XVI. SPEECH COMMUNICATION

Academic and Research Staff

Prof. Kenneth N. Stevens
Prof. Morris Halle
Dr. Sheila E. Blumstein[*]
Dr. Margaret Bullowa
Dr. William L. Henke
Dr. A. W. F. Huggins
Dr. Dennis H. Klatt
Dr. Martha Laferriere[†]
Dr. Paula Menyuk[‡]
Dr. Colin Painter[**]
Dr. Joseph S. Perkell
Dr. David B. Pisoni[††]
Dr. Stefanie Shattuck-Hufnagel
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 Goldstein
Stephen K. Holford
Shinji Maeda
Arafim Ordubadi
Edwin S. Rich, Jr.
Victor W. Zue

1. SPEECH PRODUCTION AND MODELING

National Institutes of Health (Grant 5 RO1 NS04332-13 and
Fellowship 1 F22 MH5825-01)

Dennis H. Klatt, Stefanie Shattuck-Hufnagel, Joseph S. Perkell

a. Physiological Studies

Preparations have been made to begin taking simultaneous electromyographic and cineradiographic data for use in conjunction with a dynamic model of the tongue.[1] These preparations have included running some preliminary audio and cinesynchronization studies on a facility at a nearby hospital and obtaining the assistance of an outside laboratory in taking the EMG data. A preliminary experiment on the short-latency role of feedback in the control of articulation has also been run. The results of this experiment are described in Part II, Section XVI-A. Research plans for the forthcoming year are expected to include continued physiological studies in the areas covered by physiological modeling and in the short-latency role of feedback. Simultaneous cineradiographic and electromyographic data will be taken to explore the relationship between muscle activity and tongue movement in conjunction with the use of the tongue model, and the feedback paradigm will be extended to include electromyographic and movement measurements.

In a related study of speech production, we are examining the acoustic characteristics of vowels with distorted jaw position and with reduced tactile feedback on dental surfaces. In these experiments we are attempting to determine what strategies a speaker uses to achieve an apparently invariant acoustic target in the face of voluntary or imposed distortion of jaw position.

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

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

[‡] Professor of Special Education, Boston University.

[**] Associate Professor of Communicative Disorders, Emerson College.

[††] Associate Professor, on leave from Department of Psychology, Indiana University.

b. Speech Synthesis by Rule

A synthesis-by-rule program can be thought of as a functional model of human sentence production.[2,3] The phonological component of this model accepts as input a linear string of symbols that have been produced by semantic, syntactic, and lexical components of a grammar of English. This abstract representation of an utterance is transformed by the phonological component into a phonetic transcription and a specification of the fundamental-frequency contour, segmental durations, and stress levels. During the past year we proposed a structure for a phonological component and a way to represent information in a computer simulation of a set of phonological rules. Some rules using this notation have been implemented within this framework, and a more complete set of rules is being developed.

References

1. J. S. Perkell, "A Physiologically Oriented Model of Tongue Activity in Speech Production," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., September 1974.
2. D. H. Klatt, "The Phonological Rule Component of a Speech Understanding System" (submitted to IEEE Trans., Vol. ASSP).
3. D. H. Klatt, "Speech Synthesis from an Abstract Linguistic Description," a paper to be presented at the 4th Annual New England Bioengineering Conference, Yale University, New Haven, May 7-8, 1976.

2. LARYNX MECHANISMS AND FUNDAMENTAL-FREQUENCY VARIATIONS IN SPEECH

National Institutes of Health (Grant 5 RO1 NS04332-13)

Bertrand Delgutte, Morris Halle, Shinji Maeda, Kenneth N. Stevens, Jacqueline Vaissière

Our research on laryngeal mechanisms and fundamental-frequency variations in speech is aimed toward the formulation of a more complete theory that describes the states of operation of the larynx during speech, and specifies how sequences of these states are utilized to produce distinctive acoustic outputs. An aspect of this research is an examination of the physical processes involved in vocal-fold vibration. This part of the work included completion of a doctoral thesis[1] describing in detail the vibration patterns and other aspects of the behavior of excised larynges, a Master's thesis[2] that was a study of the excitation of waves on a membrane by rapid airflow along the membrane surface, and further theoretical investigation of the energy exchange between aerodynamic and mechanical processes during vocal-fold vibration. In another part of this work, we continue to examine in detail the fundamental-frequency (F_o) variations during sentence utterances in English and in French.[3,4] These studies are leading to a description of such F_o variations in terms of sequences of a small inventory of attributes that we believe are related to underlying mechanisms of laryngeal control. Rules are being formulated to delineate in what manner the syntactic and semantic grouping of lexical items in a sentence, together with a principle (still inadequately defined) of "least physiological effort," imposes constraints on the sequences of these attributes. In order to examine the extent to which these F_o variations are processed and interpreted by listeners, we are carrying out a series of experiments in which certain aspects of the F_0 contours in sentences are systematically manipulated, and we are obtaining listener judgments of discrimination, naturalness, and other aspects such as emphasis of the utterances.

References

1. Thomas Baer, "Investigation of Phonation Using Excised Larynxes," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., February 1975.
2. Edwin Rich, "Air-Driven Oscillations of Fluid-Loaded Membranes," S.M. Thesis, Department of Electrical Engineering, M.I.T., October 1975.
3. Shinji Maeda, "A Characterization of Fundamental Frequency Contours of Speech," Quarterly Progress Report No. 114, Research Laboratory of Electronics, M.I.T., July 15, 1974, pp. 193-211.
4. Jacqueline Vaissière, "On French Prosody," Quarterly Progress Report No. 114, Research Laboratory of Electronics, M.I.T., July 15, 1974, pp. 212-223.

3. PRODUCTION AND PERCEPTION OF STOP CONSONANTS

National Institutes of Health (Grants 5 RO1 NS04332-13 and 1 T32 NS07040-01;
Fellowships 1 F22 NS007960 and 1 F22 HD019120)

Sheila Blumstein, Marcia A. Bush, William E. Cooper, William F. Ganong III,
Dennis H. Klatt, Kenneth N. Stevens, Katherine L. Williams, Catherine Wolf

We continue to conduct experimental studies of speech perception with different stimulus configurations, and to explore theoretical notions that will lead to a more complete model of the processing that occurs when a listener interprets a speech signal in terms of underlying segmental phonetic features. A guiding concept in this research is that the auditory system is predisposed to respond in a distinctive fashion to particular patterns of properties in the acoustic signal, and that these properties bear a rather simple and direct relation to the underlying phonetic features or to configurations of such features.[1] Among the questions that must be answered by further research are the following.

What is the inventory of properties to which specific detecting mechanisms respond?

Do the detectors only respond to simple patterns that are visually salient in an intensity-frequency-time representation of the signal, or do they respond to more global properties that represent combinations of simpler attributes?

Are the property detecting mechanisms available in the early weeks of human life or are they shaped on the basis of exposure to speech stimuli?

What is the relation between the property detecting mechanisms and the underlying phonetic features?

These and other questions are being examined currently with reference to the stop consonants.

a. Bursts and Transitions as Cues for Place of Articulation

Experiments are in progress to delineate the cues for place of articulation for English stop consonants in syllable-initial and syllable-final positions. In these studies we examine the perception of real speech that is altered by removing cues, as well as synthetic speech in which the bursts and transitions are systematically manipulated. We anticipate being able to specify the global properties that operate to signal place of articulation distinctions, and the degree to which these properties are independent of phonetic context.

In another series of experiments we are using a selective adaptation paradigm in which stimuli with place-of-articulation cues signaled by formant transitions are adapted by stimuli to which additional cues that enhance or conflict with the transition cues are appended. The results of these and other related adaptation experiments should provide some evidence for the existence of integrated property detectors that respond to a combination of acoustic attributes. Along similar lines, in experiments with noise bursts added to stimuli and with place of articulation cued by transitions we are determining the range of intensity, spectrum, and time location of the bursts in which the quality of the stimuli is enhanced.

b. Acoustic Analysis and Perception of Voicing Contrast for Spanish Stop Consonants

We are carrying on studies to examine the acoustic properties that distinguish between voiced and voiceless stops in Spanish, to determine the cues for voicing based on listener identification by Spanish speakers of a series of synthetic stimuli in which voice-onset time (VOT) is manipulated through a series of negative and positive values, and to investigate the cues for voicing based on listener identification of naturally produced speech sounds from which certain acoustic attributes have been removed. The acoustic analysis has shown that for several dialectal groups, in both voiced and voiceless categories, there is a systematic variation in VOT according to place of articulation. A tentative explanation of these phenomena can be given in terms of an aerodynamic model of speech production. The listener-identification experiments with synthetic stimuli show that the boundary between voiced and voiceless responses occurs at a VOT value that is consistent with data from acoustic analysis of Spanish stop consonants. The identification test with edited real-speech utterances shows that although prevoicing is a cue for voicing for many stop consonants, the removal of prevoicing from naturally produced voiced stops does not always lead to the perception of a voiceless stop.

c. Voicing Cues for Plosives

The voice-onset time (VOT) and the duration of the burst of frication noise at the release of a plosive consonant have been measured from spectrograms of word-initial consonant clusters.[2] The report of that work contains a review of the perceptual literature on cues to voicing. Six perceptual cues are proposed: (i) burst and voicing onset as one or two perceptually distinct events, (ii) presence or absence of low-frequency energy after release caused by voiced excitation of low first formant, (iii) durations of preceding vowel and plosive closure interval, (iv) burst intensity, (v) fundamental frequency change at voicing onset, and (vi) presence or absence of low-frequency energy during closure. Variations in VOT are explained in terms of articulatory mechanisms, perceptual constraints, and phonological rules. Some VOT data obtained from a connected discourse have also been analyzed and organized into a set of rules for predicting voice-onset time in any sentence context.

References

1. K. N. Stevens, "Speech Perception," in D. B. Tower (Ed.), Human Communication and Its Disorders, Vol. 3 (Raven Press, New York, 1975), pp. 163-171.

2. D. H. Klatt, "Voice Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters" (to appear in J. Speech Hearing Res.).

4. STUDIES RELATING TO SPEECH TIMING AND MEMORY

National Institutes of Health (Grant 5 RO1 NS04332-13)

William E. Cooper, A. W. F. Huggins, Dennis H. Klatt

a. Timing

Recent work on timing in speech production has been concerned with the pattern of segmental durations observed in a connected discourse,[1] the perception of changes in the timing of sentences,[2] and the development of a quantitative description of the uses of segmental duration to cue linguistic contrasts in English.[3] In the first study broadband spectrograms of a connected discourse read by a single speaker of American English were made. Segmental durations were measured for each segment type in stressed and unstressed environments. Segments significantly longer than their median duration were found to mark the ends of major syntactic units, including the boundary between a noun phrase and a verb phrase. Further analysis indicated that the voicing feature of a postvocalic consonant only has great influence on vowel duration in phrase-final syllables. In the perceptual study a category judgment technique was used to quantify the preferred duration for phonetic segments in words placed in different syntactic environments.[2] Results indicated that listeners adjust their expectation of segment duration in accordance with the syntactic position of the word in the sentence. In particular, the segmental lengthening that is seen in the production of a phrase-final syllable is represented in a listener's perceptual schema.

b. Temporally Segmented Speech

We have completed a study of the recovery of intelligibility as a function of the duration of the silent intervals used to break up continuous speech into speech intervals of any desired length[4] (see Part II, Sec. XVI-B). The critical parameter of the intelligibility seems to be the interval from the start of one speech interval to the start of the next. The implications of the result for models of auditory processing are being explored.

c. Dichotic Perception of Iterated Segments of Noise

Earlier studies with both clicks and speech have suggested that preliminary processing of the acoustic input is performed separately for both ears before the results of the two are combined.[5] This idea has been challenged, so an experiment was performed to test it directly. Iterated segments of noise are heard as a noise with a "motorboating" quality. If a sequence of noise segments ABABAB ... is presented to the left ear, while the sequence BABABABA ... is presented to the right, the heard period of the motorboating tells us whether the repetition is directed separately for each ear (repeating segment = (AB) in each ear) or centrally (repeating segment = A in one ear, B in the other). Preliminary results show that the period is detected for each ear, and is <u>not</u> central, at least up to segment lengths of 100-150 ms.

References

1. D. H. Klatt, "Vowel Lengthening Is Syntactically Determined in a Connected Discourse," J. Phonetics 3, 129-140 (1975).

2. D. H. Klatt and W. E. Cooper, "Perception of Segment Duration in Sentence Contexts" (to be published in A. Cohen and S. G. Nooteboom (Eds.), Structure and Process in Speech Perception, Proceedings of a Symposium on Dynamic Aspects of Speech Perception, I. P. O., Eindhoven, The Netherlands, August 4-6, 1975). This work was also reported in Progress Report No. 116, Research Laboratory of Electronics, M.I.T., July 1975, pp. 189-205.

