Journal of Experimental Psychology:
Human Perception and Performance
1980, Vol. 6, No. 1, 110–125

Phonetic Categorization in Auditory Word Perception

William F. Ganong III
Brown University

To investigate the interaction in speech perception of auditory information and lexical knowledge (in particular, knowledge of which phonetic sequences are words), acoustic continua varying in voice onset time were constructed so that for each acoustic continuum, one of the two possible phonetic categorizations made a word and the other did not. For example, one continuum ranged between the word *dash* and the nonword *tash*; another used the nonword *dask* and the word *task*. In two experiments, subjects showed a significant lexical effect—that is, a tendency to make phonetic categorizations that make words. This lexical effect was greater at the phoneme boundary (where auditory information is ambiguous) than at the ends of the continua. Hence the lexical effect must arise at a stage of processing sensitive to both lexical knowledge and auditory information.

Linguistic context has long been known to aid and bias the identification of speech. For example, the identification of sequences of words in noise is substantially aided if the words form sentences (Miller, Heise, & Lichten, 1951). The phoneme restoration effect (Warren, 1970) shows that context can control the perception of individual segments: When a single phone of an utterance is replaced by a noise burst or a cough, a listener is not aware which phone was replaced. Both previous and subsequent context influence this effect (Warren & Sherman, 1974). Perception is influenced not only by immediate linguistic context (e.g., the role a word plays in a sentence) but also by the frequency of a word's use in the language. For example, there is a large word-frequency effect in the identification of words presented in noise (Broadbent, 1967).

This article examines the influence on phonetic categorization of a rather simple linguistic variable: the lexical status of a phonetic sequence (i.e., whether the phonetic sequence is a word). Lexical status is already known to influence phonetic processing. Reaction time for phoneme detection is faster if the target appears in a word than if it appears in a nonword (Rubin, Turvey, & Van Gelder, 1976). Presumably this word advantage reflects a perceptual bias in favor of words similar to the perceptual bias that favors high-frequency words over low-frequency words in noise. The purpose of this article is to determine the stage in the perceptual process at which the biasing effect of lexical status appears. Does it follow phonetic categorization, or, alternatively, can it influence the interpre-

This work was supported by Grant HD 05331 from the National Institute of Child Health and Human Development to Peter D. Eimas, by National Science Foundation (NSF) Grant BMS 75 08439 to Richard B. Millward, by National Institutes of Health Grant NO1 HD 12420 to Haskins Laboratories, and by an NSF postdoctoral fellowship to the author (SMI 77-12336).

Alvin Liberman made the facilities of Haskins Laboratories available to me. Terry Halwes's and Patricia Keating's aid in generating the stimuli was essential, and discussion with them, Susan Carey, and Peter D. Eimas helped to clarify the issues discussed here. Several reviewers, including Neal Macmillan, Bruno Repp, and Michael Posner, made comments that improved the presentation of this material. I thank them all. This work was reported at the 95th meeting of the Acoustical Society of America at Providence, Rhode Island, May 1978.

Requests for reprints should be sent to William F. Ganong III, who is now at the Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Pennsylvania 19104.

tation of acoustic cues, which underlies phonetic categorization?

Asking this question presupposes that there is a stage of processing in normal speech perception that carries out phonetic categorization. This is a substantive assumption—it is possible, for example, to construct models for word perception that do not include a stage of phonetic categorization (Klatt, 1979). But the assumption must certainly be correct for experiments in which subjects are required to make phonetic categorizations. Evidence for this assumption in other experimental contexts derives primarily from work on categorical perception. The first psychologists to investigate the perception of synthetically constructed stop consonants discovered that listeners were able to discriminate with ease only those stimuli that they perceived as belonging to different phonetic categories (Liberman, Harris, Hoffman, & Griffith, 1957). Discrimination of stimuli from the same phonetic category was only slightly above chance. This phenomenon was named categorical perception because subjects acted as if the only information available was the phonetic category of each stimulus.

The concept of categorical perception has since undergone modification. It is now clear that subjects can use some auditory information as well as information about phonetic categories in discrimination (Carney, Widin, & Viemeister, 1977; Fujisaki & Kawashima, 1970), but under many conditions, the perception of speech, especially the stop consonants, is very nearly categorical. Is this auditory information of any use in the normal course of speech perception (the purpose of which is clearly the identification of words and sentences, not their discrimination)? Liberman, Mattingly, and Turvey (1972) proposed, instead, that the role of phonetic categorization is the substitution of a phonetic label for the auditory information representing a stop consonant in order to facilitate higher level linguistic processing.[1]

The phonetic categorization of an acoustic continuum between different stop consonants is typically characterized by two unambiguous regions (which are consistently given one or another phonetic categorization) separated by a narrow boundary region containing phonetically ambiguous stimuli. An acoustic variable frequently used to construct such acoustic continua is voice onset time (VOT). VOT is an acoustic cue for voicing in syllable-initial stops in many languages (Lisker & Abramson, 1964). Perception of such a continuum is often described by a single number, the locus of the phoneme boundary, which is the (interpolated) point on the acoustic continuum that would receive each phonetic categorization on half of the trials. For an English-speaking subject, the phoneme boundary between d and t responses for a [da–ta] continuum is at about 35-msec VOT. Stimuli with VOT of less than 30 msec are consistently labelled d, and stimuli with VOT greater than 40 msec are consistently labeled t.

A model in which lexical status affects phonetic processing only after phonetic categorization has occurred can be called a *categorical* model. In an *interactive* model, lexical status would be allowed to direct and bias the processing of the auditory information specifying phonetic categories. One example of such a model is the *criterion-shift* model, in which lexical status would

[1] The situation is clearly different for noncategorical phonetic distinctions. Many phonetic distinctions (such as differences between vowels) are conveyed by large acoustic differences that can be easily discriminated regardless of phonetic category. Context can bias the interpretation of these differences in words. In an experiment on the perception of disyllabic words in noise, Pollack (1959) showed that some additional auditory information beyond a simple phonetic transcription is available to subjects for a few seconds after a word is presented. He had subjects identify words from a response set given at various delays after the stimulus. For delays under 5 sec, performance depended on the delay between presentation of the stimulus and the response set and was better than performance based on subjects' immediate identification of the stimulus presented. Thus, some form of auditory memory must have been involved. However, Pollack's stimuli differed in both consonants and vowels. The auditory memory used may have been primarily memory for vowels. The present experiment extends Pollack's results to the class of speech sounds that are least likely to allow such auditory memory: the stop consonants.

