLATION vs WORD CONTEXT

1. Introduction

Previous experiments in phoneme perception have been concerned with identification and discrimination along a synthetic speech continuum whose end points sound like two particular phonemes. An example will serve to describe the main results of these experiments. In a recent study Stevens and co-workers[1] examined, in part, perception along a speech continuum from /dɛ/ to /gɛ/ by varying the starting location of the second-formant center frequency in approximately 7 equal frequency steps. If the first stimulus sounded like /dɛ/, and the seventh stimulus sounded like /gɛ/, how were intermediate stimuli perceived? The observed effect was quantal, in that only two distinct categories of sounds (/dɛ/ and /gɛ/) could be heard over the whole continuum, and the change from hearing /dɛ/ to hearing /gɛ/ occurred over a range of only one or two stimuli. This effect is known as the phoneme boundary, and can be measured by examining the extent to which two stimuli from the same category are confused with each other, compared with two stimuli drawn from different categories.

Specifically, by measuring at each region of the continuum some "index of confusion" between two adjacent stimuli in that region, a function can be obtained. A sketch of this function is shown in Fig. X-7a for the /dɛ -gɛ/ example. If a larger index corresponds to a lesser degree of confusion between the adjacent stimuli involved, then this figure

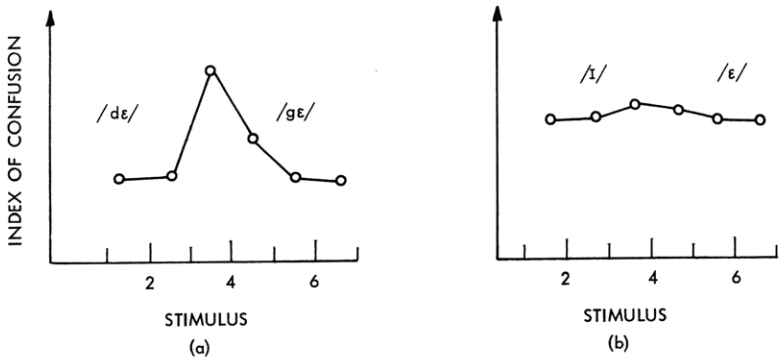


Fig. X-7. Index of confusion vs adjacent stimulus pair. A sketch for two situations: (a) /dɛ -gɛ/ continuum; (b) /ɪ-ɛ/ continuum.

shows the presence of a phoneme boundary: A small amount of confusion occurs between stimuli 3 and 4, compared with confusion between, for example, stimuli 1 and 2, or stimuli 6 and 7.

Similar experiments were performed[1] using isolated vowel stimuli (/ɪ/ to /ɛ/ for

example), and the phoneme boundary effect was much less apparent, if apparent at all, as shown in Fig. X-7b. The difference in these two results can be explained if we compare the two sets of stimuli. For the consonants, stimulus formant frequencies are changing rapidly with time as they approach a steady state for the following vowel; for the isolated vowels, the formant frequencies are already in a steady state.

If this difference in stimuli is the reason for a presence of the phoneme boundary in one case but not the other, then would such a boundary be present for vowels in consonantal or word context? The dynamic nature of vowel formant transitions caused by surrounding consonants[2] would seem to imply the presence of a phoneme boundary, and indeed, observations by Stevens[3] support this view. The present study answers the question quantitatively for the particular vowel pair /ɑ-æ/ in isolation vs the word context bottle-battle (written phonetically as /bɑdəl-bædəl/).

Two different methods were used by the author to measure the degree of confusion in a particular stimulus region, the chief method being the use of a standard Absolute Identification (A.I.) paradigm. If there are n stimuli along a given continuum, subjects are allowed to use one of n responses (a number from 1 to n). The resulting Stimulus-Response confusion matrix contains enough information so that the degree of confusion between any two stimuli can be measured by examining the distribution of responses for these two stimuli. In this case, the index of confusion is d'.[4] A plot of d' against adjacent stimulus pair is defined as an identification function for this A. I. experiment.