3. D. H. Klatt, "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," a paper presented at the Annual Meeting of the American Speech and Hearing Association, Washington, D. C., November 21-23, 1975.

4. A. W. F. Huggins, "Temporally Segmented Speech and Echoic Storage" (to be published in A. Cohen and S. G. Nooteboom (Eds.), Structure and Process in Speech Perception, Proceedings of a Symposium on Dynamic Aspects of Speech Perception, I. P. O., Eindhoven, The Netherlands, August 4-6, 1975).

5. A. W. F. Huggins, "Temporally Segmented Speech," Percept. Psychophys. 18, 149-157 (1975).

5. STUDIES OF SPEECH PRODUCTION AND SPEECH DISCRIMINATION IN CHILDREN

National Institutes of Health (Grant 1 T32 NS07040-01 and
Fellowship 1 F22 HD01919-01)

Ursula Goldstein, Lise Menn, Katherine L. Williams

a. Acoustic Analysis of Vowels Produced by Children

A study is being made of the feasibility of applying linear prediction analysis procedures to the problem of measuring formant frequencies and other acoustic parameters in utterances produced by infants and children. Various versions of linear prediction methods are being attempted, including variation of window size and number of coefficients. Preliminary results suggest that the first two formant frequencies can be extracted reliably when formants are steady or moving slowly. Further attempts to increase accuracy and reliability are in progress.

b. Longitudinal Study of the Acquisition of English as a First Language

A child of English-speaking parents has been observed intensively over a basic period of 8 1/2 months, beginning at age 12 1/2 months.[1] Extensive instrumental analysis of tape recordings made during this period has begun, and is planned for the forthcoming months. These instrumental data, together with other observational data, are being examined in an attempt to quantify the processes involved in the learning of phonemic contrasts, phonetic targets, intonation patterns, and other phonological aspects of language development.

c. Infant Speech Sound Discrimination Using the High-Amplitude Suck Procedure

This investigation focuses on three questions: (i) Can young infants discriminate a synthetically produced place of articulation that is cued by various combinations of release burst and formant transitions? (ii) What is the relative effectiveness of contingent and noncontingent sound presentation when using the High-Amplitude Suck procedure in an infant speech sound discrimination study? (iii) When using the High-Amplitude Suck procedure, is there an interaction between the infant's behavioral state and either the infant's ability to discriminate or the method of sound presentation? In the High-Amplitude Suck procedure a speech sound is presented to the infant repeatedly while the sucking rate is recorded. A significant increase in the sucking rate following a sound shift, compared with no change or with insignificant change in sucking rate when no shift in sound occurs, furnishes evidence of discrimination between the two sounds that were contrasted in the shift condition and presumably reflects the infant's attention to a new sound. The data collection phase of this study is nearly complete, and data analysis will be carried out very soon.

References

1. L. Menn, "Preliminary Report on Longitudinal Intensive Observation of the Acquisition of English as a First Language," Progress Report No. 116, Research Laboratory of Electronics, M.I.T., July 1975, pp. 225-227.

6. ACOUSTIC STUDIES OF SPEECH SOUNDS: INVARIANT ATTRIBUTES AND SPEAKER DIFFERENCES

U.S. Army (Contract DAAB03-75-C-0489)

William L. Henke, Kenneth N. Stevens, Victor W. Zue

The aim of these acoustic studies is to obtain a substantial body of data on various acoustic characteristics of segmental and prosodic speech events produced in words and in sentences by different speakers. These data are being examined and interpreted in a way that specifies invariant aspects of the data and that shows the variability of the attributes resulting from differences in phonetic context and from speaker differences. Our objectives are to gain better insight into the production and perception of speech, to provide a more solid base for linguistic and phonological theories, and to provide data and models that will be of practical use in speech recognition and speaker recognition applications.

The data base for studies of attributes of segmental events has two parts: (i) a large corpus of nonsense utterances where the phonetic context is carefully controlled, and (ii) facilities through which various aspects of the acoustic characteristics of these utterances can be examined systematically and conveniently. Using this acoustic-phonetic data base, spectral and durational characteristics of English stops in prestressed position, both in consonant-vowel context and in clusters, are studied in detail. Results indicate systematic variations as a function of voicing characteristics, place of articulation of the stops, and the phonetic environment. The variations in timing and spectral characteristics of stop burst can be interpreted in terms of the articulatory positions and movements utilized to produce the sounds. A second, more limited, data base has been designed to examine the effects of consonant voicing on fundamental frequency and glottal onset characteristics for the vowel following the consonant. Initial analysis of these data[1,2] show the expected substantial influence of consonant voicing on these parameters, with some systematic interspeaker differences.

The data base of sentence material and connected text produced by a number of speakers is being analyzed in an attempt to extract systematic interspeaker differences in parameters that describe fundamental frequency contours and other larynx-related attributes.

References

1. V. W. Zue, "Acoustic Phonetic Data Base for the Study of Selected English Consonants, Consonant Clusters, and Vowels," J. Acoust. Soc. Am. 57, S34(A) (1975).
2. V. W. Zue, "Duration of English Stops in Prestressed Position," J. Acoust. Soc. Am. 58, S96(A) (1975).

XVI. SPEECH COMMUNICATION

A. RESPONSES TO AN UNEXPECTED SUDDENLY INDUCED CHANGE IN THE STATE OF THE VOCAL TRACT

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

Joseph S. Perkell

[A slightly different form of this report was presented as a paper under the title "Feedback of Intraoral Air Pressure?" at the 90th Meeting of the Acoustical Society of America, San Francisco, California, 3-7 November, 1975.]

1. Introduction

Since laryngeal,[1] swallowing,[2] facial,[3] and mandibular reflex mechanisms exist at the brainstem level, it would not be surprising if these mechanisms were utilized to some extent in the coordination of a complex series of speech motor gestures. More specifically, it is possible that the elaboration of some details of articulatory commands might be partially dependent on feedback, some of it in the form of information from peripheral receptors in the vocal tract. Motor coordination of this nature has been hypothesized in a general way[4] and it has been demonstrated very elegantly in at least one complex motor task, coordination of head and eye movements.[5] Even if peripheral feedback mechanisms do not operate on a moment-to-moment basis in running speech, demonstration of their operation in experimental situations might at least suggest that they are available for use in the acquisition or establishment of more preplanned or centrally controlled motor sequencing.

Two recent experiments have utilized unexpected, suddenly induced alterations in the state of the vocal tract to explore the short-latency role of feedback. By interrupting mandibular closure during the onset of a /p/ imbedded in a carrier phrase, Folkins and Abbs[6] found that the lips responded with increased vertical movement to complete the closure. Even though no EMG changes were found and no detailed temporal analyses of the movement trajectories were presented, this finding suggests the possibility of fast-acting, low-level motor reorganization; and the phenomenon should be explored further. On the other hand, Putnam and Shipp[7] have reported no measurable change in the duration of laryngeal (PCA) electromyographic activity in response to an unexpected partial venting of intraoral air pressure during the production of a prestressed /p/ imbedded in a carrier phrase. This result appears to indicate that intraoral air pressure is not sensed and used in the short-latency control over the timing of the cessation of laryngeal abduction during the stop. Pressure venting, however, was only partial; there was no interference with alternate, possibly redundant feedback modalities, and other measurements of articulatory activity were not reported.

The experiment reported here is an extension of the pressure-venting paradigm. It also represents an attempt to look for evidence of short-latency changes in response to an unexpected, suddenly induced change in the state of the vocal tract.

2. Procedure

In the experiment, the subject repeatedly pronounces a test utterance, "hapapa again." It is hypothesized that the timing of some aspects of the articulation sequence "hapapa" may depend partially on completion of preceding components as signaled by lip closure and the consequent increase in intraoral air pressure during the bilabial stops. An unexpected reduction of pressure during the stops might then affect the timing relationships and/or the magnitude of commands to articulators, such as the lips, the larynx, or the mandibular musculature. Since feedback from the lips would also provide information about lip closure, pressure reduction should be tried with and without labial anesthesia. Presumably, reduced sensation from the lips might force a stronger reliance on other feedback modalities, thereby enhancing short-latency responses to an unexpected suddenly induced change in the state of the vocal tract. In other words, any observed pressure-alteration effect might be greater under a carefully controlled anesthesia condition.

The initial experiment was designed to explore the feasibility and possible usefulness of this paradigm. Only two measurements were made, timing of lip closure and timing of the onset of regular glottal vibration during the vowels.

Figure XVI-1 is a diagram of the experimental apparatus.[8] A vacuum system is used to ensure complete pressure evacuation during bilabial closure. The system consists of a reservoir tank that is evacuated by a water-faucet aspirator. The tank is connected through a series of three valves to a translabial tube. The oval-shaped open end of the

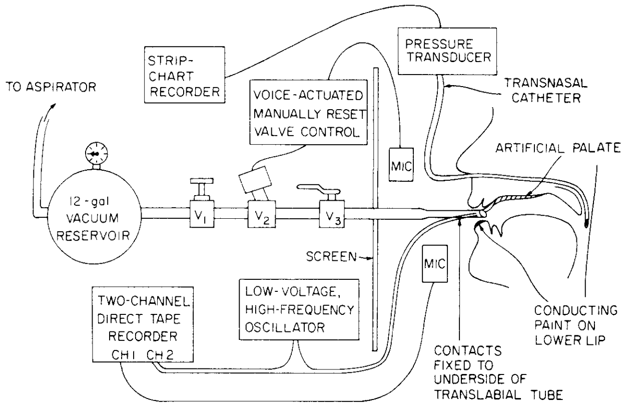

TO ASPIRATOR
STRIP-CHART RECORDER
VOICE-ACTUATED MANUALLY RESET VALVE CONTROL
PRESSURE TRANSDUCER
TRANSNASAL CATHETER
ARTIFICIAL PALATE
MIC
12-gal VACUUM RESERVOIR
V_1
V_2
V_3
SCREEN
MIC
TWO-CHANNEL DIRECT TAPE RECORDER CH1 CH2
LOW-VOLTAGE, HIGH-FREQUENCY OSCILLATOR
CONDUCTING PAINT ON LOWER LIP
CONTACTS FIXED TO UNDERSIDE OF TRANSLABIAL TUBE

Fig. XVI-1. Diagram of the experimental apparatus. The subject is located at the right, behind a screen which prevents him from observing the state of the ball valve (V_3) as it is set for each repetition by the experimenter.

tube is attached to a custom-made artificial palate that holds it in a fixed relationship with the maxilla while allowing freedom of movement for the lips and the mandible. The first valve, V_1, is used to regulate air flow to approximately the rate that would occur through the open glottis upon release of the prestressed stops. The second valve, V_2, is voice-actuated and solenoid-operated set to open always during the first /a/ and close during the final /a/ of /hapapa/. The third valve, V_3, is a ball valve set by the experimenter between each utterance to determine whether or not air will flow through the tube. The state of the ball valve, and thus whether or not there will be pressure evacuation, is unknown to the subject for each utterance.

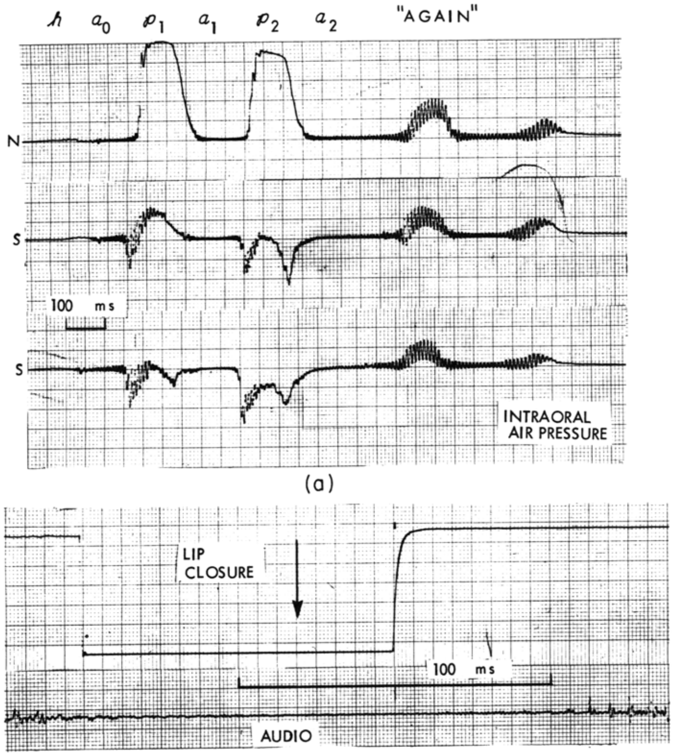
Intraoral air pressure is monitored with a transnasal catheter connected to a transducer, the output of which is recorded directly on a strip-chart recorder. To record lip closure, the lower lip is painted with a conducting paint. The painted area serves as a switch to close a loop formed by two contacts fixed to the underside of the translabial tube, a signal generator, and the input to one channel of a tape recorder. The audio signal is recorded on the second tape channel.