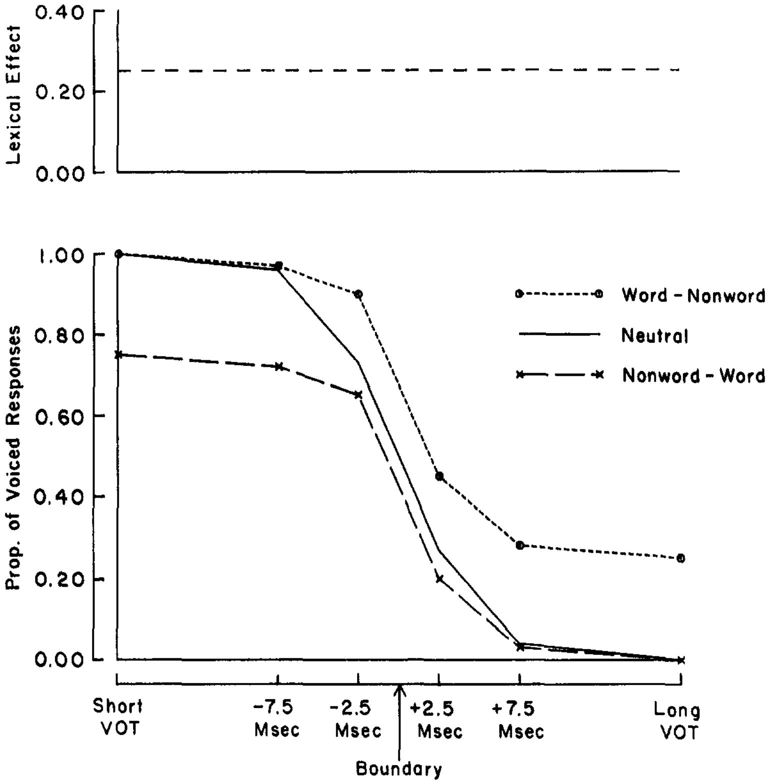

Figure 1. The lexical effect, as predicted by the categorical model. (The lower portion of the graph shows the proportion of stimuli given a voiced phonetic categorization as a function of voice onset time (VOT) for three different types of continua. The solid line shows idealized data from a neutral continuum, whose perception is not biased by lexical effects. The line with short dashes shows the the categorical model's prediction of a lexical effect on a continuum whose voiced end is a word and whose voiceless end is a nonword. Similarly, the line with long dashes shows the result of a post-categorical tendency to make phonetic categorizations voiceless, on a continuum in which voiceless responses make words. The top part of the figure shows, as a function of VOT, a measure of the lexical effect.)

change the criterion by which the auditory information is judged. The categorical model can be tested by determining whether the effect of lexical status is concentrated at the phoneme boundary.

In the categorical model, a bias toward phonetic categorizations that make words would operate as a correction process. Candidate phonetic categorizations that did not happen to make words would be changed to phonetic categorizations that did make words. Thus, subjects presented with speech stimuli for phonetic classification as beginning with [d] or [t] might, upon hearing the nonword *tash*, correct the [t] categorization to [d] so as to make the word *dash*. Since a strict categorical model assumes no acoustic information is available when lexical status has its effect (i.e., after phonetic categorization), for a particular acoustic continuum between words and nonwords, the probability of a categorization being corrected depends only on the categorization and not on the acoustic information that specified it. A formal description of this model is given in Appendix A.

In the present experiment, the phonetic categorization of a lexically biased continuum is not compared with the categorization of an unbiased continuum, but with a continuum biased in the opposite direction. Figure 1 shows the result of suck biases, according to the categorical model

The figure shows a hypothetical phonetic categorization function (the solid line) based on a cumulative normal function, assuming a difference of 5-msec VOT corresponds to a *z* score difference of 1.2. The categorical model's predictions for the shape of the identification functions, assuming a probability of correcting a nonword response of .25, is also shown. In the top panel of Figure 1 is the resulting lexical-effect function, which is simply the difference between the two identification functions. The shape of this lexical-effect function, as predicted by the categorical model, is derived in Appendix A. In Figure 1 it is assumed that the two lexical biases are equal, so the lexical-effect function does not depend on VOT. If the biases are not equal, the resulting lexical-effect function will be a monotonic function of VOT, and its value at the phoneme boundary will be the average of the values at the ends of the continuum.

In a criterion-shift model, on the other hand, a bias toward hearing words could affect the interpretation of acoustic information at a stage of processing before phonetic categorization. The criterion for making a phonetic categorization would depend on lexical status. This change in criterion would produce a shift in the location of the phoneme boundary. For example, a subject whose phoneme boundary for a [da–ta] continuum was at 35-msec VOT might require 40-msec VOT to hear a [t] in the environment *_ash* (because *dash* is a word and *tash* is not). Such a shift in the location of the phoneme boundary would not produce a uniform effect throughout the continuum but would produce an effect concentrated near the phoneme boundary, just as a change in threshold affects the probability of detection of near-threshold stimuli far more than subthreshold or suprathreshold stimuli. The categorization of stimuli far from the phoneme boundary would not be influenced much because the shift in criterion would not affect the interpretation of unambiguous acoustic evidence. However, for the acoustically ambiguous stimuli near the boundary, a change in criterion would produce large effects on categorization. Such a change could shift a stimulus from the ambiguous region to the word region or from the part of the nonword phonetic category near the boundary into the ambiguous region. This model is also described quantitatively in Appendix A.

Figure 2 shows the predictions of the criterion-shift model. This figure assumes equal but opposite direction lexical biases for the two continua, as did Figure 1. Assuming that the phonetic categorization function is sigmoid, with the steepest slope near the phoneme boundary, the lexical-effect function will generally reach a maximum in the neighborhood of the phoneme boundary and thus be greater near the phoneme boundary than at either end of the continuum. This is unlike the shape of the lexical-effect function according to the categorical model, for which the value at the boundary must be between the values at the ends of the continua. Thus, to test the categorical model, it is only necessary to determine whether the effect of lexical bias is spread equally throughout the continuum (as the categorical model predicts) or concentrated at the phoneme boundary (as the criterion-shift model predicts.)

The most direct way to determine whether lexical effects are stronger near the phoneme boundary would be to compare phonetic categorizations of lexically biased acoustic continua with the phonetic categorizations of neutral continua, which are not biased by lexical factors. Alexander (Note 1) has done this with limited success.[2] The present study, instead, compared categorizations of matched pairs of continua chosen so that lexical biases on the two continua operate in opposite directions. Pairs of VOT continua between monosyllabic words and nonwords were synthesized so that the voiced end of one continuum of each pair was a word and the voiceless end

[2] The problem with this approach is that biases in the perception of the neutral continuum against which the lexically biased series are judged can confuse and hide lexical effects. This seems to have happened in Alexander's (Note 1) study. It showed an overall bias toward phonetic categorizations that make words, but this bias was not reliable across continua.

