Perception & Psychophysics
1974, Vol. 15, No. 2, 215-220

Auditory evoked potential correlates of speech sound discrimination*

MICHAEL F. DORMAN†
University of Connecticut, Storrs, Connecticut 06268

An electrophysiological correlate of the discrimination of stop consonants drawn from within and across phonetic categories was investigated by an auditory evoked response (AER) technique. Ss were presented a string of stimuli from the phonetic category [ba] (the standard stimulus) and were asked to detect the occurrence of a stimulus from the same phonetic category (within-category shift), or the occurrence of a stimulus from a different phonetic category [pa] (across-category shift). Both the across- and within-category shift stimuli differed equally from the standard stimulus in the time of onset of the first formant and in the amount of aspiration in the second and third formants. The N1P2 response of the AER was larger to the across-category shift than to the within-category shift. The within-category shift did not differ from a no-shift control. These findings suggest (1) that the AER can reflect the relative discriminability of stop consonants drawn from the same or different phonetic categories in a manner similar to other behavioral measures; (2) that the detailed acoustic representation of stop consonants is transformed into a categorized phonetic representation within 200 msec after stimulus onset.

Numerous studies have indicated that the sounds of speech enjoy a special mode of perception, distinct from that of nonspeech signals (Liberman, 1970; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). One set of investigations supporting this view has examined the relationship between identification and discrimination of speech and nonspeech signals. Listeners can discriminate many more nonspeech stimuli than they can identify absolutely (Miller, 1956; Pollack, 1952). However, certain speech sounds, the stop consonants [b, d, g, p, t, k], tend to be discriminated no better than they can be identified (Pisoni, 1971; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). This unique relationship between identification and discrimination is termed "categorical perception."

In a typical experiment, Lisker and Abramson (1970) presented to Ss for identification and discrimination a series of computer-synthesized stop consonants which differed solely along the physical continuum of voice onset time (VOT).[1] Listeners identified these stimuli exclusively as members of the phonetic category [ba] or [pa]. Ss discriminated almost perfectly between stimuli which were assigned to different phonetic categories. However, when physically different stimuli were drawn from the same phonetic category, discrimination was only slightly better than chance. Thus, equal acoustic differences (for example, 20 msec) along the VOT were not equally discriminable. Only when stimuli were drawn from different phonetic categories could listeners discriminate accurately between physically different stimuli. Discrimination, then, was constrained by a phonetic coding of the acoustic signal.

In contrast to the categorical perception of the stop consonants, nonspeech signals and steady-state vowels are perceived "continuously." Signals drawn from the same nonspeech or vowel category are discriminated equally well or poorly as signals drawn from different categories (Mattingly, Liberman, Syrdal, & Halwes, 1971).

The purpose of the present study was to determine whether components of the human cortical averaged auditory evoked response (AER) would reflect the categorical perception of different stop consonant signals or the equal physical differences between the different signals.

Very few studies have explored AERs to speech stimuli (Cohen, 1971; Greenberg & Graham, 1970; Wood, Goff, & Day, 1971). However, previous studies with nonspeech signals have indicated that when a target stimulus is detected in a signal detection task, the amplitude of both the N1-P2 and P300 components of the AER are larger than in response to nontarget stimuli (Davis, 1964; Karlin, 1970; Ritter & Vaughan, 1969; Sheatz & Chapman, 1969; Wilkinson & Lee, 1972). If the vertex AER responds to the discrimination of speech stimuli in a similar manner, then the AER to discriminably different stop consonant signals should be larger than the corresponding response to signals which are not discriminably different.

The use of the AER technique has another purpose which bears directly on the nature of categorical perception and its interpretation. It is possible that a listener may hear two physically distinct stimuli from within the same phonetic category as slightly different. However, because the listener knows that the two stimuli are both labeled the same in conventional speech and orthography, he may respond that the two stimuli are the same.

*This paper reports a portion of the research carried out for a PhD dissertation accepted by the University of Connecticut in 1971. The author wishes to thank F. Cooper and A. M. Liberman for making available the facilities of the Haskins Laboratories for the generation of the stimulus material. R. Hoffman and P. Morse are due acknowledgement for their part in the execution of this study. B. Karmel graciously allowed the use of his laboratory for the study.

†Currently at Haskins Laboratories, New Haven, Conn. 06510, and Herbert Lehman College of CUNY, Bronx, New York.