A trial included 50 utterances, 25 with the ball valve open, and 25 closed. The open and closed conditions were altered randomly. Four trials were run on two separate days. On each day there was a normal run followed by a run with labial anesthesia. Both topical and injection techniques were tried and each presented a unique set of problems.[9] The injection technique ultimately produced a subjectively greater loss of awareness of lip closure with adequate motor control. The results of the injection trial and the normal run that preceded it have been analyzed.

3. Results and Discussion

A tape dubbing was made for the analysis to obtain a speed reduction, and graphical records were made from the dubbing. Figure XVI-2a shows three intraoral pressure records, and Fig. XVI-2b shows a time-expanded audio trace and a lip-closure trace. The phonetic sequence corresponding to the pressure traces is shown in Fig. XVI-2a. The pressure trace labeled N (normal) was produced with the ball valve closed, and normal increases in intraoral pressure can be observed during the two stops p_1 and p_2. The two traces labeled S (suction) were produced when the ball valve was open and the oral cavity was connected to the vacuum reservoir. Variation in the average pressure from a slight rise to a moderate decrease with respect to atmospheric pressure can be seen for the four examples shown. When the pressure was below atmospheric, a subjective impression of suction was produced, and it is evident that the net amount of suction varied, depending presumably on the respiratory effort with which each stop was produced.

The suction (S) records also show periodic fluctuations that last throughout most of the closure intervals. These fluctuations should be due to glottal vibrations induced in



(a)

(b)

Fig. XVI-2. (a) Traces of intraoral air pressure as a function of time for three repetitions. The phonetic sequence corresponding to the traces is shown. The trace labeled N was produced with the ball valve closed ("normal" intraoral air pressure increases during the stops), and the two traces labeled S were produced with the ball valve open ("suction" condition during the stops).

(b) Time-expanded lip-closure trace and a synchronous audio trace. The lip-closure trace was obtained by envelope detecting and thresholding the recorded and dubbed output of the oscillator. The audio trace is retouched for clarity.

the spread configuration by Bernoulli forces associated with a high transglottal air flow. Once the bilabial constriction is released and the suction-induced air flow ceases, we presume that the glottis assumes a normal configuration, and a normal-appearing delay of onset of voicing shows on the audio record. Temporal measurements were made from the lip-closure and audio records.

Figure XVI-3 is a composite bar graph of the temporal measurements for four conditions: normal sensation with and without suction and anesthesia with and without suction. The lengths of the bars represent means (ms) calculated from the number of utterances indicated on the right. Lip closure for the stops p_1 and p_2 is indicated by

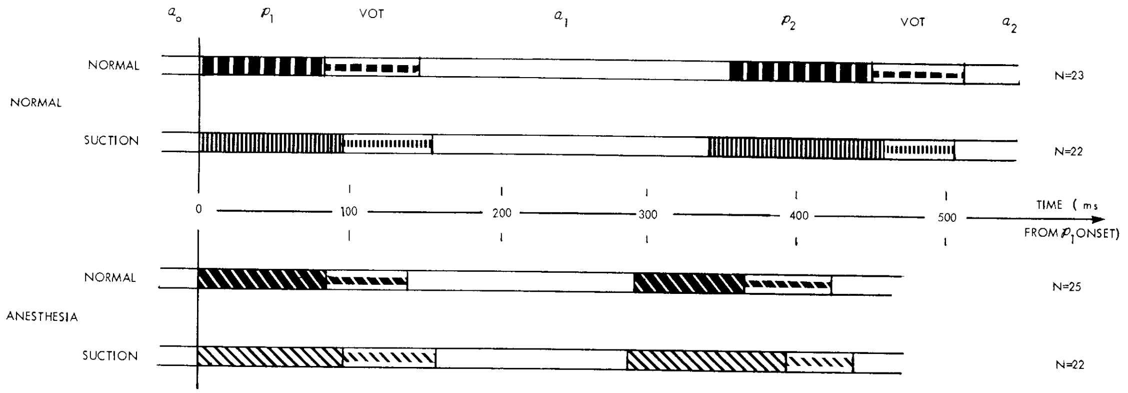


Fig. XVI-3. Composite bar graph of the temporal measurements for 4 conditions: normal lip sensation with and without suction, and labial anesthesia with and without suction. Lengths of bars represent means (in milliseconds) calculated from the number of utterances indicated on the right. Lip closure for the stops p_1 and p_2 is indicated by wide filled bars, voicing onset time (VOT) by half-filled bars, and the vowels a_1 and a_2 by the unfilled portions.

wide filled bars, voicing onset time (VOT) by half-filled bars, and the vowels a_1 and a_2 by the unfilled portions. It is obvious that the speaking rate was higher during the anesthesia run. Detailed observations can be made more readily from the following figures.

Figure XVI-4a and 4b shows closure durations for p_1 and p_2 for the four conditions. The mean duration (ms) is given at the end of each bar, and the difference between each suction vs normal pair of means with the two-tailed t test confidence level is given between each pair of bars. The width of two standard deviations is indicated by brackets. For each suction vs normal pair the stops pronounced under suction are longer, with the differences being greater for p_2 than for p_1. These differences can probably be explained simply on the physical basis that suction produces a force working in opposition to lip opening. The greater differences for p_2 are probably due to the lesser degree of stress and lower net intraoral pressure with which the second stop was usually produced.

Figure XVI-4c shows the mean p_2 VOT durations (the intervals between p_2 release and a_2 voicing onset) for comparison with the p_2 durations shown in Fig. XVI-4b. In both anesthesia and normal sensation conditions, when p_2 closure is lengthened by suction VOT is somewhat shorter. This result indicates that closure lengthening tends, to some extent, to take place at the expense of VOT. But this relationship does not hold for p_1 and the next figure shows that the trading relationship is not complete.

Figure XVI-5a shows for p_1 and Fig. XVI-5b for p_2 the sum of closure duration and the following VOT. The sum of the two durations is greater under the suction

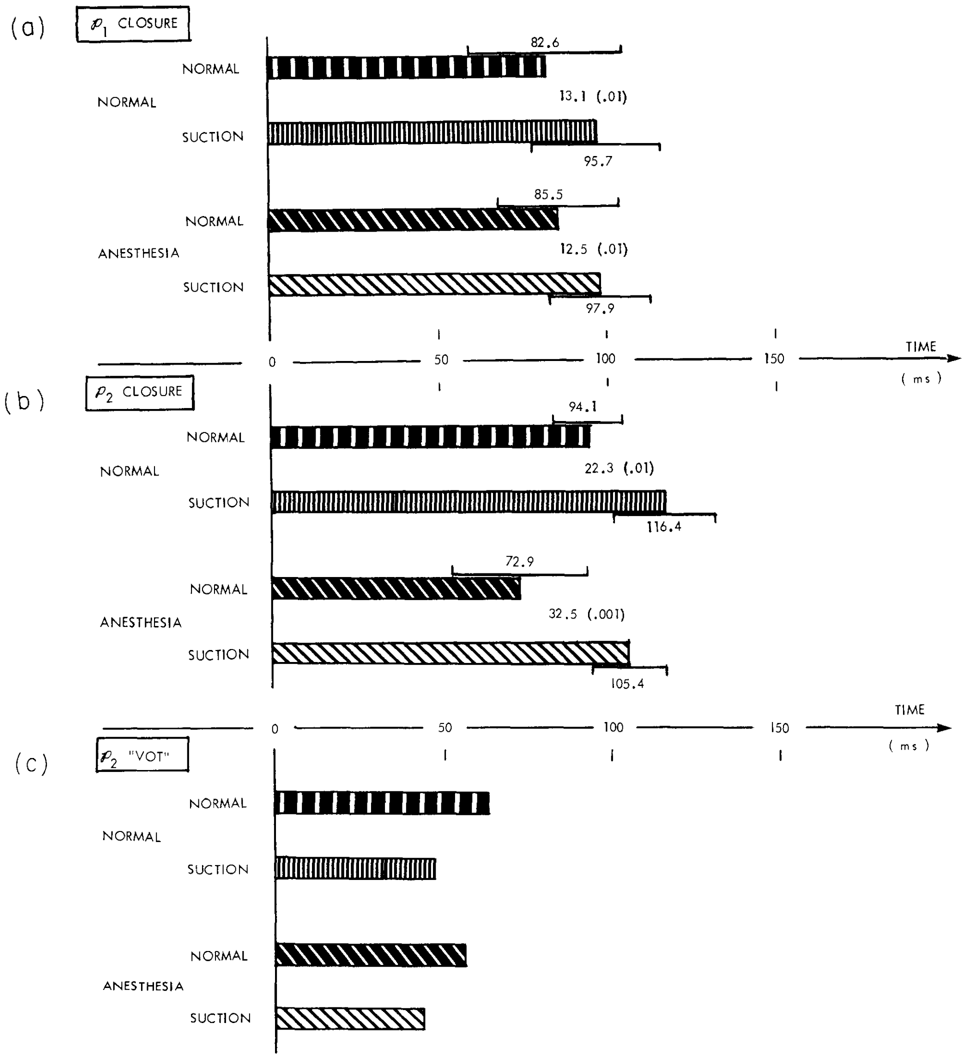


Fig. XVI-4. (a) Lip-closure durations for p_1 under the four conditions listed in Fig. XVI-3. The mean duration (ms) is given at the end of each bar, and the difference between each suction vs normal pair of means with the two-tailed t test confidence level is given between each pair of bars. The width of two standard deviations is indicated by the brackets.

(b) Lip-closure durations for p_2.

(c) Durations for p_2 VOT (the intervals between p_2 release and voicing onset of the following vowel a_2).

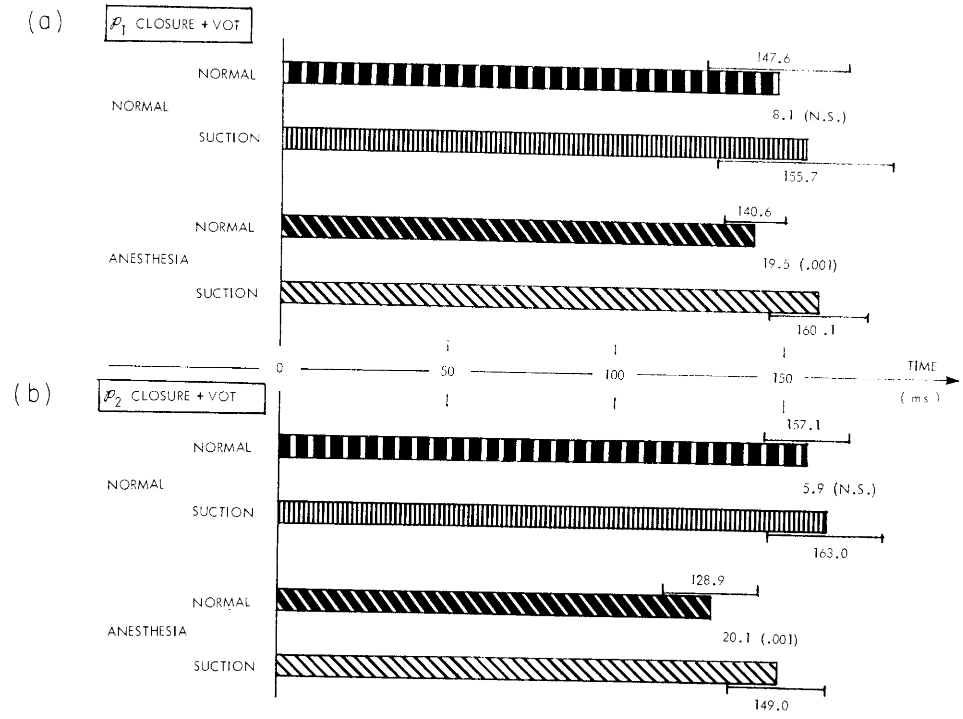


Fig. XVI-5. The sum of lip-closure duration and the following VOT for p_1 (a) and p_2 (b) under the four conditions.