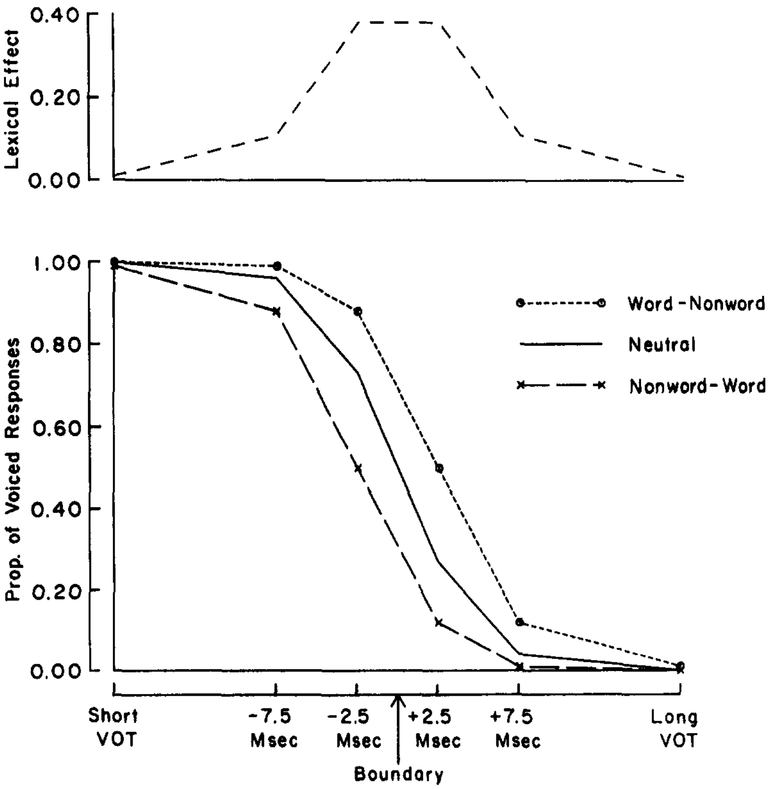

Figure 2. The lexical effect, as predicted by the criterion-shift model. (This figure is analogous to Figure 1, except that it shows predictions of the criterion-shift model.)

was not, and vice versa for the other continuum. For example, one continuum ranged from the word *dash* to the nonword *tash*, and its matched pair ranged from *dask* (nonword) to *task* (word). Thus, it is only possible to measure the lexical effect for the two continua combined. The continua of each pair were carefully matched to have the same vowel and to be as similar as possible in postvocalic consonants.

Experiment 1

Method

Stimuli. Seven pairs of continua were synthesized using the Haskins Laboratories speech synthesis by rule program, FOVE (Ingemann, Note 2). Four alveolar continuum pairs were synthesized. One pair was based on the words *dash* and *task* (that is, one continuum ranged between the word *dash* and the nonword *tash*; the other ranged between the nonword *dask* and the word *task*). Other continuum pairs were based on *dust* and *tuft*, *dirt* and *turf*, and *dose* and *toast*. Three velar continua, based on *gift* and *kiss*, *geese* and *keep*, and *gush* and *cusp* were also synthesized. Each continuum had seven members, with VOTs of 15, 25, 30, 35, 40, 45, and 55 msec. The stimuli were recorded on audiotape in the format that subjects later heard. The alveolar and velar stimuli were presented in different blocks. Thus, there was a block containing a randomization of all the alveolar stimuli (presented with a 3-sec interstimulus interval), which was followed by a block containing all the velar stimuli. This pattern was repeated six times. Another tape was constructed that contained all the endpoint stimuli, with a 5-sec interstimulus interval.

Procedure. Seventeen subjects[3] (paid volunteers from the psychology department's subject pool) participated in the experiment. First, they were presented with a block of trials containing each stimulus from the alveolar continua. The subjects were instructed to write, for each stimulus, their first impression as to whether the syllable began with d or t. This was followed by a block of velar trials. Six

[3] An 18th subject, who refused to categorize any of the velar stimuli as g or k, was immediately dropped.

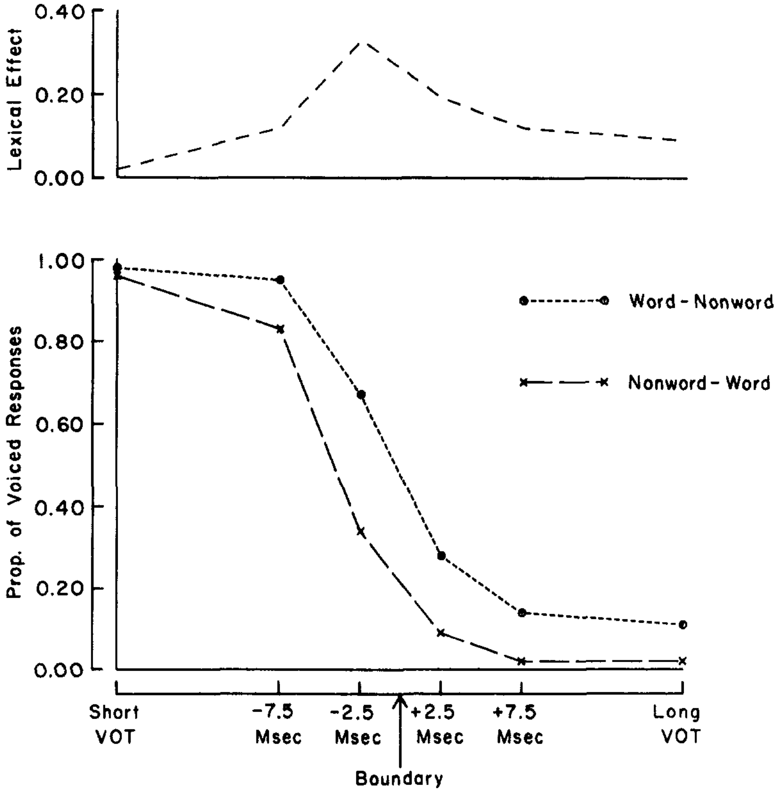

Figure 3. Results of Experiment 1. (Phonetic categorizations pooled as described in the text.)

blocks of alveolar and six blocks of velar trials were presented to each subject. The subjects were not told which words and nonwords would be presented.