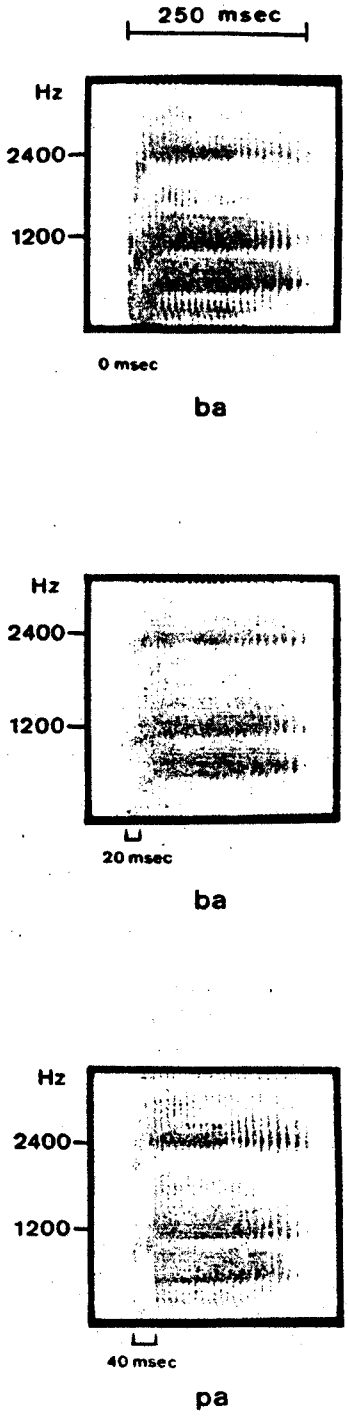

Fig. 1. Sound spectrograms of the speech stimuli 0 VOT [ba], 20 VOT [ba], and 40 VOT [pa].

In the context of the present study, an estimate of the time necessary to code the acoustic signal into a categorized phonetic description can be made by assessing whether the N1-P2 component of the AER reflects continuous or categorical perception. If the N1-P2 component reflects a categorical response, i.e., a larger response to the stimuli from a different phonemic category than to the stimuli from the same phonemic category, then within 200 msec after stimulus onset the acoustic signal has been recoded into a phonetic representation. This would suggest that a categorized phonetic coding is an immediate and obligatory transformation of the acoustic signal.

METHOD

Subjects

Fifty undergraduate students at the University of Connecticut served as Ss. No S had previously participated in research involving synthetic speech or electroencephalographic (EEG) recording.

Apparatus

The Ss sat in a comfortable chair within a dimly lit, electrically shielded room, and listened to tape-recorded stimuli presented via stereo headphones (Koss 600A). The sound level at the headphones was 65 dB.

Recording of the EEG was made from the scalp, using a single silver-disk electrode, located at the vertex, which was referenced to the right earlobe Resistance between electrodes was always less than 5K ohms.

The EEG signals were transmitted by telemetry (Narco FM-110-E3) to an ac preamplifier (W-P Instruments DAM 6) and oscilloscope amplifier (Tektronix RM 502A), which also served to monitor the EEG. The frequency response of the system after amplification was flat between 2.0 and 30 Hz. The amplified EEG was stored for later analysis using a Vetter FM adapter (FM-3) and a Sony 355 tape deck.

The extraction of the evoked response from the EEG was carried out both on- and off-line by a computer of average transients (Fabri-Tek 1072). The sweep duration was 1 sec. The averaging cycle of the computer was triggered by a pulse from the second channel of the stimulus presentation tape. The onsets of the cuing pulses and the synthetic speech stimuli were simultaneous. The AER records were written out on an X-Y plotter (Hewlett-Packard 7035b).

Stimuli

The three synthetic stop consonant-vowel syllables used in this study are shown in Fig. 1. These stimuli were generated on the Haskins Laboratories computer-controlled parallel-resonance synthesizer (Cooper & Mattingly, 1969).