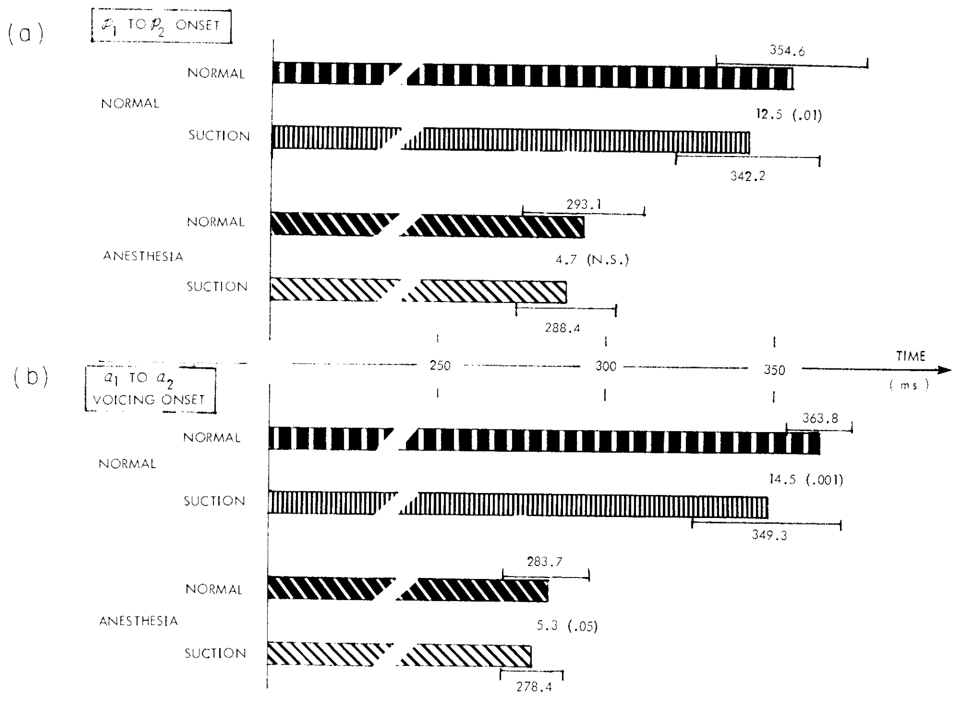


Fig. XVI-6. (a) The interval between the onset of p_1 and p_2 and (b) the onset of voicing for a_1 and a_2 under the four conditions.

than under the normal condition. This difference is more pronounced for the anesthesia cases and, in fact, the differences are not significant in the nonanesthesia cases.

It has been mentioned that the glottis must be forced into an abnormal state during the stop by the suction-induced air flow. It is also possible that some more complicated physical explanation could account for the delayed VOT. Nevertheless, at this point the possibility of altered articulatory commands to the larynx cannot be precluded without appropriate EMG data taken in conjunction with lip-closure measurements.

Figure XVI-6 shows two measurements of intersyllabic intervals, the interval between the onset of p_1 and p_2 (Fig. XVI-6a), and the interval between voicing onsets for a_1 and a_2 (Fig. XVI-6b). In all cases the intervals are shorter during the suction conditions. This shortening could be caused by any number of factors, and it may be due to an attempt to abbreviate the utterance in response to an awareness of unusual conditions during the first closure. This idea might be reinforced slightly by the fact that utterance shortening appears to be greater in the nonanesthesia condition, presumably when more information is available to stimulate a conscious reaction. Since the intervals between syllables do exceed established reaction times, these data are less relevant to the hypothesis.

4. Conclusions

Sudden unexpected decreases of intraoral air pressure during the production of voiceless, prestressed, intervocalic bilabial stops appear to induce short latency changes in the timing of lip and glottal gestures. These changes must be due in part to aerodynamic effects. The larger, significant increase in closure duration plus VOT produced under the anesthesia condition, however, is difficult to explain on a purely aerodynamic basis. The possibility of feedback-related alterations in motor commands cannot be precluded without accompanying detailed EMG measurements.

The apparent aerodynamic effect on lip movement indicates that the lip musculature must be operating in force ranges that are roughly equivalent to aerodynamic forces normally generated in speech.

The results are interesting enough to suggest that the paradigm might be used in conjunction with EMG measurements, other measurements of articulatory movements, and carefully controlled applications of anesthesia to study possible feedback effects, as well as the interactions between speech aerodynamics and physiological properties of the articulators.

Footnotes and References

[1] M. Sawashima, "Laryngeal Research in Experimental Phonetics," Status Report on Speech Research SR-23, Haskins Laboratories, New Haven, Connecticut, 1970, pp. 69-115.

2. R. W. Doty, "Neural Organization of Deglutition," Handbook of Physiology IV, 1861-1902 (1968), Chap. 92.
3. E. Kungelberg, "Facial Reflexes," Brain 75, 385-396 (1971).
4. R. E. Burke, "Control Systems Operating On Spinal Reflex Mechanisms," Chap. IV, "Central Control of Movement," Neurosciences Research Program Bulletin, Vol. No. 9, No. 1, pp. 60-85, 1971.
5. E. Bizzi, R. E. Kalil, and U. Tagliasco, "Eye-Head Coordination in Monkeys: Evidence for Centrally-Patterned Organization," Science 173, 452-454 (1971).
6. J. W. Folkins and J. H. Abbs, "Lip and Jaw Motor Control during Speech: Responses to Resistive Loading of the Jaw," J. Speech Hearing Res. 18, 208-220 (1975).
7. A. H. B. Putnam and T. Shipp, "EMG-Aerodynamic Measures for Intervocalic /p/ Production," J. Acoust. Soc. Am., Vol. 57, Suppl. 1, p. 570, 1975 (Abstract).
8. I am grateful to Donald K. North for his invaluable assistance in constructing the experimental apparatus and as an extremely helpful experimenter.
9. Topical anesthesia was attempted by applying an experimental 30% lidocaine cream to the lips for 1 hour. This procedure produced some loss of sensation with a diminution in two-point discrimination ability, but without significant loss of awareness of lip contact. Bilateral infraorbital and mandibular blocks were injected with a 2% xylocaine solution containing epinephrine at a concentration of 1:100, 000. The injections produced a loss of motor control profound enough to preclude performance of the task for approximately 2 hours. The trial was run when the subject regained sufficient motor control but still retained some uncertainty of an awareness of lip closure.

B. TEMPORALLY SEGMENTED SPEECH AND "ECHOIC" STORAGE

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

A. W. F. Huggins

This report defends the argument that successive events tend to be integrated in perception if they are similar and follow each other rapidly, but they are perceived as separate if the succession is more leisurely. This idea is far from new; for example, in one of the important papers that led to the development of Gestalt psychology, Wertheimer's study of the phi-phenomenon (apparent movement),[1] exactly this conclusion was reached. Two lines, viewed sequentially in a tachistoscope, were seen as simultaneous when separated by less than 30 ms, but as successive when separated by 200 ms or more. At intermediate separations, a single moving line was often seen. Wertheimer's results have been revised and expanded since then, but the principle is clear. More recently, Stroud[2] proposed that psychological time is not continuous but occurs in discrete "moments." Two events that occur within a single moment are integrated; they can only be perceived as separate if they occur in separate moments. Stroud's theory has been hard to sustain; most of the evidence points to a continuous, rather than a discrete, perceptual process. But the same principle is present: rapid succession leads to integration, slower succession to separation.

Before we turn to audition, let us consider briefly the implications of this principle in visual perception. We shall start with two of the germinal studies in the growth of the information-processing approach to perception and cognition. Sperling[3] showed that all of the information in a brief visual display was available for a short time after the display was switched off but that it decayed rapidly. His display was an array of 9 or 12 letters in 3 rows. The subject was told which row to report by a high, middle, or low-pitched tone that followed the display offset. With short delays, any row was correctly reported, but with delays longer than half a second, only 3 or 4 letters in the array could be reported, and there was no further decay. Averbach and Coriell[4] asked for a single letter to be identified in an array, and used a visually presented bar or circle to point to the required letter after the array had disappeared. Their results were similar to Sperling's, but there was an additional finding that the circular marker sometimes erased the letter to which it was pointing. These results, with others reviewed by Neisser,[5] suggest that visual input is stored initially in a rich raw form in "iconic" memory, from which it is read out sequentially for subsequent processing. Information in storage decays quite rapidly, and must be read out quickly if it is to be perceived. A new visual input interferes retroactively with the processing of earlier inputs.

These findings seem to lend support to ideas like Stroud's, that perception occurs in a blink-by-blink fashion. But the findings may not be correct for <u>all</u> modes of perception. It is possible that discrete processing <u>only</u> takes place when the input is discrete. Since the perceptual apparatus was molded by evolutionary pressures, stimulus situations that do not occur in the real world often provide a way of penetrating the perceptual armor because ability to handle these situations is of no advantage to the organism. But it may be dangerous to lose sight of the fact that the stimulus is unusual: normal stimuli may be processed differently. Thus the finding that a later stimulus can interfere with processing an earlier stimulus may not mean that this always happens. If it did, how would we see anything at all? Erasure of the earlier stimulus is obviously appropriate when visual attention is being switched off an old object and onto a new one, but it is <u>not</u> appropriate as long as the focus remains on one object. Although there are some situations requiring repeated rapid shifts of fixation – reading is a good example – there are many other activities where changes of fixation are much less frequent, and the interesting things happen during fixation. Any activity that requires visual tracking falls in the second class; for example, watching any object that is itself moving.

Consider the success of the motion picture industry. Successive frames of a film do not erase the earlier frames but are sequentially integrated with them into an ongoing perception of action. The situation is different from that producing erasure, in that successive frames of a motion picture are very similar. The developing action is continuous. If the interval between exposing successive frames is lengthened, so that the similarity between them disappears, then perception is indeed disrupted. The effect is striking and has led to a popular party game of illuminating the party with a slow-running strobe light.

Thus we may conclude that the perceptual result of following one stimulus with another will depend on the intervening interval, the similarity between the stimuli, and whether the task requires that the stimuli be integrated or separated. Short intervals and similar stimuli will assist their integration into one event, and make it very difficult to treat the two as separate percepts. Longer intervals and dissimilar stimuli will have reverse effects.

We know much less about auditory processing than about visual processing. Until the advent of computers it was much more difficult to obtain complete control of the stimulus. Although there are striking differences between audition and vision, the same model of peripheral raw storage followed by sequential read-out has been proposed for audition.[6] The need for peripheral storage is much more obvious in audition: sound travels past the observer at 1100 ft/s, so there is no possibility of taking a second look, as often there is in vision where the objects of fixation tend to be relatively permanent. Also, auditory input is already sequential, unlike vision. Sound is composed of pressure changes over time, and the short-term spectrum of the sound itself changes over time. One can think of meaningful sound as sequential in both microstructure and macrostructure. Vision, on the other hand, is parallel in microstructure, and sequential in macrostructure. I have argued elsewhere[7] that a storage mechanism that permits stored information to decay over time may not be appropriate for a medium in which time is so inextricably involved in the definition of the stimulus. Decay would produce qualitative as well as quantitative changes in the stimulus.

Many of the foregoing arguments for vision apply also to audition. Similar events that follow each other rapidly tend to be integrated into a single percept, whereas less similar events are perceived as separate. For example, Miller and Heise[8] found that two tones alternated at 5 Hz were heard as a single warbling tone when the frequency separation of the tones was small, but when the separation was increased two separate interrupted tones were heard. Musicians have known of this effect for centuries: Bach's music is full of examples.

The disruption of auditory perception by lack of continuity is best exemplified perhaps by how difficult it is to report the temporal order of a cyclically repeating set of disparate sounds. If the sounds are as different as buzzes, hisses, tones, and vowels, each sound must last 300-700 ms for listeners to report the sequence correctly.[9] On the other hand, when four vowel sounds with abrupt boundaries were used, temporal order was reported correctly when each vowel lasted 125 ms but not when it lasted 100 ms.[10] In the latter case the lack of continuity was much less severe, since all sounds were from a single speaker at the same fundamental frequency. The phenomenon of primary auditory stream segregation[11] provides further evidence that the perception of temporal order is strongly influenced by the similarity of the component tones. It can be argued that a large discontinuity in the input forces the perceptual apparatus to treat the sounds on each side as separate, and the rate at which new unrelated percepts can be handled is limited. Although the sequential sounds in the foregoing studies interfere with each other to the extent that they disrupt perception of the temporal order, subjects report that they hear each of the sounds clearly. Perhaps this is not surprising, since the sounds each last 100 ms or more.

When the sounds are made very short, however, an auditory erasure effect is found that is similar to the visual erasure mentioned above. Thus Massaro[6] has reported that 20-ms bursts of tone that are followed after a variable silent interval by a masking tone at the same intensity cannot be reliably identified as being slightly higher or lower in frequency than the masker when the silent interval is less than 150-200 ms. Similar results have been observed with vowel sounds. These findings, and those that we have outlined for temporal order, would seem to create severe problems in explaining how speech is perceived. The different acoustic events produced by the articulation of speech are brief (often less than 60 ms) and follow each other very rapidly. Why do they not mask each other, and how are they different from other sounds, in that we have no difficulty perceiving their temporal order?