Unfortunately, FOVE does not always produce perfectly intelligible speech. To be able to show an advantage for phonetic categorizations that make words, a subject must hear the words as words. Hence, in a second condition, (which always followed the first) the endpoints of the continua were presented and subjects spelled out the words and nonwords they heard. The subjects were instructed to spell words correctly and to make a rough guess for the spelling of nonwords. The data from the second condition were used to determine which continua each subject heard correctly. For each subject, the phonetic categorization data were analyzed for only those continuum pairs for which the lexical status (e.g., word or noword) of three of the four endpoint stimuli was correctly identified.

Results

Application of the criterion to subject's spelling of the words and nonwords resulted in elimination of from 0 to 3 of the 7 continuum pairs for each subject, for a total of 22 of the 119 continuum pairs (18%). Different continuum pairs passed the criteria to quite different degrees. The most successful pair, dash–task, was spelled acceptably by all of the 17 subjects, whereas the geese–keep pair was spelled correctly by only 6 of the subjects.

The results of phonetic categorization, pooled across subjects and continua, are shown in Figure 3. Phoneme boundaries differed across subjects and across continua. Thus, each subject's data from a given continuum pair was pooled, and the phoneme boundary for that subject on that continuum pair was determined. The position of each phoneme boundary was estimated by finding the VOT that received the proportion of voiced responses closest to one half. Henceforth, the data are considered relative to the position of these phoneme boundaries. The line with short dashes in the figure shows the proportion of voiced responses to continua whose voiced end is a word, and the line with long dashes shows responses to continua whose voiceless end is a word. The leftmost and rightmost points of each line show responses to the endpoints of the continua: stimuli with VOTs of 15 and 55 msec. The four middle points of each line show responses to stimuli with various differences in VOT from the phoneme boundary. For example, the third point from the left in each line was determined by pooling data from the first stimulus to the left of each subject's phoneme boundary for each continuum pair. There was a small but consistent lexical effect in all seven continuum pairs ($p < .01$, sign test). This effect was also consistent across subjects (pooling data across different continua); 16 of the 17 subjects showed a lexical effect ($p < .001$, sign test).

In Figure 3, the lexical effect seems to be stronger at the voiceless end of the continua than at the voiced end. This pattern was reliable across subjects (11 subjects had stronger effects at the voiceless end, 2 at the voiced end, and 4 had equal-size effects; $p < .05$, sign test) but not across continua (five continua showed the effect, one did not, and one showed equal-size effects). Most of the effect was caused by the gift–kiss continuum pair. Most subjects categorized almost all of the stimuli from the gift–kift continuum as beginning with g.

The data were next examined to determine whether the lexical effect was spread throughout the continuum, as predicted by the categorical model, or concentrated at the phoneme boundary, as predicted by the interactive model. For each subject, the size of the lexical effect exhibited at the phoneme boundary was compared with the sum of the two lexical effects measured at the ends of the continua. Sixteen of the subjects showed more lexical effect at the phoneme boundary than at either endpoint, and 1 showed the opposite tendency ($p < .005$, sign test). Five of the seven continuum pairs showed more of an effect at the boundaries than at the endpoints. One showed equal-size effects (the dose–toast pair), and one (the problematical geese–keep pair) showed the opposite effect. Only 6 of the subjects passed the (quite weak) screening criteria for this continuum, and none of them heard all four of the endpoint stimuli from these continua correctly. The concentration of the lexical effect at the phoneme boundary is, then, not significant across continua ($p > .1$). However, this failure to obtain significance is probably due to the small number of different continuum pairs tested and to the poor quality of the geese–keep pair.

Experiment 2

Experiment 2 was designed to provide a replication of Experiment 1, with more continuum pairs, to determine whether the lexical effect arises the first time a subject hears stimuli from a particular continuum,[4] and to determine whether the effect is robust enough to appear even if subjects know the stimulus set.

Method

Stimuli. The stimuli for Experiment 2 were produced by digitally cross splicing tokens of natural speech. Any stimulus with a positive VOT contains two acoustic segments: a voiceless, noisy segment (consisting of a burst and aspiration) and a voiced segment which begins, by definition, at the voice onset time. In VOT continua produced with a speech synthesizer (as used in most previous experiments on the perception of VOT), the voiceless segment is produced by passing aspiration noise through the formant filters. At the onset of voicing, glottal pulsing replaces aspiration noise as the source for the formant filters, and a voiced segment results. For the stimuli for Experiment 2, these two segments were excised from natural speech rather than produced synthetically (Lisker, 1976; Spencer & Halwes, Note 3). Thus the stimuli were considerably more intelligible than those of Experiment 1.

For each continuum pair, two voiced stimuli with exactly the same acoustic waveform for the first 100 msec were constructed by digital splicing. Two voiceless endpoints were similarly constructed. These tokens were used to supply the voiced and aspirated segments to construct the stimuli with different VOTs.

[4] This condition was included to assure that the lexical effect is truly perceptual, that is, that it arises before subjects have learned the set of continua used in the experiment. Otherwise, the lexical effect might be due to processes peculiar to a situation in which a closed set of words and nonwords is presented repeatedly.

For each continuum, six stimuli were produced by cross splicing. One had the shortest VOT possible using the particular natural tokens on which the stimuli were based. The next four stimuli for each continuum were chosen to span the phoneme boundary (as measured in a pilot experiment) in approximately 5-msec steps. The sixth stimulus had the longest VOT possible given the tokens used. The VOTs of the first, second, and sixth stimulus of each continuum are given in Table 1. The VOTs of the third, fourth, and fifth stimuli were approximately 5, 10, or 15 msec more than the VOT of the second stimulus. The details of the construction and selection of these stimuli is given in Appendix B.

Eighteen new subjects (again, paid volunteers from the psychology department's subject pool) participated in Experiment 2. The experiment was conducted using a PDP-8 minicomputer, which generated the stimuli on-line by digital splicing, presented the stimuli to subjects (at a 10-kHz sampling rate) and collected responses. Subjects were run in groups of 3 or fewer. They heard stimuli presented over headphones and typed their responses. Each subject was assigned to one of six groups, each group containing 3 subjects. All subjects in each group heard the stimuli in the same order.

Each subject participated in one session consisting of seven blocks of trials. In the first block of trials, subjects were presented with a randomization of all 48 labial stimuli and pushed the b or p keys to indicate their first impression of the first segment of each stimulus. The order in which stimuli were presented was constrained so that for each stimulus of each continuum, one of the six groups of subjects heard that stimulus before they heard the other stimuli from the same continuum. The second and third blocks presented the alveolar and velar stimuli in an analogous fashion for phonetic categorization.