The three stimuli differed solely along the VOT continuum: 0 msec VOT (0 VOT); 20 msec VOT (20 VOT); and 40 msec VOT (40 VOT). Stimulus duration was 250 msec. For Stimulus 0 VOT, the onsets of the first (F1), second (F2), and third (F3) formants were simultaneous; for Stimulus 20 VOT, F1 began 20 msec after F2 and F3; for Stimulus 40 VOT, F1 began 40 msec after F2 and F3. Aspiration was added to the upper formant frequencies during the period of F1 delay for Stimuli 20 and 40 VOT. Thus, each adjacent pair of stimuli along the VOT continuum differed by exactly 20 msec VOT (i.e., 20-0 VOT and 20-40 VOT). Previous identification studies have indicated that stimuli with 0 and 20 VOT are identified as members of the phonetic category [ba] and that the stimulus 40 VOT is identified as a member of the phonetic category [pa] (Lisker & Abramson, 1970).[2] Discrimination tests have indicated that the pair 20-40 VOT is discriminated essentially perfectly. The pair 20-0 VOT is discriminated just slightly better than chance (Abramson & Lisker, 1970). In the following account, Stimulus 20 VOT will be termed the "standard," Stimulus 0

VOT the "within-category" shift stimulus, and Stimulus 40 VOT the "across-category" shift stimulus.

Preparation of the Stimulus Tapes

With the aid of the computer-controlled synthesizer, four stimulus sequences were recorded on audio tape. Two of the stimulus sequences were composed of varying length runs of standard stimuli (20 VOT), separated by pairs of either within- or across-category shift stimuli. There was a total of 154 standard stimuli and 16 pairs of shift stimuli in each sequence. The pairs of shift stimuli occurred on the average once every 10 successive standard stimuli (range 6-14). In one sequence, the pairs of shift stimuli were within-category stimuli; in the other, they were across-category stimuli. A third stimulus tape consisted of a single sequence of 186 standard stimuli.

The fourth stimulus sequence contained an alternating sequence of blocks of 10 within-category stimuli and 10 across-category stimuli separated by 30-sec interblock intervals. There were three blocks of each shift category. The interstimulus interval (onset to onset) for all sequences was 2 sec.

Design

The Ss were assigned to five groups (10 Ss per group). The groups were run successively. The experimental task for the Ss in Groups 1, 2, 3, and 4 was to detect the occurrence of shift stimuli embedded in the sequence of standard stimuli.

The Ss in Group 1 listened first to the within-category shift sequence (20-0 VOT), then, on the following day, to the across-category shift sequence (20-40 VOT). The Ss in Group 2 also listened to both sequences on successive days, but in the reverse order.

Group 3 was given 20 practice trials with both the standard and within-category stimuli before listening to the within-category shift sequence. Pretraining consisted of 20 presentations of a group of four stimuli; two standard stimuli followed by two within-category stimuli. The interval between the groups was 5 sec. The Ss were told the order of the different stimuli and were instructed to try to detect any difference between the sounds. The within-category shift sequence was begun immediately after pretraining. The Ss were given pretraining to determine whether increased familiarity with the "unfamiliar" nonphonemic distinction would improve performance.

In a no-shift condition (Group 4), the Ss listened to the tape which contained all standard stimuli. The purpose of this control was to establish a baseline from which to assess the effects of the different shift conditions. In the other control condition (Group 5), the Ss listened to the randomized sequence of blocks of within- and across-category stimuli (the fourth stimulus sequence). The purpose of this control was to determine the amplitude of the AER to the across- and within-category stimuli in a setting unrelated to the discrimination tasks and thus to assess the "inherent" amplitude of the AERs to the 0 and 40 VOT stimuli.

Groups 3, 4, and 5 were tested in a single session. The session duration was approximately 7 min.

Analysis of the Evoked Potentials

The amplitude differences between the N1 and P2 responses was determined from the X-Y plots by measuring the difference in millimeters between the maximum wave of negativity between 75 and 125 msec after stimulus onset (N1) and the maximum wave of positivity between 175 and 225 msec (P2).

Each AER was the sum of 16 individual responses. Responses to the standard and shift stimuli were averaged separately in all conditions. A separate AER was accumulated for each member of the shift pairs. The AER to the last standard stimulus before the shift pair was designated as the AER to the standard stimulus. In the no-shift condition (Group 4), evoked responses were accumulated for the stimuli which occurred in the same positions as the standard and shift stimuli in the shift conditions. For the stimulus control condition (Group 5), separate evoked responses were accumulated for the within- and across-category stimuli by summing over blocks of trials.