The answer, I argue, is that speech is not to be regarded as made up of a sequence of unrelated events, but rather as a single continuously developing process. Backward recognition masking should only take place when the focus of attention between one sound source and another is changed. When auditory attention is fixed on an ongoing sound, even quite large disturbances fail to produce noticeable masking. Listeners make large errors when asked to locate a click within a sentence[12] or even a loud buzz or cough that completely replaces a sound in the sentence.[13] On the other hand, when a silent interval is used to replace a sound, no errors are made in locating it. Auditory perception was forced to evolve in such a way as to be able to cope with extraneous noises that occur frequently in life. But there is no naturally occurring event that produces a short silence in the middle of an ongoing sound. On the contrary, a silence indicates that a sound-generating event has finished. Thus if a silent interval is artificially introduced into speech, it is immediately perceived as unnatural, unless it can be integrated into the speech and perceived as a stop.[14]

The introduction of silent intervals does not necessarily interfere with speech intelligibility. In an early study of interrupted speech, Miller and Licklider[15] reported that at rates of interruption above 10 per second (the speech was switched off for 50 ms every 100 ms) intelligibility of PB words was little affected. They likened the experience to viewing a scene through a picket fence. The result has been repeated with running speech.[16, 17] Longer silent intervals did reduce intelligibility, of course, since substantial parts of the speech wave were suppressed by the interrupting switch. But intelligibility can be degraded by inserting silent intervals even when none of the speech is suppressed.[7] A continuous-speech message was "temporally segmented," that is, divided into speech intervals by effectively splicing 200 ms silent intervals into the recorded speech, without discarding any of the speech. Intelligibility declined approximately 30% for each halving of speech interval duration below 200 ms. On the other hand, if relatively short speech intervals of 63 ms were separated by silent intervals of variable duration, intelligibility remained low (~50%) as long as silent intervals lasted 125 ms or more. But as silent intervals were shortened from 125 ms to 63 ms, intelligibility abruptly recovered to ~100%, where it remained for shorter silent intervals. We shall refer to the last finding as the gap-bridging result.

These results fit in exactly with the arguments made here. When silent intervals are short, it is easy to integrate successive similar sounds into an ongoing percept, but hard to treat successive sounds as different. As silent intervals are lengthened it becomes progressively harder to integrate the successive speech intervals into an ongoing percept, and easier to treat them as separate independent events. If the foregoing interpretation is accepted, the present results agree remarkably well with Thomas's study on temporal order in vowels.[10] Thomas and his co-workers found that temporal order was reported correctly when each of the four cyclically repeated vowels lasted 125 ms, but not when they lasted 100 ms. That is, each vowel could be processed separately when it lasted 125 ms. In the gap-bridging result, silent intervals of 125 ms could not be bridged, so successive speech intervals had to be processed separately, and intelligibility was low. On the other hand, when the combined duration of speech and silent interval was 125 ms, the successive speech intervals could be integrated, and intelligibility was high. The results seem highly compatible with an account in terms of a short-term storage mechanism, or a "processing time." As such they are reminiscent of earlier theories put forward to account for the loss of intelligibility that occurs when speech is cyclically alternated between the ears.[16] Unfortunately, the processing-time argument has been severely undermined by the finding that in both alternated speech and the very similar temporally segmented speech, the effects are tied to speech rate rather than to absolute duration.[17-19]

Therefore before making too much of the agreement between the gap-bridging result and temporal-order results with vowels, we should explore the parameters of the gap-bridging result, and confirm that the result is not tied to speech rate.

1. Method

The experimental approach involves separating the duration of the speech intervals from the amount of speech that they contain by varying the playback speed of the speech, and measuring the effect of silent interval duration on intelligibility. The speech materials were the same two sets of nine 150-word passages used in earlier studies. Each of 18 passages was temporally segmented in two ways, as diagrammed in Fig. XVI-7.

First, a complete 150-word passage was digitized, and stored in the Speech Communication Group's PDP-9 computer. The waveform was then "marked," in unused low-order data bits, to divide it into speech intervals lasting 62.5 ms. Next, the sampling rate was increased by a factor of 1.189 ($=2^{1/4}$), and the silent interval duration was set, say to 125 ms. The stored speech was then played out at the increased sampling rate, and the computer inserted 125 ms of silence whenever it encountered a segmenting mark. Next, the sampling rate was reduced below the rate used for input by the same factor of 1.189. The silent interval duration was reset to 125 ms, and the segmented speech was played out and recorded. The same segmenting marks were used for both versions of a passage, so that corresponding speech intervals had exactly the same waveform. They differed only in time scale: speech intervals lasted 74 ms in the SLOW played version, and 52 ms in the FAST played version. Silent intervals lasted 44 ms in the first two passages of each set, and 52, 62.5, 74, 88, 110, and 177 ms in the seven remaining passages. Twelve subjects each shadowed both sets of passages, one in the SLOW version and one in the FAST version, with order of presentation and passage sets appropriately counterbalanced. The subject was told in detail how his shadowing performance would be used to assess intelligibility, and he was asked to try

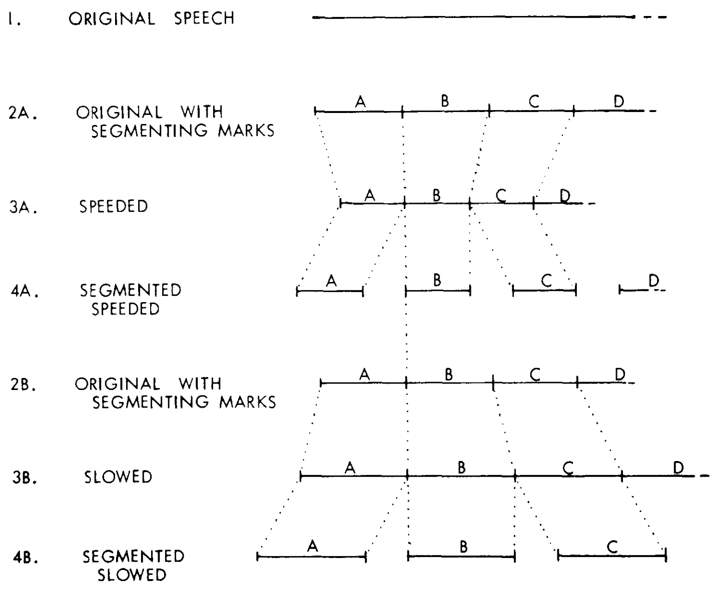


Fig. XVI-7. Experimental design.

to maximize his score. He was warned against trying to shadow with very short delays because most subjects find this very difficult, and against long delays, because of the risk of forgetting several stored-up words upon encountering something unintelligible. He was told that what he said did not have to make sense – he would score 50% either if he repeated every word in the first half of the passage, or if he repeated alternate words all through. He was given several minutes practice of shadowing undegraded speech (taken from the same master tape), and the first two passages in each set (44 ms silent intervals) were regarded as practice for segmented speech. The subject's performance was recorded for scoring: one point was earned for each word correctly repeated from the middle 100 words of each 150-word passage. The remaining 50 words were discarded, from the beginning and the end of the passage, to eliminate start-up and recency effects.

2. Results and Discussion

The mean shadowing scores are plotted twice in Fig. XVI-8, so that their fit to two hypotheses can be compared. If the gap-bridging effect is independent of speech rate, then the ABSOLUTE duration of the silent intervals is the critical parameter controlling the recovery of intelligibility, and the two sets of data, labeled FAST and SLOW, should be in good agreement in the upper part of Fig. XVI-8. Here the scores are plotted against the absolute duration of the silent intervals, as they were presented to the subjects. If, on the other hand, the critical parameter is the duration of the silent intervals relative to the duration of the speech events, then the two sets of data should be in good agreement as they are plotted in the lower part of Fig. XVI-8. Here the scores are plotted against the silent interval durations that would have been heard, had the speech and silent intervals been played back at a speed that RESTORED speech interval duration to 62.5 ms.

Clearly, the two sets of data are in better agreement in the upper part than in the lower part. Therefore, if the two options are (i) ABSOLUTE: the recovery of intelligibility is independent of speech rate, and takes place at the same silent interval duration when speech rate is increased, or (ii) RELATIVE: the critical silent intervals are a fixed proportion of phoneme duration, so that recovery takes place at shorter silence intervals when speech rate is increased, we must clearly choose the former. The two sets of data are not in good agreement with either hypothesis. Inspection of Fig. XVI-8 suggests that the recovery, rather than taking place at shorter silence intervals when speech rate is increased, occurs at longer silence intervals. That is, the FAST data lie to the left of the SLOW data in Fig. XVI-8, rather than to the right, as would be required by the RELATIVE hypothesis. To test whether the lengthening of the silent intervals from SLOW to FAST data was compensating for the shortening of the speech intervals caused by the change in playback speed, the two sets of data are plotted in another way in Fig. XVI-9. Here the two sets of data are plotted against duration in such a way as to equate the duration between successive onsets (or offsets) of speech intervals across the FAST and SLOW versions. In fact, the plotted silent intervals were ADJUSTED for the FAST data by increasing the true silence duration by the decrement in speech interval duration resulting from playing it FAST, and for the SLOW data by decreasing the true silence duration by the increment in speech interval duration resulting from playing it SLOW. The two sets of data seem to be in better agreement in Fig. XVI-9 than in the upper part of Fig. XVI-8, which suggests that the critical parameter controlling the recovery of intelligibility in gap-bridging is not the time between the end of a speech interval and the start of the next, but rather the time from the start of a speech interval to the start of the next.

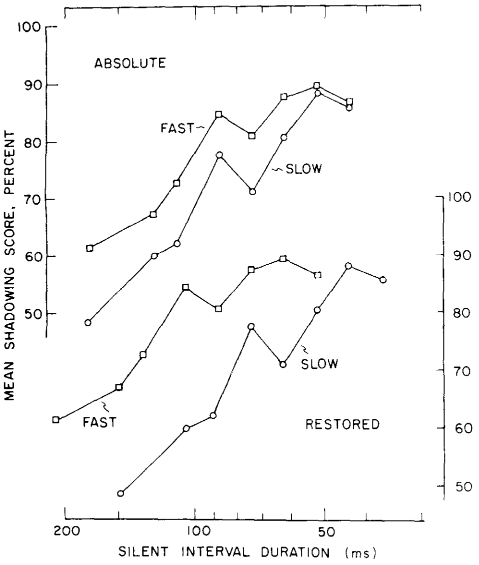


Fig. XVI-8.

Mean shadowing scores plotted against ABSOLUTE silent interval duration (upper) and against RESTORED duration relative to speech rate (lower).

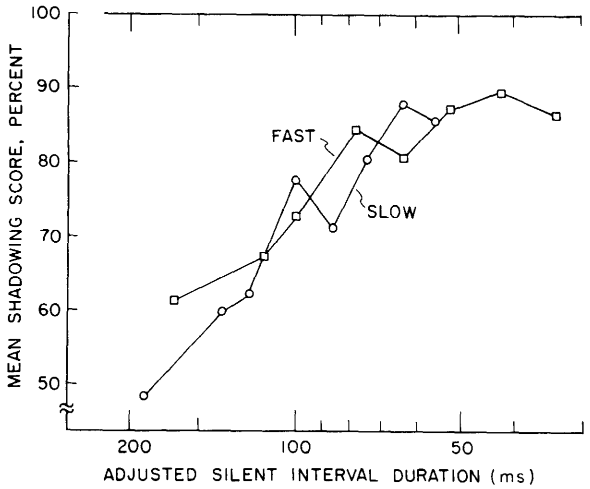


Fig. XVI-9.

Mean shadowing scores plotted against silent interval duration ADJUSTED to equate time from onset to onset of speech intervals.

Before exploring briefly the implications of this result, we should note that it was obtained by post hoc analysis, and therefore needs to be confirmed. The data are not as smooth and regular as might be expected from the results of earlier experiments using the same materials, and it is not obvious whether smoother data would improve or worsen the agreement in Fig. XVI-9. It is possible that the inversion in both sets of data [middle two data points in Fig. XVI-8 (upper)] might disappear if the number of subjects were increased. Otherwise, the inversion adds to the need for a confirming experiment, with better counterbalancing of the speech materials.

To return to the result obtained from Fig. XVI-9: Suppose we are correct in concluding that the critical parameter controlling gap-bridging is the time interval between the onsets of successive speech intervals. What are the implications of this finding? The fact that the two sets of data were brought into agreement by plotting them against a single temporal parameter for two different speech rates suggests that the recovery of intelligibility may reflect some relatively fixed temporal property of auditory processing at a stage more primitive than speech, or phonetic processing. There is some evidence that "acoustic" and "phonetic" stages of processing are separate. For example, Wood[20] showed that reaction time in classifying an acoustic parameter such as fundamental frequency is not affected by irrelevant phonetic variation such as place of articulation on a continuum between /b, d/. On the other hand, reaction time in classifying place of articulation is increased by irrelevant variation in fundamental frequency. Reaction times are fastest when phonetic and acoustic variations are correlated.[21] Wood argues from these results that acoustic and phonetic processing proceed in parallel, with the results of acoustic processing being available earlier than those of phonetic processing.