The fourth block was a whole-syllable identification condition. Subjects were presented with endpoint stimuli from all of the continua in a random order and responded by indicating whether the stimulus was a word or nonword (by pushing the w or n keys) and spelling out the stimulus.

In the fifth through seventh blocks, subjects again phonetically categorized the labial, alveolar, and velar stimuli. However, this time the stimuli were presented in groups of trials containing only stimuli from a particular continuum pair, and the first four stimuli of each such group of trials were the four endpoints of the two continua. Within each block, the different groups of trials were not explicitly separated, but the way in which the stimuli were grouped was explained to the subjects. Throughout the experiment, each stimulus was presented 3 sec after the last subject finished responding to the previous stimulus.

Table1
Stimuli for Experiment 2

Word		Voice onset times		
		Stimulus		
Voiced	Voiceless	1	2	6
	Labials			
bash	past	10.0	19.0	61.7
boat	pope	8.6	18.0	42.8
babe	page	7.2	16.2	59.4
beef	peace	8.0	16.7	56.1
	Alveolars			
dark	tarp	18.6	36.6	84.4
deep	teach	22.1	36.0	89.0
depth	text	19.3	46.9	82.6
dirt	turf	19.5	52.2	86.0
	Velars			
garb	cars	18.2	40.6	64.3
gorge	corpse	35.4	53.2	86.1
gulp	cult	35.6	39.4	69.8
gout	couch	14.9	37.9	66.7
gift	kiss	23.1	40.6	85.1
geese	keep	21.9	39.0	88.8

Note. Stimuli 1 and 6 had the shortest and longest voice onset times (VOTs) possible, given the particular tokens used. Stimuli 2, 3, 4, and 5 were chosen to span the phoneme boundary in 5-msec steps. Thus the voice onset times (VOTs) of stimuli 3, 4, and 5 can be obtained by adding 5, 10, or 15 msec to the VOT for Stimulus 2.

Results

The same criteria used in Experiment 1 were applied to subjects' identification of the lexical status of the endpoints (collected in the fourth block). The stimuli for Experiment 2 were more successful than those of Experiment 1—only 24 of the 252 continuum pairs were eliminated (10%). Figure 4 shows the phonetic categorization data, pooled in the same way as the data of Figure 3 (i.e., with respect to phoneme boundary locations for each continuum pair), for the first three blocks of trials, in which the stimuli were presented in random order for phonetic categorization. Again, there is a lexical effect that is significant across subjects (17 of 18, $p < .001$, sign test) and across continua (14 of 14, $p < .001$). This time the effect is significantly stronger at the boundary than at the end of each continuum, showing a larger effect both for subjects (15 of the 18 subjects show the effect, $p < .005$) and for continua (14 of 14, $p < .001$). Thus, in Experiment

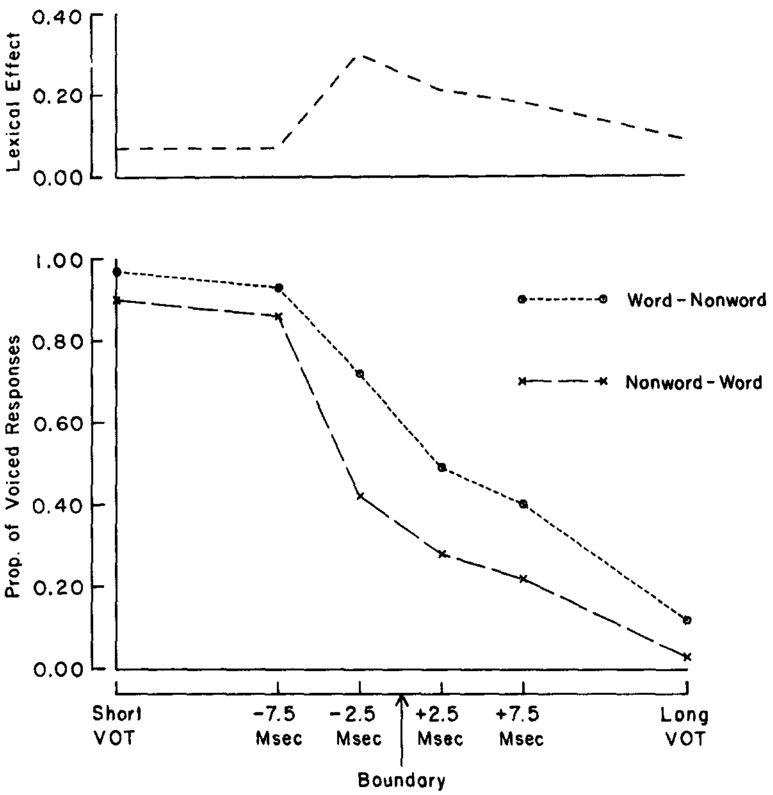

Figure 4. Results of Experiment 2. (Phonetic categorizations of each subject during the first three blocks of the experiment, when stimuli from different continua were presented together in random order.)

2 as in Experiment 1, there is a lexical effect concentrated at the phoneme boundary.

The data were next examined to see if the lexical effect was present the first time that subjects heard a stimulus from a particular continuum. Figure 5 shows the resulting data, pooled with respect to phoneme boundary location tor the first three blocks of trials. Again, there is a significant lexical effect across continua (nine continua show the lexical effect, one shows the opposite effect, and four show no effect, $p <$ .05, by a sign test), and, again, it appears stronger at the phoneme boundary than at the ends of the continua.

Finally, data gathered in the last three blocks of trials were examined. Responses pooled in the usual way are shown in Figure 6. Again, there is a significant lexical effect that is stronger at the boundary than at the endpoints. This effect is not apparently different from the effects during the first blocks.

The failure of blocking to reduce or eliminate the word advantage was surprising, since most previous studies of the word advantage in vision or word frequency effects in auditory word perception have found that the effects are eliminated when the message set is known. However, the word advantage in visual perception can be maintained in the face of perfect knowledge of the stimulus set (Smith & Haviland, 1972) if the visual angle of the words is small enough (Purcell, Stanovich, & Spector, 1978).

In Figure 4 (as in Figure 3), there seems to be a slightly larger lexical effect for the voiceless stimuli than for the voiced stimuli. On closer examination, however, this tendency is not reliable across subjects (of the 14 subjects showing different amounts of lexical bias at the ends of the continua, 7 showed more effect at the voiced end) nor continua (five of the nine unequally affected continuum pairs showed stronger effects at the voiced end). The tendency for the lexical effect to be greater at the voiceless end of the continuum in Experiment 1 thus seems to have been a product of the particular words or speech-synthesis strategy used there.