Table 1
Average Ratio of the Standard Stimulus N1-P2 Amplitude to the N1-P2 Amplitude of the Shift Stimuli

		Position in Shift Pair	
	Shift Category	First	Second
Group 1	Across	1.36	1.60
	Within	0.92	0.88
Group 2	Within	0.95	0.90
	Across	1.35	1.51
Group 3	Pretrained Within	0.92	0.90
Group 4	No Shift	0.95	0.90

Procedure

All Ss were instructed to remain as motionless as possible during the experiment and to fixate on a point in front of them. The Ss in Groups 1, 2, 3, and 4 were instructed to "listen for" the occurrence of "any change" from the standard stimuli. The Ss were not told which pair of shift stimuli would occur in a given test sequence. The Ss in Group 3, after practice with the within-category and standard stimuli, were told to "listen for" the same changes in the test sequence as they had listened to in the practice sessions. The Ss in Group 5 were told that they would hear separate blocks of [pa] and [ba], and were instructed simply to listen to the stimuli.

RESULTS

Amplitude of N1-P2

For each S, the amplitude scores for both shift stimuli were expressed as the ratio of the shift-stimulus amplitude to the standard-stimulus amplitude. A ratio score of 1.0 indicated that the amplitudes of the standard and shift stimuli were identical. A ratio score greater than 1.0 indicated a larger shift-stimulus amplitude than standard-stimulus amplitude. For the Ss in Group 1 (across-shift then within-shift) and Group 2 (within-shift then across-shift), separate ratio scores were computed for the within- and across-category shift conditions. The ratio scores for Groups 1-4, collapsed across Ss, are shown in Table 1.

For Groups 1 and 2, the effects of presentation order (within-shift then across-shift vs across-shift then within-shift), shift type (within vs across), and location in the shift pair (first vs second) were compared in an analysis of variance. A reliable main effect due to shift type was obtained [$F(1,18) = 25.00$, $p < .01$; $\bar{X}$ (within-shift) = .91, $\bar{X}$ (between-shift) = 1.45]. No other main effects were significant. A Shift Type by Location interaction was also obtained [$F(1,18) = 4.66$, $p < .05$].

The difference in N1-P2 amplitude to the within- and across-category shifts is illustrated for a representative S

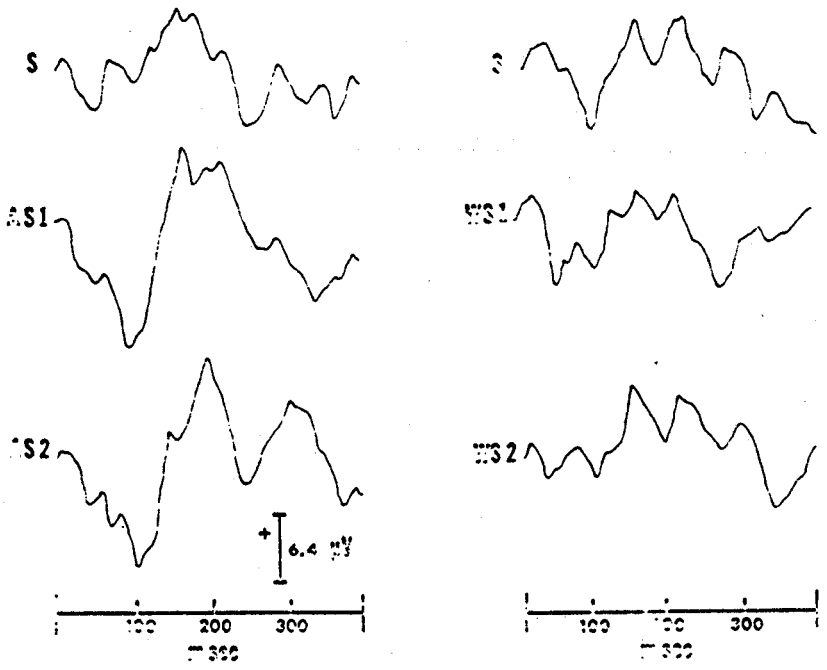

Fig. 2. Auditory evoked responses to the within- and across-category shifts for a representative S. The standard stimuli are labeled S. The within- and across-category shift pairs are labeled WS and AS, respectively.

in Fig. 2. In the across-category shift, amplitude of both members of the shift pair (AS1 and AS2) exceeded that of the standard stimulus (S). For the within-category shift, neither member of the shift pair was larger than the standard stimulus.

Since the analysis of variance showed no significant effect due to presentation order, the data for the within- and across-category shifts were pooled over Groups 1 and 2. Two additional analyses of variance were then computed with the pooled data.