If the foregoing arguments are correct, and phonetic processing does not make use of the results of acoustic processing, then any temporal parameter that affects both acoustic and phonetic processing must be more primitive than either.

Moreover, the fact that the temporal parameter underlying the recovery of intelligibility corresponds to the sum of speech and silent interval durations suggests a storage mechanism with a fixed capacity like Archimedes' bath: The more speech the storage contains, the less silence is required to fill it up. I have argued elsewhere that raw "echoic" storage in the auditory system might have some of the properties of a delay line.[7] The interpretation of the present results appears to be compatible with this concept.

References

1. M. Wertheimer, "Experimentelle Studien über das Sehen von Bewegung," Z. Psychol. 61, 161-265 (1912).
2. J. Stroud, "The Fine Structure of Psychological Time," in H. Quastler (Ed.), Information Theory in Psychology (Free Press, New York, 1955).
3. G. Sperling, "The Information Available in Brief Visual Presentations," Psychol. Monogr., Vol. 74, No. 11, 1960.
4. E. Averbach and A. S. Coriell, "Short-Term Memory in Vision," Bell Syst. Tech. J. 40, 309-328 (1961).
5. U. Neisser, Cognitive Psychology (Appleton-Century Crofts, New York, 1967).
6. D. W. Massaro, "Preperceptual Images, Processing Time, and Perceptual Units in Auditory Perception," Psychol. Rev. 79, 124-145 (1972).

7. A. W. F. Huggins, "Temporally Segmented Speech," Percept. Psychophys. 18, 149-157 (1975).

8. G. A. Miller and G. A. Heise, "The Trill Threshold," J. Acoust. Soc. Am. 22, 637-638 (1950).

9. R. M. Warren, C. J. Obusek, R. M. Farmer, and R. P. Warren, "Auditory Sequence: Confusions of Patterns Other than Speech or Music," Science 164, 586-587 (1969).

10. I. B. Thomas, P. B. Hill, F. S. Carroll, and B. Garcia, "Temporal Order in the Perception of Vowels," J. Acoust. Soc. Am. 48, 1010-1013 (1970).

11. A. S. Bregman and J. Campbell, "Primary Auditory Stream Segregation and Perception of Order in Rapid Sequences of Tones," J. Exptl. Psychol. 89, 244-249 (1971).

12. P. Ladefoged and D. E. Broadbent, "Perception of Sequence in Auditory Events," Quart. J. Exptl. Psychol. 12, 162-170 (1960).

13. R. M. Warren, "Perceptual Restoration of Missing Speech Sounds," Science 167, 392-393 (1970).

14. J. Bastian, "Silent Intervals as Closure Cues in the Perception of Stop Phonemes," Quarterly Progress Report No. 33, Haskins Laboratories, New Haven, Connecticut, 1960, see Appendix I.

15. G. A. Miller and J. C. R. Licklider, "The Intelligibility of Interrupted Speech," J. Acoust. Soc. Am. 27, 167-173 (1950).

16. E. C. Cherry and W. K. Taylor, "Some Further Experiments on the Recognition of Speech, with One and with Two Ears," J. Acoust Soc. Am. 26, 554-559 (1954).

17. A. W. F. Huggins, "Distortion of the Temporal Pattern of Speech: Interruption and Alternation," J. Acoust. Soc. Am. 36, 1055-1064 (1964).

18. A. W. F. Huggins, "More Temporally Segmented Speech: Is Duration or Speech Content the Critical Variable in Its Loss of Intelligibility?" Quarterly Progress Report No. 114, Research Laboratory of Electronics, M. I. T., July 15, 1974, pp. 185-193.

19. A. Wingfield and J. Wheale, "Word Rate and Intelligibility of Alternated Speech" (to appear in Percept. Psychophys.).

20. C. C. Wood, "Levels of Processing in Speech Perception: Neurophysiological and Information-Processing Analyses," Status Report on Speech SR-35/36, Haskins Laboratories, New Haven, Connecticut, 1973.

21. C. C. Wood, "Parallel Processing of Auditory and Phonetic Information in Speech Perception II: Application of a Probabilistic Model," J. Acoust. Soc. Am. 55, S88 (1974).

C. PERCEPTUAL IMPORTANCE OF THE SECOND FORMANT DURING RAPID SPECTRUM CHANGES

National Institutes of Health (Grant 5 RO1 NS04332-13)

Dennis H. Klatt, Stefanie R. Shattuck-Hufnagel

1. Introduction

This study began as an attempt to see whether mirror-image acoustic patterns are perceived as similar. Our motivation arose from observations of formant patterns in prevocalic and postvocalic stop consonants. Figure XVI-10 illustrates typical vowel-consonant and consonant-vowel formant transitions for /b, d, g/. The three lowest formants fall in frequency during the closure gesture for /b/, that is, postvocalically, and the three formants rise in frequency during release. The first formant follows the same pattern for /d/ and /g/, but the second- and third-formant motions differ. The second and third formants usually rise during closure for /d/ and fall during release. The second and third formants come together during closure for /g/ and diverge on release.

In each case illustrated in Fig. XVI-10, the formant pattern at closure is the mirror image of the release pattern. These formant transitions are important cues to place of articulation for plosive consonants.[1, 2] Formant transitions are the primary cues to place of articulation for nasals because there is no frication burst at release, and the nasal murmur portions of /m, n, n/ are perceptually ambiguous.[3] Formant transitions are also the primary cues for final unreleased stops.[4] A plosive is normally not released in English if it is the first element in a sequence of two plosives, such as the /g/ in the words "big doll."

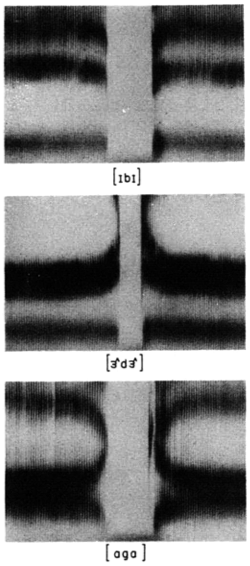


Fig. XVI-10.

Broadband spectrograms of the motions of the three lowest formant frequencies for intervocalic /b, d, g/.

The test for mirror-image similarity failed. We found that mirror-image acoustic patterns representing the same place of articulation were judged less similar to each other than to patterns representing different places of articulation.[5, 6] The results suggest that the child who has mastered the production and perception of [ba] will not automatically assign the consonant in [ab˺] to the same perceptual category. The phonetic identity of [ba] and [ab˺] must be learned, presumably by associating the two allophones with the same articulatory gesture, as in [aba].

To our surprise, similarity judgments were based on acoustic similarities in the frequency region of the second formant, even when components in this region were reduced in amplitude by as much as 20 dB relative to higher frequency rapid spectrum changes. This report reviews these results and presents new data on the origin of the effect.

2. Experimental Paradigm

In order to avoid the difficulties of using natural speech stimuli or speech responses, stimuli composed of two simultaneous pure tones of rapidly changing frequency are used.[7, 8] These brief pure-tone glissandos are subjectively nonspeechlike. Therefore similarity judgments can be obtained that are relatively uncolored by adult linguistic biases.

Pure-tone glissandos are very similar to formant transitions with respect to short-term spectral composition. A formant frequency transition in a consonant-vowel syllable and a pure-tone glissando with the same frequency transition have in common a rapid spectrum change such that the center frequency of an energy concentration shifts rapidly in time. While a formant transition and glissando differ in other acoustic dimensions, they share the aspect of rapid spectrum change,[9] which is probably the most important aspect for perception of place of articulation.

The stimulus ensemble for a typical experiment is shown in Fig. XVI-11. There are 4 stimuli, each having two 40-ms glissandos. A glissando is defined here as a brief tone burst with linearly changing frequency throughout its duration. The frequency of a glissando component varies about an average frequency $\overline{F}$ by an amount ΔF, which was set to 350 Hz in Experiment 1. This frequency is a moderately large motion for a stop-vowel formant transition.

The amplitudes of glissando components are varied from subtest to subtest, but in a particular subtest involving 4 stimuli, the amplitude of the higher frequency glissando component A_H and the amplitude of the lower frequency component A_L are each fixed. The more intense component is set to produce 70 dB peak SPL at the output of a pair of binaural headphones.

The first stimulus in Fig. XVI-11 is an idealization of the motions of the second and third formants during the 40 ms following release of a /g/ before a vowel such as /Λ/. The second stimulus is a mirror image of the first, and corresponds to a postvocalic /g/ pattern. The third stimulus in Fig. XVI-11 resembles the second- and third-formant transitions for prevocalic /b/, and the fourth stimulus resembles postvocalic /b/ if the vowel is a front vowel such as /I/. Stimuli 3 and 4 also resemble second- and third-formant transitions for postvocalic and prevocalic /d/ if the vowel is a back vowel such as /o/ or /ɝ/.

A listening test using these four stimuli was generated by computer and recorded on audio tape for playback to a number of subjects. In each trial, three out of the four stimuli were presented in random order with 250-ms silent interstimulus intervals. Listeners were asked to indicate which of the three stimuli was most different from the other two.

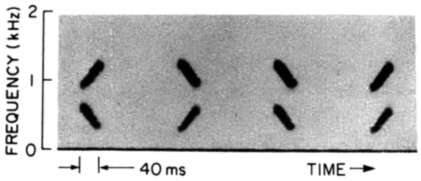


Fig. XVI-11.

Narrow-band spectrograms of the four stimuli used in one of the subtests of Experiment 1.

Figure XVI-12 illustrates two sample trials from the experiment. If a listener judged stimulus B to be most different in either of these trials, he would be making a mirror-image response, since stimuli A and C have no common acoustic attributes except that they are mirror images of one another. Stimuli A and B have one common glissando component and so do stimuli B and C. Therefore an A or C response may be interpreted as being caused by the perceptual dominance of the higher or lower frequency component.

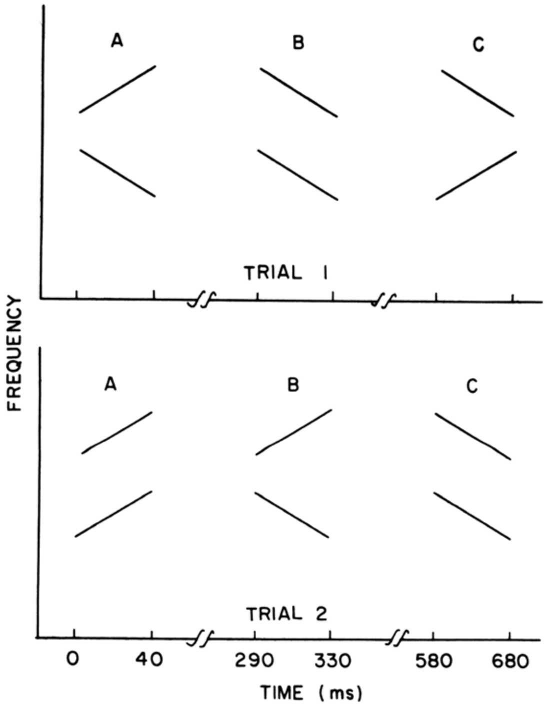


Fig. XVI-12.

Examples of two trials in the experiments involving judgments of relative similarity among 3 stimuli. Stimulus C is usually judged most different in trial 1 and stimulus A most different in trial 2. A B response would be a mirror-image response in both trials.