The data were also examined to determine whether the size of the lexical effect is correlated with the frequency of occurrence of the words defining the continua. The sum of the logarithm of the frequency of occurrence (as measured using the Kucera & Francis, 1967, norms) of the words defining each continuum pair was correlated with the size of the lexical effect measured at the phoneme boundary for each pair. The correlation coefficient (—.04) was not reliably different from zero. Thus, no word-frequency effect is apparent in these data. However, the experiment was not designed to test for word-frequency effects, so little importance should be attached to this result.

Discussion

It is clear from the results of Experiments 1 and 2 that lexical status affects phonetic categorizations much more for acoustically ambiguous (i.e., boundary) stimuli than for acoustically unambiguous (endpoint) stimuli. Hence, lexical status has an effect before acoustic information is replaced by a phonetic categorization. This demonstration that lexical effects are stronger at the boundary than at the endpoints does not rule out the possibility that some lexical effects occur after phonetic categorization. However, the concentration of the lexical effect at the phoneme bound-

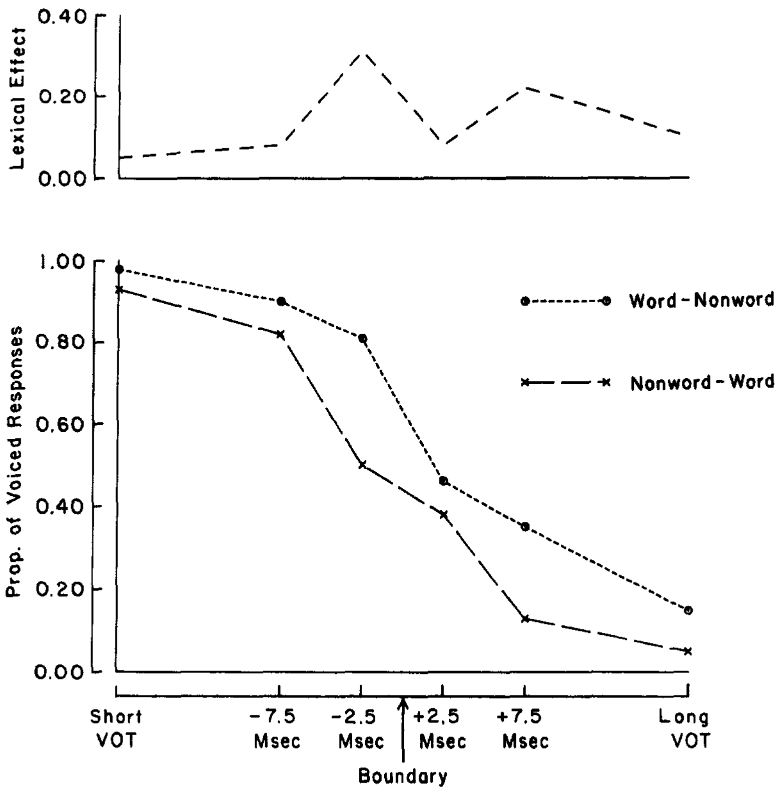

Figure 5. Phonetic categorization when stimuli are unfamiliar. (These data represent subjects' responses to the first stimulus presented to them from each continuum.)

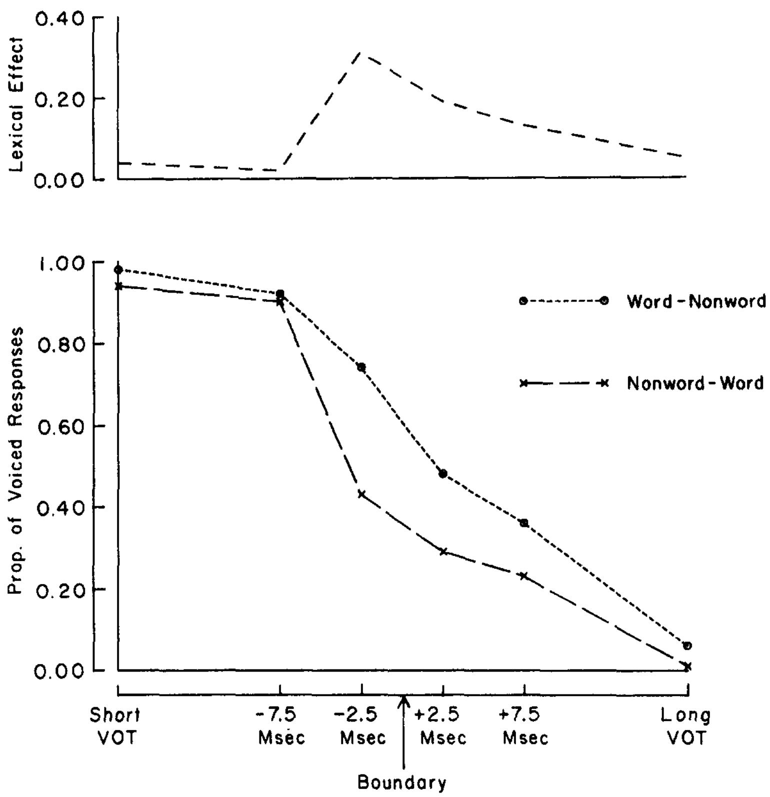

Figure 6. Phonetic categorization with grouped presentation of stimuli. (The results of the last three blocks of Experiment 2, when stimuli from each continuum pair were presented in the same group of trials.)

ary certainly does show that some acoustic information is available when lexical knowledge comes into play. So the categorical model is incorrect, although there may be postcategorical effects as well as precategorical ones.[5]

Lexical status is a fairly simple form of higher level linguistic knowledge that might be expected to affect the interpretation of acoustic evidence. Other studies have shown effects comparable to the results reported here, using semantic (rather than lexical) information to bias phonetic categorizations. Garnes and Bond (1977) showed that sentence context can bias phonetic categorization, and Spencer and Halwes (Note 3) have shown that the size of these effects depends, to some extent at least, on subjects' expectations about and knowledge of the experimental situation. Similarly, Marslen-Wilson and Welsh (1978) and Cole and Jakimik (1978) have shown that the detection and shadowing of mispronunciations depends on the predictability (on syntactic and semantic grounds) of the mispronounced word. Thus, it seems that the additional auditory information tapped by the lexical effect can interact with much higher level linguistic constraints.

There are at least three different levels at which this information could be coded. The information could be coded as extra candidate phonetic categorizations; it could be

[5] It could be argued that the results of the present experiments are "merely" due to response bias. But this claim is irrelevant to the question under consideration here: The categorical model prohibits response bias (or any other tendency to perceive speech sounds as words) from being concentrated at the phoneme boundary.

kept in a raw, uninterpreted form; or it could be coded as confidence ratings for various phonetic categorizations.