The first analysis compared the pooled across-category shift condition from Groups 1 and 2 with that from Group 3 (pretrained within-category shift) and Group 4 (no shift). In the Groups by Location analysis of variance, only the groups effect was significant [$F(2,37) = 13.16$, $p < .01$]. Post hoc comparisons according to Scheffé revealed that the pooled across-category shift condition ($\bar{X} = 1.46$) differed from both the pretrained within-category condition ($\bar{X} = .91$) and the no-shift condition ($\bar{X} = .92$) at the .05 level. A second analysis of variance compared the pooled within-category shift condition from Groups 1 and 2 with that from Groups 3 and 4. The analysis of variance showed no reliable effects.

For Group 5, the absolute N1-P2 amplitude difference of the AERs to the within-category stimulus (0 VOT) and to the across-category stimulus (40 VOT) were compared by a correlated t test. The amplitudes of the two stimuli were not significantly different ($T_9 = 1.01$, n.s.).

P300 Amplitude

Due to variations in the records prior to N1, it was not possible to determine accurately a baseline from which to measure P300. Thus, no statistical analysis of P300 was attempted. However, visual inspection of the AERs indicated that P300 of the shape and magnitude reported by Ritter and Vaughan (1969) was not frequently elicited in any of the conditions. A large P300 was noted for 4 of 20 Ss in the pooled across-category shift condition. A large P300 was not noted in any of the other conditions.

DISCUSSION

The comparison of the within- and across-category shift conditions demonstrated that the across-category shift elicited a larger N1P2 response than did the within-category shift. The differences in N1-P2 amplitude in the two shift conditions cannot be attributed to an "inherently" larger N1-P2 response to the across-category stimulus than to the within-category stimulus, since in the stimulus control condition the amplitude of the N1-P2 response to 0 VOT [ba] and to 40 VOT [pa] did not differ. This outcome suggests that the difference in N1-P2 amplitude in the within- and across-category shift conditions was due to the difference in discriminability of the two types ofshift.

The comparison involving the within-category shift group and the no-shift control (Group 4) revealed that the N1-P2 response in the two conditions did not differ. Furthermore, pretraining with the within-category and standard stimuli (Group 3) did not alter the amplitude of the N1-P2 response in the within-category shift situation.

Thus, the behavior of the N1P2 component of the AER, under the conditions of the present study,

mirrored the phonetic difference between the standard and shift stimuli, not the equal physical differences between the stimuli.

P300 Amplitude

The relative absence of a large P300 component in the across-category shift condition was due at least in part to the limited bandwidth of the recording system. The 2-Hz high-pass filter severely attenuated components of the evoked response, such as the contingent negative variation, which have been implicated in the P300 response (Donchin & Smith, 1970). In subsequent studies of speech sound discrimination, using .06-Hz high-pass filters, we have consistently recorded large P300 responses in across-category shift conditions. The P300 response remains absent in within-category shift conditions.

Auditory to Phonetic Recoding

The "categorical" response of the N1-P2 component of the AER suggests that within 200 msec after the onset of a stop consonant, the finely detailed acoustic stimulus has been recoded into a categorized phonetic representation. The data from the present study do not support the suggestion that a categorical response is generated at a "long" interval after stimulus onset as a function of an arbitrary labeling of two discriminably different stimuli as belonging to the same phonetic category.

This interpretation of the data bears directly on the nature of the processing of the highly encoded stop consonants. After a stop consonant has been recoded into a categorized phonetic representation, a listener knows very little about the detailed acoustic structure of the auditory signal (e.g., VOT). The processing mechanism for the stop consonants appears to act like a "digitizing" device, accepting as input a highly variable and finely detailed auditory signal and then rapidly recording it into a quantized phonetic representation (Mattingly et al, 1971). After recoding, the detailed auditory information does not seem to be stored in any accessible form.

This interpretation of the data is supported by two recent studies exploring differences in the processing of stop consonants and steady-state vowels. Crowder (1971), using a serial recall task, found that if the vowel portions of CV syllables were varied in a serial list, then a large recency effect was obtained during recall. If, however, the consonant portions of the syllables were varied in the lists, then no recency effect was obtained. If the recency effect is contingent upon an "echoic" or "precategorical" auditory memory store of 2-3 sec duration, as Crowder and Morton (1969) have suggested, then the representation of a stop consonant does not persist 2-3 sec in "precategorical" auditory memory.