Another method of measuring confusions involves discrimination ability in a particular stimulus region independent of that ability in other regions. In other words, this method tests a subject's ability to discriminate between two sounds in one particular stimulus region during one complete experiment. The discrimination index can be measured by examining the information in the 2 × 2 Stimulus-Response matrix from a Two-Interval Two-Alternative Forced-Choice (2AFC) paradigm (the index obtained is d', just as in the A. I. paradigm). The d' values for several stimulus regions can be collected on one graph to form a discrimination function (d' vs stimulus region) for a given continuum. A comparison of data from an A. I. paradigm and a corresponding ZAFC test will be described.

A summary of the basic experiments in this study follows.

1. Using A.I., measure the identification function for the /bɑdəl-bædəl/ continuum.

2. Using A.I., measure the identification function for the isolated vowel continuum /ɑ-æ/ and compare with the results of Experiment No. 1.

3. Using 2AFC, measure an approximate discrimination function for the /bɑdəl-bædəl/ continuum and compare with the results of Experiment No. 1.

2. Procedure: Absolute Identification

Spectrograms of the end points of the two vowel continua (/bɑdəl-bædəl/ and /ɑ-æ/) are shown in Fig. X-8. Two durations of the same isolated vowel stimuli were used,[5] one set lasting approximately 150 msec (approximately the duration of the vowel in word context), and another set lasting 250 msec. Thus, three A.I. experiments were performed corresponding to 3 stimulus continua: (i) /bɑdəl-bædəl/; (ii) short duration /ɑ-æ/; and (iii) long duration /ɑ-æ/. By using a synthesis-by-rule computer program to generate the stimuli,[6] 8 stimuli were selected along each continuum so that adjacent stimuli corresponded to equal logarithmic increments of the second-formant center frequency, that is,

$$\Delta f_2/f_2 = \text{constant for any two adjacent stimuli.}$$

For stimulus 1 (bɑdəl or ɑ), f_2 = 1000 Hz; for stimulus 8 (bædel or æ), f_2 = 1426 Hz.

In each experiment, 3 subjects were presented with a random sequence of stimuli — one per trial — in an A.I. format, and were asked to identify the stimulus they heard by a number, one through eight. The stimuli were played from a tape recorder and presented binaurally through headphones in a quiet listening room. The responses were recorded automatically by having the subjects pressing one of eight buttons on a response box. Feedback was given by flashing the correct number on a row of lights located in the front of the listening booth. The entire experiment was automated, each trial lasting approximately 5 1/2 sec, with feedback given during the last 2 1/2 sec of the trial. Subjects could make a response any time during the initial 5 sec of each trial.

Eight runs of 64 trials each were presented to each subject during an experimental session (so that each stimulus was heard 64 times), and the responses from these 8×64 stimuli were analyzed from the resulting Stimulus-Response confusion matrix. The distribution of responses for each adjacent pair of stimuli was compared to obtain several estimates of an "identification index" for this stimulus pair. (The identification index was d'.) The index estimates for each stimulus pair in the matrix were combined with corresponding estimates obtained from other experimental sessions in order to obtain a sample mean and standard deviation for the identification index of each adjacent stimulus pair. Each subject was exposed to 4 experimental sessions using the same stimulus set, the data from the last three sessions being used to obtain the averaged results. Examination of results from each session showed that the first session usually gave sufficient time for the subjects to learn the task.

3. Procedure: 2AFC Discrimination

A subsidiary 2AFC discrimination experiment was performed, with the /bɑdæl-bædɔl/ continuum used. In order to produce significant response errors, the two