In a simple version of sophisticated guessing theory, the extra information would be represented as a set of possible phonetic categorizations (Catlin, 1969). In the case of stimuli from an acoustic continuum, this is equivalent to adding a third categorization, *don't know*, to the two possible categorizations of the continuum. To account for the lexical effect's concentration at the phoneme boundary, this model must assume that only items labeled as ambiguous are susceptible to lexical effects and that the probability of an item receiving an ambiguous categorization increases near the phoneme boundary. Of the various possible representations of the additional auditory information, this is the most linguistic.

At the opposite pole is the possibility that the perceptual system stores the additional auditory information in a raw, unprocessed, echoic form. In this model, when the output of the phonetic recognizer is not a word, the acoustic evidence specifying that phonetic sequence would be reexamined to determine whether the evidence was consistent with any phonetic strings that were words. This model would explain the concentration of the lexical effect near the phoneme boundary by postulating that stimuli far from the phoneme boundary would be consistent with only one phonetic categorization. Stimuli near the boundary, on the other hand, would allow two categorizations, and the one that made a word would be favored.

It is necessary to posit the existence of an echoic store not only to explain the evidence for the stimulus suffix effect (Crowder & Morton, 1969) but also to explain the trading relations shown in the integration of different acoustic cues into a phonetic percept (Repp, Liberman, Eccardt, & Pesetsky, 1978). There are two problems with the assumption that the additional auditory information involved in the lexical effect is stored in echoic memory. One is the apparent coarseness of the code in echoic memory. Distinctions among stop consonants cannot be retrieved from the precategorical acoustical store. Thus, a stimulus suffix interferes with memory for acoustically dissimilar vowels but not for acoustically similar vowels (Darwin & Baddeley, 1974) or stop consonants (Crowder, 1971). Another potential problem is the short duration over which information stored in echoic memory is available. Estimates vary, but the duration of echoic memory for consonants is probably under one second.

A third form of coding, intermediate in processing depth between the candidate phonetic categorizations of sophisticated guessing theory and the raw sensory information of the echoic store, would use goodness of fit ratings of different possible phonetic categorizations. In this model, a bias toward categorizations that make words would show up as a lower rating threshold for words than for nonwords. For stimuli near the phoneme boundary, this difference in thresholds would affect phonetic categorization. However, the phonetic categorization of stimuli far from the boundary would not be influenced by these differences in threshold because the ratings of the correct phonetic categorization would be quite high. Information in this form is neither strictly phonetic nor auditory but, rather, expresses the relation between the auditory data and the phonetic possibilities in a succinct way. It is interesting that the most successful speech understanding system to date, HARPY, works in this way (Klatt, 1977; Lowerre & Reddy, in press).

Also, implicit confidence rating responses of this sort have been used in many models of linguistic processing. Morton's (1969) logogens measure the acoustic/phonetic fit of each word in the language to the stimulus.[6] Marslen-Wilson and Welsh's (1978) modifications of the logogen view preserve this feature. And Massaro's (Note 4) model of the integration of information

[6] The logogen model, as stated by Morton (1969), simply counts phonetic features. To account for the data presented here, the model must be slightly modified to sum the goodness of fit ratings of each segment of the word. This modification seems in the spirit of the original logogen model.

from different knowledge sources also provides estimates of the goodness of fit of various options. Massaro's work is of particular interest because he has carried out experiments on reading that are analagous to the experiments described here. He has examined the visual perception of letterlike forms drawn from a (visual) continuum from the letter *c* to the letter *e* (Naus & Shillman, 1976). The perception of such stimuli is biased in favor of orthographically regular strings, just as in the present experiment phonetic categorization is biased toward words.

Deciding between these different representations of the additional auditory information will be a difficult task because they are all modifications of the simple categorical model with the same goal. They can perhaps best be distinguished by examining the way in which the processing of phonetically ambiguous items is influenced by later arriving biasing information. All of the work described above on syntactic and semantic context in phonetic decisions provides the biasing context before the phonetic ambiguity. Providing the biasing information at various delays after the ambiguous item should provide information about the nature of the coding of this additional auditory information. If the information is stored in some sort of echoic store, phonetic categorization of the ambiguous acoustic information should only be susceptible to bias over a short period of time, regardless of the linguistic structure of the utterance. On the other hand, if the information is stored as confidence ratings or as sets of possible phonetic categorizations, the linguistic structure of the utterance (particularly, whether the ambiguous information and biasing information are in the same clause) would be of great importance.

Reference Notes

1. Alexander, D. *The effect of semantic value on perceived b–p boundary.* Unpublished manuscript, Massachusetts Institute of Technology, 1972.
2. Ingemann, F. *Speech synthesis by rule using the* FOVE *program.* Paper presented at the International Congress of Phonetic Sciences, Miami, 1977.
3. Spencer, N. J., & Halwes, T. *Relating categorical speech perception to ordinary language: Boundary shifts on a "t" to "d" continuum in nonsense, word, and sentence context.* Manuscript in preparation, 1978.
4. Massaro, D. W. *Reading and listening* (Tech. Rep. 423). Madison: University of Wisconsin, Wisconsin Research and Development Center for Cognitive Learning, December, 1977.

References

Broadbent, D. E. Word-frequency effect and response bias. *Psychological Review*, 1967, *74*, 1–15.

Carney, A. E., Widin, G. P., & Viemeister, N. F. Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, 1977, *62*, 961–970.

Catlin, J. On the word-frequency effect. *Psychological Review*, 1969, *76*, 504–506.

Cole, R. A., & Jakimik, J. Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of information processing.* London: Academic Press, 1978.

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

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

Darwin, C. J., & Baddeley, A. D. Acoustic memory and the perception of speech. *Cognitive Psychology*, 1974, *6*, 41–60.

Fujisaki, H., & Kawashima, T. Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute* (University of Tokyo), 1970, *29*, 207–214.

Garnes, S., & Bond, Z. S. The relationship between semantic expectation and acoustic information. In W. Dressler, O. Pfeiffer, & T. Herok (Eds.), *Phonologica 1976 Akten der dritten Internationalen Phonologie-Tagung Wien, 1.–4. September, 1976.* Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck, 1977.

Klatt, D. H. Review of the ARPA speech understanding project. *Journal of the Acoustical Society of America*, 1977, *62*, 1345–1366.

Klatt, D. H. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 1979, *7*, 279–312.

Kucera, H., & Francis, W. N. *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press, 1967.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, *54*, 358–368.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. Language codes and memory codes. In A. W. Melton & E. Martin (Eds.), *Coding processes in* *human memory*. Washington, D.C.: V. H. Winston, 1972.