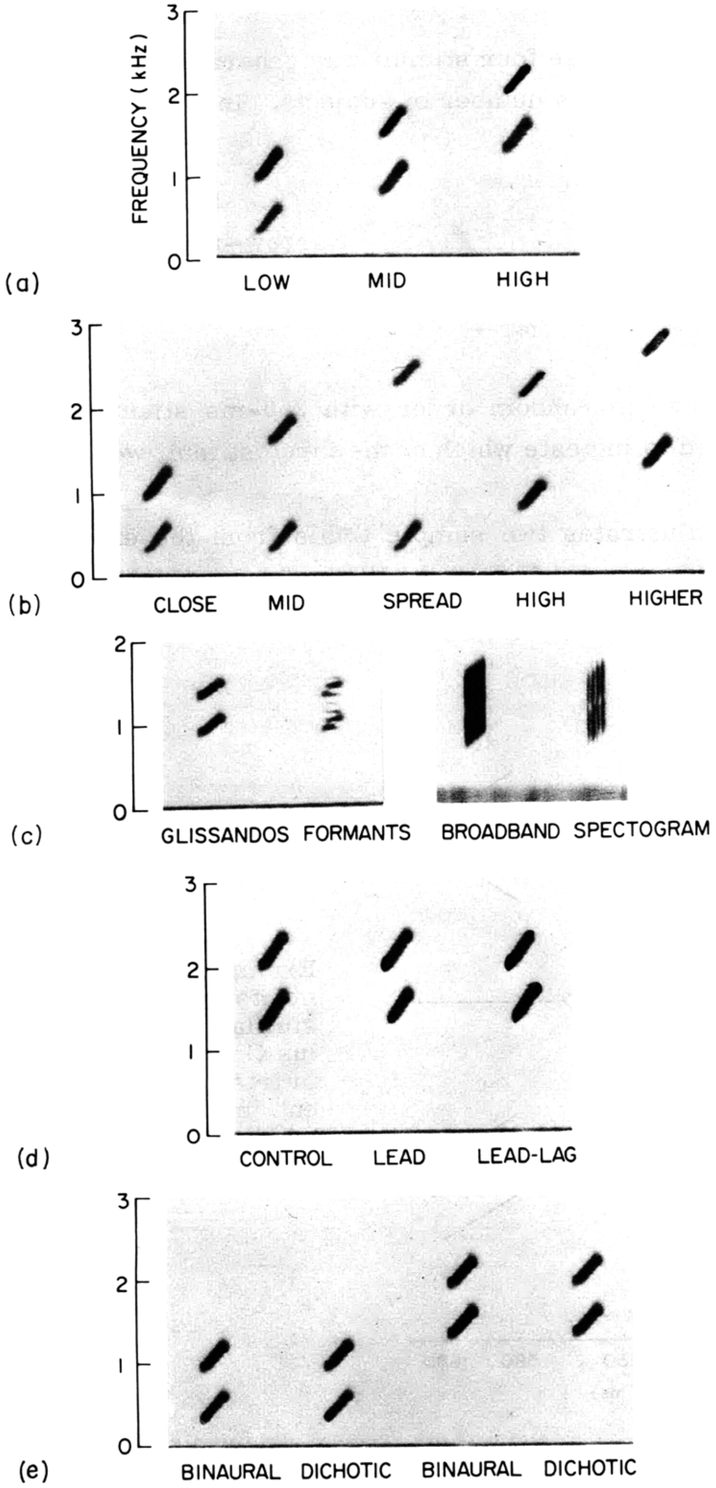


Fig. XVI-13. One of the four stimuli is shown for each condition of each of the experiments involving judgments of dissimilarity.

Experiment 1: Dominance of the Second-Formant Region

In the first experiment dissimilarity judgments were obtained for stimuli drawn from 3 different frequency ranges. The frequency values used in these subtests are illustrated in Fig. XVI-13a. Spectrograms are shown in Fig. XVI-13 of one of 4 stimuli used in a subtest in which the amplitudes of the two components were roughly equal. The four stimuli corresponding to the example labeled "LOW" in Fig. XVI-13a have average frequencies analogous to the first and second formants. Glissando components labeled "HIGH" in Fig. XVI-13 have component frequencies that correspond to the second and third formants, and components labeled "MID" provide an intermediate case. For each of these cases, 6 subtests were designed in which the relative amplitudes of A_H and A_L were varied systematically.

Before listening to the experimental tape, subjects were given a 30-trial pretest to determine whether they were capable of making the required dissimilarity judgment when only one glissando component is present. Among 18 listeners, 10 performed the task perfectly and 7 had scores ranging from 96% to 62% correct. One subject produced essentially random responses (37% correct) and his data were discarded.

Results from the remaining listeners when presented with the two-component stimuli revealed that there were very few mirror-image responses. (The mirror-image hypothesis will be discussed in the sequel.) The average responses of 17 subjects for one of the conditions of Experiment 1 are plotted in Fig. XVI-14. The horizontal axis corresponds to the relative levels of the two glissando components A_H/A_L on the subtests involving $\overline{F}_L$ = 475 Hz. The vertical axis indicates the percentage of trials on which the lower frequency glissando component, the higher frequency glissando component, or mirror-image symmetry formed the basis of the dissimilarity decision.

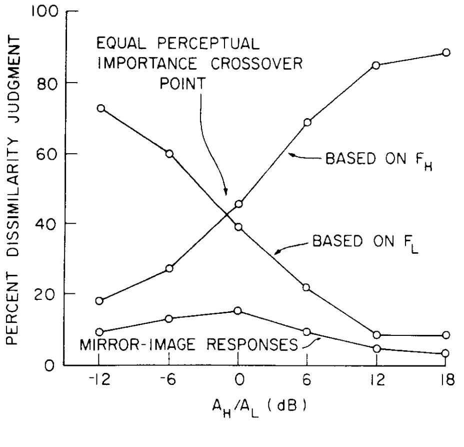


Fig. XVI-14.

Responses of 17 subjects to trials of Experiment 1 for $\overline{F}_L$ = 475 Hz plotted as a function of the relative level of the two components A_H/A_L. As indicated, dissimilarity judgments can be based on F_L, F_H, or mirror-image symmetry.

The value of A_H/A_L for which the curves labeled "BASED ON F_L" and "BASED

ON F_H" intersect is called the equal-perceptual-importance crossover point. The amplitude ratios corresponding to equal perceptual importance in each experiment reported here are summarized in Table XVI-1.

The most interesting aspect of the data is the observation that the equal-perceptual-importance crossover points occur at dramatically different amplitude ratios A_H/A_L of -1, 13, and 26 dB. This means that when F_L and F_H are in the frequency range of the first and second formants, the two components are of equal perceptual importance when they are roughly of equal amplitude. When the frequencies of the two components are raised, however, such that $\overline{F}_L$ = 1475 (i. e., when the two components correspond to the second and third formants), the component analogous to the second formant completely dominates the judgment. The F_L component must be lowered 26 dB in relative amplitude before the two components are of equal perceptual importance.

There was considerable variability in the self-consistency with which subjects performed the task and in the individual equal-perceptual-importance crossover points. Good subjects had very steep curves with crossover values that differed less than 3 dB from day to day. Subjects who were less proficient had rather irregular flat curves with poorly defined crossover points. Figure XVI-15 shows dissimilarity judgment data from a good and an inconsistent subject on those subtests in which $\overline{F}_L$ = 475 Hz. Subject MC produced several mirror-image responses, which suggested that he was guessing most of the time for those conditions on the left side in the figure.

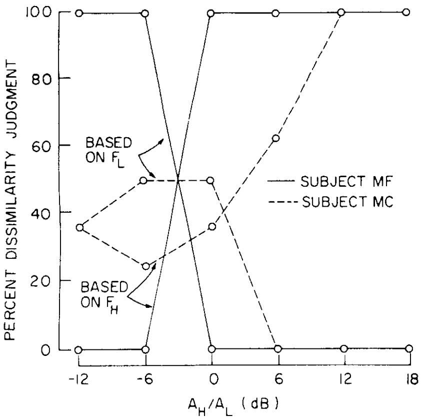


Fig. XVI-15.

Data of one good and one inconsistent subject are compared as in Fig. XVI-14. Mirror-image responses are not shown.

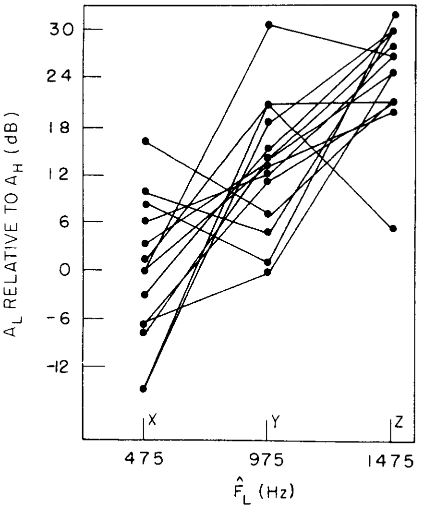


Fig. XVI-16.

Value of A_H/A_L at which the two components are of equal perceptual importance plotted for 14 subjects as a function of the average frequency of the lower component in the first experiment.

The equal-perceptual-importance crossover points of 14 subjects who had unique crossover points on low, mid, and high tests are plotted in Fig. XVI-16. Most of the subjects follow the group trend of increasing dominance of F_L as F_H is increased; those individuals who do not follow this trend are subjects whose data are less self-consistent.

Evidence for Frequency-Change Detectors

Analysis of the mirror-image responses to stimulus triads revealed a surprising pattern. Not all of the possible random orders of stimulus triads received an equal number of mirror-image responses. Stimulus pairs corresponding to /g/-like transitions received over 85% of the mirror-image responses. In no case was the total number of mirror-image responses above chance levels; hence, it appears that subjects are not equating mirror-image patterns, but rather are being confused more often when a stimulus contains both a rising and a falling glissando.

It is appealing to speculate that there are both rising and falling frequency detectors in the central nervous system, and that /g/-like transitions excite units in both the rising and falling detector populations. In order to account for the confusability of /g/-like patterns, we must postulate further that the frequency-change detectors are only moderately frequency specific[10] or that glissandos can be interchanged as represented in short-term auditory memory. Our /b/-like and /d/-like stimuli with both rising or both falling glissandos would excite only the rising or falling detectors, and thus would be less confusable as represented in short-term auditory memory.

Experiment 2: F_L and F_H More Separated in Frequency

Additional experiments were run at other values of F_L and F_H in order to examine more carefully the interactions between competing glissandos. Examples of one of the stimuli from each of these experiments are shown in Fig. XVI-13b. The results are summarized in Table XVI-1.

The perceptual equivalence crossover values of A_H/A_L for an $\bar{F}_L$ of 450 Hz were 3, 18, and 18 dB as $\bar{F}_H$ was increased from 1050 to 1750 to 2450 Hz.

The new data are consistent with the results of Experiment 1: As F_H increases in frequency beyond the region of the second formant, its perceptual importance wanes. This shows clearly that we are not dealing with a traditional masking phenomenon: increasing the separation between the two components decreases the perceptual importance of the F_H glissando.

Experiment 2 also included conditions where the separation between the two glissando components was fixed at 1300 Hz while the frequency of both components was varied. As indicated in Table XVI-1, the perceptual equivalence crossover point of A_H/A_L increased slightly from 18 to 24 to 27 dB as $\bar{F}_L$ went from 450 to 950 to 1450 Hz.

These results are consistent with a model in which the perceptual importance of a glissando decreases gradually as its average frequency increases above approximately 1200-1500 Hz. This could reflect the frequency distribution of property detectors sensitive to rapid spectrum changes of this type. Rising and falling frequency detectors are postulated to have an approximately uniform distribution below ~1500 Hz and to fall off gradually in number above this frequency.

Experiment 3: Pure-Tone Glissandos vs Formant Transitions

The third experiment was a control to see if there would be any change in the results if the pure-tone glissandos were replaced by speechlike synthetic formant patterns. Formant glissandos were generated by sending a 100-Hz impulse train through two digital formant resonators whose resonant frequencies were varied rapidly by the same linear change as the glissandos. The bandwidths of the two competing formants were set to 50 Hz and the duration of the formant patterns was adjusted to 40 ms.

Examples of one of the stimuli from the control and formant conditions are shown in Fig. XVI-13c, and the results are summarized in Table XVI-1. The equal-perceptual-importance crossover point changed only 3 dB, thereby reinforcing our conjecture that formant transitions and pure-tone glissandos share similar rapid spectrum change features.

Experiment 4: Effect of Temporal Offsets (F_H Leads and/or F_L Lags)

It was suggested by Stevens[9] that the burst of frication noise at plosive release may combine with a following formant transition to produce an effective formant transition with a larger rapid spectrum change than if the burst were not present. For example, the burst for /d/ is thought to combine with the third or fourth formants to produce the sensation of a generally falling spectrum change even if the second formant is rising, as it is in /di/ or /dI/.

A set of experiments was designed to see whether an F_H glissando corresponding in frequency to the third formant would have greater perceptual importance if it began 10 ms earlier. Examples of one of the stimuli from the experimental and control condition are shown in Fig. XVI-13d, and the results are summarized in Table XVI-1. When the 10-ms lead in F_H is produced by a 10-ms cutback in the onset of F_L there is little or no change in the perceptual importance of F_H.

The Stevens hypothesis is not confirmed; the higher frequency component does not increase in perceptual salience if it starts early. Our results indicate that in a syllable such as /di/, the perceptually dominant spectrum change will be the rising second formant. A /bi/ and a /di/ cannot be distinguished on the basis of dominant rapid spectrum change. Only the presence of a high-frequency burst of frication noise and the cluster of energy at the fourth formant at voicing onset differentiate /di/ from /bi/.