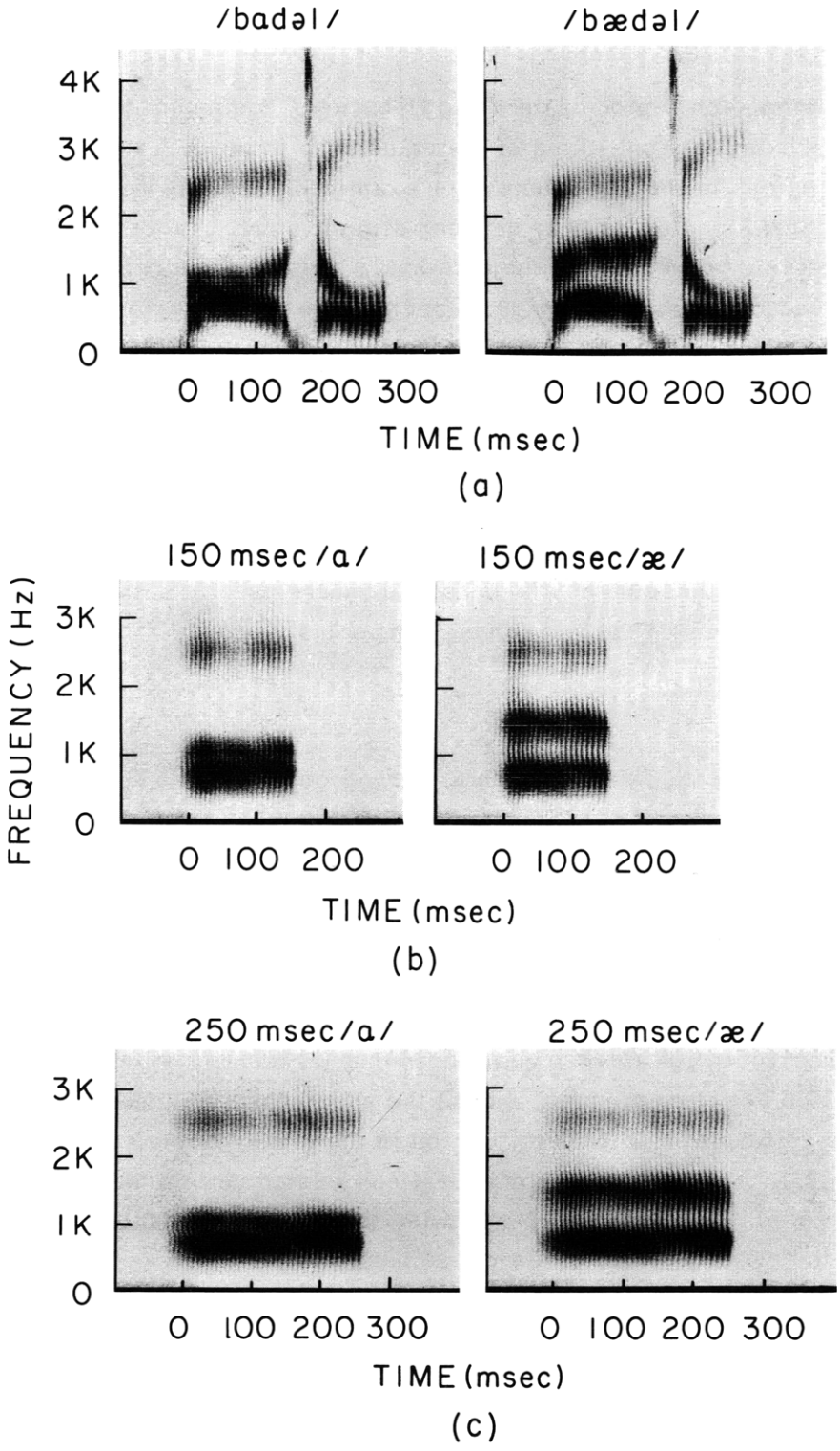


Fig. X-8. Spectrograms of synthetic speech stimuli.
(a) First and eighth stimuli for the /bɑdəl-bædəl/ experiment.
(b) First and eighth stimuli for the 150-msec /ɑ-æ/ experiment.
(c) First and eighth stimuli for the 250-msec /ɑ-æ/ experiment.

stimuli taken from each region of the /bɑdəl-bædəl/ continuum were such that the ratio $\Delta f_2/f_2$ was approximately one-third of the value used in the A.I. test. Discrimination at 3 selected regions in the continuum was examined (corresponding to regions about stimulus 2, 7, and a region halfway between stimuli 4 and 5 in the A.I. paradigm).