The life span of auditory memory for stop consonants has also been studied using recognition memory tasks. In one of a series of studies, Pisoni (1971) varied the interval (0, .25, .50, 1.0. 2.0 sec) between vowel pairs and stop consonant-vowel pairs in an A-X discrimination paradigm. The discrimination of vowel stimuli was markedly affected by the A-X interval, with longer intervals producing poorer discrimination. Stop consonant discrimination, however, was relatively unaffected by A-X interval. Pisoni concluded that "information other than a binding phonetic categorization is unavailable for use in discrimination [of stop consonants]." The results of the present study agree with those of Pisoni (1971) and Crowder (1971) and further reinforce the notion of a special mode of processing for the stop consonants characterized by the absence of a persistent noncategorical auditory image.

REFERENCES

Abramson, A., & Lisker, L. Discriminability along the voicing continuum: Cross-language tests. Proceedings of the 6th International Congress of Phonetic Sciences, Prague: Academia, 1970.

Cohen, R. Differential cerebral processing of noise and speech stimuli. Science, 1971, 172, 559-601.

Cooper, F., & Mattingly, I. Computer controlled PCM system for investigation of dichotic speech perception. Journal of the Acoustical Society of America, 1969, 46, 115(A).

Crowder, R. The sound of consonants and vowels in immediate memory. Journal of Verbal Learning & Verbal Behavior, 1971, 10, 587-596.

Crowder, R. G., & Morton, J. Precategorical acoustic storage (PAS). Perception & Psychophysics, 1969, 5, 365-373.

Davis, H. Enhancement of evoked cortical potentials in humans related to a task requiring a decision. Science, 1964, 145, 182-183.

Donchin, E., & Smith, D. The contingent negative variation and the late positive wave of the average evoked potential. Electroencephalography & Clinical Neurophysiology, 1970, 29, 201-203.

Greenberg, H., & Graham, J. EEG changes during learning of speech and nonspeech stimuli. Journal of Verbal Learning & Verbal Behavior, 1970, 9, 274-281.

Karlin, L. Cognition, preparation and sensory-evoked potentials. Psychological Bulletin, 1970, 73, 122-136.

Liberman, A. M. The grammars of speech and language. Cognitive Psychology, 1970, 1, 301-332.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. Psychological Review, 1967, 74, 431-461.

Lisker, L., & Abramson, A. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague: Academia, 1970.

Mattingly, I., Liberman, A., Syrdal, A., & Halwes, T. Discrimination in speech and nonspeech modes. Cognitive Psychology, 1971, 2, 131-157.

Miller, G. A. The magical number seven, plus or minus two, or some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-96.

Pisoni, D. On the nature of categorical perception of speech sounds. Supplement to the Haskins Laboratories' Status Report on Speech Research, November 1971.

Pollack, I. The information of elementary auditory displays. Journal of the Acoustical Society of America, 1952, 24, 745-749.

Ritter, W., & Vaughan, H. G. Averaged evoked responses in vigilance and discrimination: A reassessment. Science, 1969, 164, 326-328.

Sheatz, G. C., & Chapman, R. M. Task relevance and auditory evoked responses. Electroencephalography & Clinical Neurophysiology, 1969, 26, 468-475.

Studdert-Kennedy, M., Liberman, A., Harris, K., & Cooper, F. The motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 1970, 77, 234-249.

Wilkinson, R., & Lee, M. Auditory evoked potentials and selective attention. Electroencephalography & Clinical Neurophysiology, 1972, 33, 411-418.

Wood, C., Goff, W., & Day, R. Auditory evoked potentials during speech perception. Science, 1971, 173, 1248-1251.

NOTES

1. VOT refers to the relative timing of the release of supraglottal closure and the onset of laryngeal pulsation or "voicing." Abramson and Lisker (1970) have argued that the acoustic features of explosion energy, amount of aspiration, and first-formant intensity may all be derived from the single articulatory variable of VOT. In sound spectrograms, VOT is reflected by the onset of the first formant relative to the second and third formants and, for stop consonants with a delayed onset of the first formant, the presence of aspiration in the upper formants in the period preceding the onset of the first formant.

2. The three synthetic speech stimuli used in this study were slight modifications of the stimuli used by Lisker and Abramson (1970). Listening tests by the author and his colleagues indicated that the 20 VOT stimulus used in the present study was labeled more consistently as a [ba] than were the 20 VOT stimulus used by Lisker and Abramson. These tests also indicated that the 20 VOT stimulus was discriminated less often from the 0 VOT stimulus than was the corresponding stimulus used by Lisker and Abramson.

(Received for publication May 4, 1973;
revision received June 30, 1973.)