Table XVI-1. Ratio of glissando amplitudes A_H/A_L at which the two components of a stimulus are of equal perceptual strength tabulated as a function of the average frequency $\bar{F}_L$ of the lower frequency component, the average frequency $\bar{F}_H$ of the higher frequency component, and the magnitude of the glissando frequency change ΔF. Each experiment involved a different set of listeners.

	$\bar{F}_L$	$\bar{F}_H$	ΔF	A_H/A_L
Experiment 1. Main effect				
low frequencies	475	1125	350	-1
mid frequencies	975	1625	350	13
high frequencies	1475	2125	350	26
Experiment 2. Greater separation between the two components				
control	450	1050	300	3
F_H higher	450	1750	300	18
F_H still higher	450	2450	300	18
F_H and F_L higher	950	2250	300	24
still higher	1450	2750	300	27
Experiment 3. Formants vs glissandos				
control	1000	1400	200	5
formants	1000	1400	200	8
Experiments 4. F_H glissando begins early				
control	1500	2200	400	16
F_H leads	1450	2200	300,400	14
F_H leads, F_L lags	1450	2150	300	31
Experiment 5. Dichotic vs binaural presentation				
control Fl-F2	500	1100	300	2
dichotic Fl-F2	500	1100	300	(see text)
control F2-F3	1500	2100	300	12
dichotic F2-F3	1500	2100	300	(see text)

A second condition was included in the lead experiment in which the F_H lead was produced simply by delaying the F_L pattern 10 ms. In this case F_H leads and F_L lags. The lagging component increased in perceptual importance from 16 dB (simultaneous onset and offset) to 31 dB (F_H leads and F_L lags). The lagging component probably derives its increased salience from a pitch effect rather than from anything having to do with the rapid spectral changes in the stimulus. This interpretation will be discussed later.

Experiment 5: Dichotic vs Binaural Presentation

If the two competing glissandos are presented dichotically, that is, F_L to one ear and F_H to the other, then it is possible to test for the central/peripheral origin of observed interactions. Dichotic and binaural presentation modes were compared for two sets of glissando frequencies, as shown in Fig. XVI-13e. The results are summarized in Table XVI-1. Binaural crossover points were as expected from previous tests, but the dichotic responses were virtually insensitive to the relative levels of the two components. For most subjects, the higher frequency glissando determined the response for all relative A_H/A_L levels tested, but some subjects seemed to alternate randomly between attending to one or the other component.

The results of this experiment suggest that the dominance of a glissando in the second-formant region is due to neural interactions occurring prior to a point where inputs from the two ears are combined. If the effect is due to the distribution of frequency-change detectors, it would appear that these detectors must exist in nuclei that receive inputs primarily from one ear.

Experiment 6: Pitch Effects

Brady, House, and Stevens[11] investigated the perception of pitch and quality in harmonic complexes characterized by a rapidly changing resonant frequency. They found that judgments of the pitch or quality of an energy concentration analogous to a formant tended to skew away from the mean frequency and toward the terminal frequency.

The perceptual prominence of the final frequency has been shown to apply also to pure-tone glissandos.[12, 7] For example, the perceived pitch of a 40-ms glissando with an average or midpoint frequency of 730 Hz and an incremental frequency change of +300 Hz was found by Nabelek and Hirsh[7] to be ~820 Hz. There was a fairly large variance in the pitch judgment data (standard deviation of 38 Hz in this example). The authors argue that pitch judgments migrate toward the final frequency on the glissando and have a large variance whenever the average spectrum of a single glissando contains two or more maxima.

It is possible that in our experiments subjects are responding to differences in perceived pitch between rising and falling glissandos, rather than the rising vs falling feature, as we have argued up to this point. This kind of explanation would not account in itself for the perceptual importance of the second-formant region, but would considerably inhibit attempts to relate our findings to the perception of rapid spectral changes in speech.

Therefore we are examining glissando competition in stimuli with and without compensation for expected pitch differences in an experiment that is now in progress. Preliminary results indicate that pitch compensation, that is, where falling glissandos have an appropriately higher average frequency than rising glissandos, does not decrease the dominance of the F_L component. Whatever role pitch differences may play in our earlier experiments, the pitch effect seems unrelated to the dominance of rapid spectral changes in the second-formant region.

Experiment 7: Frequency Discrimination, Loudness, and Masking

The results shown in Table XVI-1 cannot be explained easily in terms of basic psychophysical data on constant-frequency pure tones. Equal-loudness contours for a pure tone presented at 70 dB peak SPL are nearly flat in the 300-3000 Hz frequency range.[13] Thus frequencies in the second-formant region are not inherently louder.

Pure-tone thresholds are lowest at 2-3 kHz frequency. Therefore the high-frequency glissandos in our experiments are most intense relative to threshold. Glissandos below 1500 Hz dominate the dissimilarity results, however, so physical intensity within the central nervous system does not explain our results.

Tone-on-tone masking functions indicate that a lower frequency tone can mask detection of the presence of a nearby higher frequency tone. But the masking functions do not change shape as masker and masked tones are shifted from the frequency region for the first formant to that of the second or third formant,[14] and masking does not occur for the relative intensities used in the dissimilarity judgment experiments.

The JND for frequency of a pure tone does not increase suddenly at 1500 Hz. Constant-frequency pure tones with a 40-ms duration can be distinguished if they differ in frequency by ~3%.[15] There are some data[16,7] on the discrimination of frequency transitions, but results on the threshold for detection of DIRECTION of a frequency change in a brief glissando are not available.

Several frequency-discrimination experiments were designed in an attempt to determine whether glissando transitions are more difficult to detect and/or categorize as rising or falling when the average frequency of the glissando is increased. We first presented pairs of 40-ms single-component stimuli that were a rising-frequency glissando followed by a falling-frequency glissando, or vice versa. Listeners were asked to indicate rise-fall or fall-rise in each trial.

The results indicate that a just noticeable difference (JND) (rise-fall vs fall-rise judgments correct 75% of the time) occurred at a value of ΔF of 15 Hz at an average glissando frequency of 500 Hz. There was only a small increase in the JND for glissando direction as the average frequency of the glissando was increased from 500 Hz to 2000 Hz. We conclude that the similarity judgment experiments employed frequency increments that were significantly larger than the JND.

In a second experiment the JND for slope of a glissando was determined with and without pitch compensation. Rise-fall vs fall-rise judgments were obtained as the ΔF decreased. The task was much harder in the pitch-compensated case, which indicated that pitch is used by listeners in this type of judgment. The pitch-compensated JND for rise-fall vs fall-rise discrimination occurred at a ΔF of 50 Hz, which is still considerably smaller than the ΔF employed in the dissimilarity experiments.

In both of these JND experiments there were several examples of blocks of trials in which a subject produced a consistent sequence of response reversals when confronted with a new ΔF. He could discriminate between a rise-fall and a fall-rise pattern but could not absolutely identify rise vs fall. Data from 5 of 15 subjects had to be discarded because of the presence of one or more blocks of response reversals. Therefore the discrimination JND that we observe may not be a very accurate measure of rise-fall absolute IDENTIFICATION ability.

3. Discussion

The experiments that we have described concern the perception of brief rapidly changing energy concentrations. Our results point to the existence of populations of rising and falling frequency detectors, the majority of which are concentrated in the frequency region of the first and second formant, i. e., below ~1500 Hz. The frequency-change detectors appear to be located in a nucleus of the central nervous system that receives inputs only from one ear.

Several alternative explanations involving frequency discrimination, loudness, masking, and pitch differences have been analyzed and rejected. Future experiments are planned to see whether the observed effects are sensitive to signal level, and to find out what happens when 3 glissando components are present in a stimulus to simulate interactions between the first, second, and third formants.

Stevens has argued[17] that consonants are characterized as a class by the presence of a rapid spectrum change, particularly at the onset or offset of a change in speech signal level. The spectra of vowels are stationary or slowly changing. The rapid spectrum-change detectors that have been discussed could play an important role in the determination of manner of articulation and in differentiating vowels from consonants.

Stevens[9] has postulated the existence of simple property detectors in the human auditory system that are sensitive to rising or falling rapid spectral changes. The most common places of articulation (labial, coronal, and velar) are thought to be differentiated on the basis of spectral changes above ~800 Hz and over a time interval of ~20-40 ms.

Our results suggest that the distribution of these detectors has a greater concentration at frequencies near or below the normal range of the second formant, and the detectors are broadly tuned to respond to any frequency change within a relatively wide bandwidth. According to our results, third-formant transitions must be of less perceptual importance than transitions in F2 and F1, especially if F3 transitions are competing with first- or second-formant transitions.

From the very earliest days of speech perception research, second-formant loci have been implicated as of major importance in the determination of place of articulation for consonants.[18] Our results suggest the reason for this: the distribution of detectors of frequency change is concentrated in this region.

Our results indicate that a syllable such as /di/ will produce the sensation of a generally rising rapid spectrum change at release (above 800 Hz) because of a rising second-formant frequency, while a syllable such as /da/ will produce the sensation of a generally falling rapid spectrum change because of a falling second formant. Thus the generality that invariant rapid spectrum-change cues perceptually differentiate labial, alveolar, and velar plosives,[9] while usually true, is falsified logically by the /di/ counterexample.

The /di/, /da/, and /du/ share nearly the same onset frequencies for the second and third formants[19] and have similar frication bursts. Listeners must learn to pay more attention to the burst spectrum and intensity and the spectrum at voicing onset (and less attention to the rapid spectrum-change cues) if they are to equate perceptually the /d/s in /di/, /da/, and /du/.

References

1. A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. Gerstman, "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants," Psychol. Monogr. 68, No. 8, 1954.
2. E. Fischer-Jørgensen, "Perceptual Studies of Danish Stop Consonants," Annual Report of the Institute of Phonetics, University of Copenhagen, Vol. 6, pp. 75-176, 1972.
3. A. Malecot, "Acoustic Cues for Nasal Consonants," Language 32, 274-284 (1956).
4. W. S.-Y. Wang, "Transition and Release as Perceptual Cues for Final Plosives," J. Speech Hear. Res. 2, 66-73 (1959).
5. D. H. Klatt and S. R. Shattuck, "Perception of Brief Stimuli that Resemble Rapid Formant Transitions," in G. Fant and M. A. A. Tatham (Eds.), Auditory Analysis and Perception of Speech (Academic Press, London, New York, San Francisco, 1975), pp. 294-301.
6. S. R. Shattuck and D. H. Klatt, "The Perceptual Similarity of Mirror-Image Acoustic Patterns in Speech" (submitted to Perception and Psychophysics).
7. I. Nabelek and I. J. Hirsh, "On the Discrimination of Frequency Transitions," J. Acoust. Soc. Am. 45, 1510-1519 (1969).
8. I. Nabelek, A. K. Nabelek, and I. J. Hirsh, "Pitch of Tone Bursts of Changing Frequency," J. Acoust. Soc. Am. 48, 536-552 (1970).

9. K. N. Stevens, "The Potential Role of Property Detectors in the Perception of Consonants," Quarterly Progress Report No. 110, Research Laboratory of Electronics, M.I.T., July 15, 1973, pp. 155-168.

10. R. H. Kay and D. R. Mathews, "On the Existence in Human Auditory Pathways of Channels Selectively Tuned to the Modulation Present in Frequency-Modulated Tones," J. Physiol. 225, 657-677 (1972).

11. P. T. Brady, A. S. House, and K. N. Stevens, "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency," J. Acoust. Soc. Am. 33, 1357-1362 (1961).

12. J. M. Heinz, B. E. F. Lindblom, and J. Ch. K-G. Lindquist, "Patterns of Residual Masking for Sounds with Speech-like Characteristics," IEEE Trans., Vol. AU-16, No. 1, pp. 107-111, March 1968.

13. H. Fletcher and W. A. Munson, "Loudness, Definition, Measurement, and Calculation," J. Acoust. Soc. Am. 5, 82-108 (1933).

14. R. L. Wegel and C. E. Lane, "The Auditory Masking of One Pure Tone by Another and Its Probable Relation to the Dynamics of the Inner Ear," Phys. Rev. 23, 266-285 (1924).

15. B. J. Moore, "Frequency Difference Limens for Short-Duration Tones," J. Acoust. Soc. Am. 54, 610-619 (1973).

16. I. Pollack, "Detection of Rate of Change of Auditory Frequency," J. Exptl. Psychol. 77, 535-541 (1968).

17. K. N. Stevens, "The Role of Rapid Spectrum Changes in the Production and Perception of Speech," in L. L. Hammerich and R. Jakobson (Eds.), Form and Substance: Festschrift for Eli Fischer-Jørgensen (Akademisk Forlag, Copenhagen, 1971).

18. P. C. Delettre, A. M. Liberman, and F. S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am. 27, 769-773 (1955).

19. D. H. Klatt, "The Acoustic Characteristics of English Plosives" (in preparation for publication).