In this subsidiary experiment, subjects heard the two stimuli per trial (with the two stimuli separated by approximately 100 msec), and were asked to press one of two buttons to indicate which stimulus sounded more like "battle." Each subject was immediately informed by one of two feedback lights on his response box whether his decision was right or wrong. During each experimental session, only one stimulus region was tested. Eight runs of 100 trials/run were given during the session. An estimate of the discrimination index (d') for each set of 100 trials was obtained for each subject. In order to ignore some of the effects of training and fatigue, only the three best estimates (corresponding to the three highest scores in d') were used to obtain a sample mean and standard deviation for each subject for each stimulus region.

4. Results

The results of the A.I. experiments were plotted as $d'_j \pm \sigma_s(d'_j)$ against adjacent stimulus pair, where d'_j indicates the indentification index estimate for the stimulus pair $\{j, j+1\}$, and $\sigma_s(d'_j)$ is the sample standard deviation for that estimate.[7] Figure X-9 shows three such graphs for the three experiments, the subjects being D.G. and R.S. (the author). The third subject produced results that were similar in some respects.[8] Three main conclusions can be drawn from Fig. X-9.

1. There is a noticeable difference in the A.I. response between long-duration /ɑ-æ/ and short-duration /ɑ-æ/ vowels, in that (a) identification is much better along the long-duration vowel continuum; and (b) the short-duration vowels are more characterized by a significant peak in identification in the center of the continuum.

2. The A.I. response to the short-duration isolated vowels is very similar to the word context vowels /bɑdəl-bædəl/; that is, there is a noticeable central peak in identification for the word context vowels whose height is the same as in the isolated case.

3. Using the method of averaging noted above, we found that the average standard deviation in d' was $\sigma_s(d'_j) = 0.3$ units. To a first approximation, this value was independent of both the value of d'_j and the subject.

The results of the 2AFC experiment were plotted as $d'_j \pm \sigma_s(d'_j)$ against stimulus region, where d'_j represents the discrimination index for the j^{th} stimulus region. Figure X-10 shows graphs obtained for subjects D.G., J.B., and R.S. The main conclusion drawn from comparison of Figs. X-9 and X-10 is the following.

4. There is a noticeable difference in the shape of the d' vs stimulus pair function for the word context vowels, depending on whether the A.I. test (over the whole

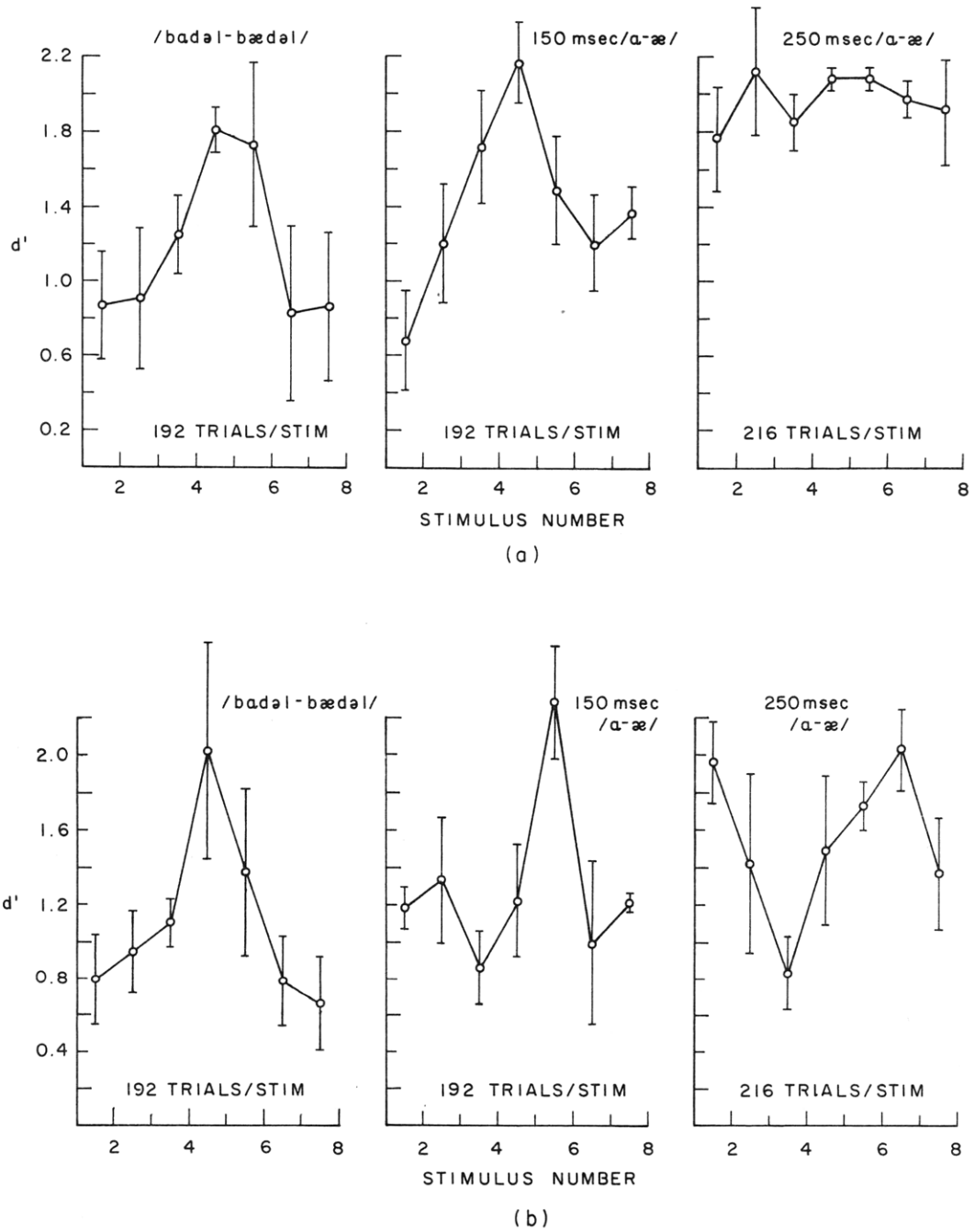


Fig. X-9. $d'_j \pm \sigma_s(d'_j)$ vs adjacent stimulus pair. Summary of three absolute identification experiments. (a) Subject D. G. (b) Subject R. S.

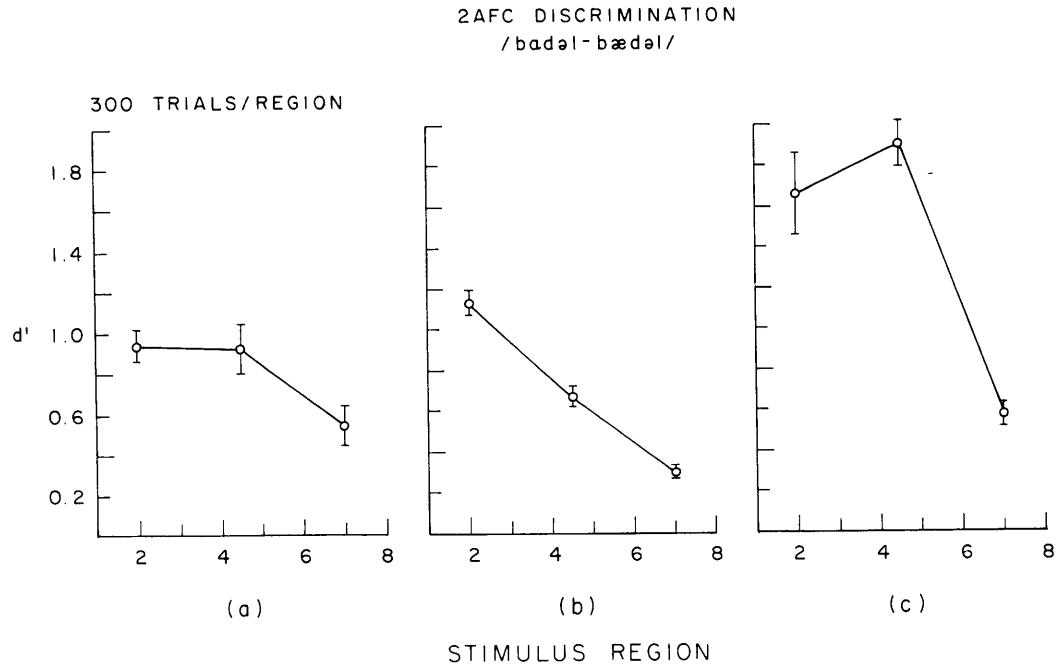


Fig. X-10. $d'_j \pm \sigma_s(d'_j)$ vs stimulus region. Summary of the two-alternative forced-choice experiments using /bɑdəl-bædəl/ stimuli. Average data of the three best 100 trial runs for each stimulus region. (a) Subject D. G. (b) Subject J. B. (c) Subject R. S.

continuum) or the 2AFC test at specific points on the continuum is used. While both low and high numbered stimulus pairs produced low d' scores (with respect to center pairs) in the A. I. test, the 2AFC test produced lower d' scores only in the high stimulus region (corresponding to /bædəl/).

5. Discussion of A. I. Results

The name "phoneme boundary," used by other investigators to denote the presence of a sharp peak in identification between two phoneme regions, seems appropriate. The presence of a phoneme boundary, then, indicates the tendency of a subject to categorize a given continuum in two parts: the narrower and higher the peak, the more pronounced the categorization. Thus, one can re-explain the major result of these experiments in the following generalization: The shorter the duration of a vowel, the more it is perceived categorially. Whether a vowel of short duration is produced in isolation or in word context, the listener is more likely to perceive the vowel in terms of binary features than with slowly spoken vowels. Of course, a vowel spoken rapidly in isolation is very unnatural, so that in fact this same generalization has been hypothesized by other observers only in contrasting long-duration isolated vowels with vowels in word or

consonantal context. For example, Stevens[3] found phoneme boundaries to be much more pronounced when subjects were exposed to vowels in word context (bil-bɪl-bɛl) than when exposed to the same vowels in long-duration isolation (i-ɪ-ɛ).

6. Identification vs Discrimination

Experiments by other investigators to measure the presence of phoneme boundaries (for both vowels and consonants) have followed a noticeably different methodology than the author's, the most striking difference being that other investigators have used a Roving Standard, Roving Increment ABX test to measure extent of pairwise differentiation of stimuli.[9] Subjects were required to match the third stimulus in a triad with either of the first two members, and it was argued that this ABX test measured, on a relative scale, the ability of subjects to make pairwise discrimination on any basis whatsoever. The index of "discrimination" was the percentage of times that a correct assignment was made as a function of stimulus pair for a given stimulus pair separation.

If, in fact, discrimination was being measured and the phoneme boundary was still observed, then the author would have a difficult time explaining the difference between the A.I. and 2AFC results of this experiment. There is strong reason, however, to believe that the complexity of this kind of ABX test made it give results very similar to an A.I. experiment. The roving standard and roving increment aspects of this experiment were obstacles in keeping a subject from directly comparing X with A and B, and these factors necessarily required a subject to keep the entire continuum in his mind in order to successfully perform the task. Because of the roving stimuli, short-term memory is very much involved.[10] And because of short-term memory, the results of Wickelgren[11] are important: that these speech sounds are probably first stored as a sequence of distinctive phonetic features. Thus, a label is probably attached to each member of the triad before any comparison can be made.

It can therefore be argued that discrimination was not being adequately measured in the ABX test, and that the 2AFC results of the author are a better indication of the effects of using speech stimuli in a discrimination test.[12] The difference between results of the A.I. and 2AFC tests (see Observation 4 in Results) might thus be interpreted as being due to a "memory noise,"[13] which, in turn, is probably due to linguistic experience. That is, the author has assumed that, unlike the A.I. test, the 2AFC test was so designed that to interpret the stimuli as speech would have had little effect on performance: the two stimuli were so close together as to sound, effectively, like identical phonemic sequences.

Why then did discrimination results show poor performance only in the /bædəl/

region? While these results are difficult to explain in terms of the use of speech stimuli, a close look at the detailed changes in the physical stimuli as they affect the hearing mechanism have provided one possible explanation.[14]

R. M. Sachs

Footnotes and References

1. K. N. Stevens, A. M. Liberman, M. Studdert-Kennedy, and S. E. G. Öhman, "Crosslanguage Study of Vowel Perception," (to appear in Language & Speech).
2. B. E. F. Lindblom, "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. 35, 1773-1781 (1963).
3. K. N. Stevens, "On the Relations between Speech Movements and Speech Perception," Proc. XVIII International Congress of Psychology, Moscow, U.S.S.R., August 1966.
4. d' is a normalized statistic reflecting the degree of confusion between two stimuli (see Durlach and Braida[13]). d' = 0 corresponds to complete confusion, and the larger the value of d', the greater the ability to differentiate between the two stimuli. d' can also be negative, corresponding to a confusion as to the ordering of the stimuli.
5. An initial experiment employed the 150-msec isolated vowels, and when the results obtained were not as expected (see Observation 4 in Results), a longer-duration vowel continuum was employed.
6. In other words, a word could be synthesized by typing in its phonetic spelling. See J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech Synthesis by Rule," Language & Speech 1, 127 (1964) for a good general discussion of synthesis-by-rule programs. The specific program used here was written by D. H. Klatt for a software synthesizer. See R. M. Sachs[15] for a description of stimulus generation for this experiment.
7. More precisely,

   d'_{ij} = i^{th} estimate of d'_j

   $d'_j = \frac{1}{n}\sum_{i=1}^{n} d'_{ij}$ = Sample average of the identification index for the stimulus set j, j+1. The value of n depends on the number of measurements taken.

   $\sigma_s(d'_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d'_{ij}-d'_j)^2}$ = Sample standard deviation of d'_j for n estimates.
8. The results of the third subject (J. B.) for the A. I. task were not presented because of an over-all low level of performance. For this subject, a slight peak in identification occurred in the center of all stimulus continua. Justification for not presenting the A. I. results of this subject is given in Sachs.[15]
9. See, for example, A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith, "The Discrimination of Speech Sounds within and across Phoneme Boundaries," J. Exptl. Psychol. 54, 358-368 (1957) and K. N. Stevens et al.,[1] as well as other publications on phoneme boundary experiments by the Haskins Laboratory. The task of responding to this ABX test is made difficult by making any region of the stimulus continuum equally likely to occur during a run (roving standard) and by varying the stimulus separation between A and B (roving increment) during the run.

10. J. E. Berliner, "Short-term Memory in Intensity Discrimination of Tones," S. M. Thesis, M. I. T., June 1968.

11. See W. A. Wickelgren, "Distinctive Features and Errors in S. T. M. for English Vowels," J. Acoust. Soc. Am. 38, 583-588 (1965) or "Distinctive Features and Errors in S. T. M. for English Consonants," J. Acoust. Soc. Am. 39, 388-398 (1966).

12. Of course, there is no reason to expect any generalization from the 2AFC results. They merely reflect the ability to discriminate along this particular (bɑdəl-bædəl) second-formant continuum.

13. N. I. Durlach and L. D. Braida, "Psychophysics of Intensity Resolution," Quarterly Progress Report No. 91, Research Laboratory of Electronics, M. I. T., October 15, 1968, pp. 240-249.

14. Such an explanation has been given by Sachs,[15] by examining the relationship between the fixed first formant (f1) and the changing second formant (see Fig. 2(a) or 2(b)), and arguing that detecting small changes in f2 would be more successful if f1 were closer to f2.

15. R. M. Sachs, "Vowel Identification and Discrimination in Isolation vs Word Content," S. M. Thesis, M. I. T., February 1969.