Lisker, L. Stop voicing production: Natural outputs and synthesized inputs. In, *Haskins Laboratories: Status report on speech research, 1976, SR-47.* (ERIC Document Reproduction Service No. ED-128-870; NTIS No. AD A031789)

Lisker, L., & Abramson, A. Cross-language study of voicing in initial stops. *Word*, 1964, *30*, 384–422.

Lowerre, B., & Reddy, D. R. The HARPY speech understanding system. In W. Lea (Ed.), *Trends in speech recognition*. Englewood Cliffs, N.J.: Prentice-Hall, in press.

Marslen-Wilson, W. D., & Welsh, A. Processing interaction and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 1978, *10*, 29–63.

Miller, G. A., Heise, G., & Lichten, W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 1951, *41*, 329–335.

Morton, J. A. Interaction of information in word perception. *Psychological Review*, 1969, *76*, 165–178.

Myerow, S., & Millward, R. SPLIT—A sound editor for a PDP-8 computer. *Behavior Research Methods and Instrumentation*, 1978, *10*, 281–284.

Naus, M. J., & Shillman, R. J. Why a Y is not a V. A new look at the distinctive features of letters. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, *2*, 394–400.

Pollack, I. Message uncertainty and message reception. *Journal of the Acoustical Society of America*, 1959, *31*, 1500–1508.

Purcell, D. G., Stanovich, K. E., & Spector, A. Visual angle and the word superiority effect. *Memory & Cognition*, 1978, *6*, 3–8.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 621–637.

Rubin, P., Turvey, M. T., & Van Gelder, P. Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics*, 1976, *19*, 394–398.

Smith, E. E., & Haviland, S. E. Why words are perceived more accurate.y than nonwords. *Journal of Experimental Psychology*, 1972, *92*, 59–64.

Stevens, K. N., & Klatt, D. H. Role of formant transitions in the voice-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, *55*, 653–659.

Warren, R. M. Perceptual restoration of missing speech sounds. *Science*, 1970, *167*, 392–393.

Warren, R. M., & Sherman, G. L. Phonemic restoration based on subsequent context. *Perception & Psychophysics*, 1974, *16*, 150–156.

Appendix A

This appendix describes quantitatively the predictions of the categorical and criterion-shift models.

The categorical model can be described by two equations. For a continuum C1, for which a response of d makes a word and t does not (such as the *dash–tash* series), the categorical model predicts

$$P(\mathrm{d}|v) = P_0(\mathrm{d}|v) + P_{\mathrm{change:C1}}*[1 - P_0(\mathrm{d}|v)]$$

where $P(\mathrm{d}|v)$ is the probability of responding d to a stimulus with voice onset time (VOT) v, $P_0(\mathrm{d}|v)$ is the probability of a response of d to a stimulus with VOT $= v$ in an unbiased situation (which is also the probability of an internal response of d before the correction process), and $P_{\mathrm{change:C1}}$ is the probability of correcting a t response to d for Continuum C1. Similarly, the predictions for a continuum in which a t response makes a word (such as the *dask–task* continuum) are given by

$$P(\mathrm{d}|v) = (1 - P_{\mathrm{change:C2}})*P_0(\mathrm{d}|v).$$

The lexical effect function $L(v)$ is simply the difference in response probabilities for the two continua of a matched pair. Thus,

$$\begin{aligned} L(v) &= P_0(\mathrm{d}|v) + P_{\mathrm{change:C1}}*[1 - P_0(\mathrm{d}|v)] \\ &\quad - (1 - P_{\mathrm{change:C2}})*P_0(\mathrm{d}|v) \\ &= P_{\mathrm{change:C1}} + P_0(\mathrm{d}|v)*(P_{\mathrm{change:C2}} - P_{\mathrm{change:C1}}) \end{aligned}$$

In Figure 1, it is assumed that $P_{\mathrm{change:C2}} = P_{\mathrm{change:C1}}$, so $L(v)$ is constant (i.e., does not depend on v). This is not an essential feature of the model; there is no reason to think biases are equal. However, for any value of $P_{\mathrm{change:C2}} - P_{\mathrm{change:C1}}$, the value of $L(v)$ at the boundary (where $P_0(v) = .5$) will be equal to the mean of the values of $L(v)$ at the ends of the continuum and certainly less than their sum.

The quantitative predictions of the criterion-shift model are described by

$$P(\mathrm{d}|v) = P_0(\mathrm{d}|v + b_{\mathrm{C1}}),$$

where $P_0(\mathrm{d}|x)$ is the probability of responding d to a stimulus with VOT $= x$ in a situation in which there is no lexical bias, and b_{C1} is the criterion shift for this continuum. b_{C1} will be negative when the voiced end of Continuum C1 is a word (resulting in more voice responses) and positive when the voiceless end is a word.

The shape of $L(v)$, according to the criterion shift model, is simply

$$L(v) = P_0(\mathrm{d}|v + b_{C1}) - P_0(\mathrm{d}|v + b_{C2}).$$

Figure 2 shows the predictions of the criterion-shift model, assuming that P_0 $(\mathrm{d}|v)$ is a cumulative normal function. Figure 2 assumes equal but opposite direction lexical biases for the two continua, as did Figure 1. Assuming that $P_0(\mathrm{d}|v)$ is sigmoid, with steepest slope near the phoneme boundary, $L(v)$ will generally reach a maximum in the neighborhood of the phoneme boundary and, thus, be greater near the phoneme boundary than at either end of the continuum. This is unlike the shape of $L(v)$ according to the categorical model, for which the value at the boundary must be between the values at the ends of the continua.

Appendix B

I am aware of only two previous studies (Lisker, 1976; Spencer & Halwes, Note 3) that used voice onset time (VOT) continua produced by cross splicing natural speech. Since these studies are not readily available, it seems important to describe the technique in some detail.

The splicing method uses segments cut from voiced and voiceless natural tokens to supply the aspirated and voiced segments of each stimulus. Good tokens of each stimulus are recorded and digitized, and pitch periods are marked in the voiced stimulus. Stimuli from the VOT continuum are produced by replacing the segment before a particular pitch period in the voiced token with an equal duration segment from the beginning of the voiceless token. For these purposes, the beginning of a pitch period is taken to be the last upward-going zero crossing before a pitch pulse. This choice minimizes clicks due to splicing and assures that a pitch pulse occurs at each nominal VOT. Stimuli produced by this method can only have VOTs at times that are the beginnings of pitch periods in the voiced stimulus.)

For the present experiment, tokens of the voiced and voiceless syllables used to construct the continua were spoken by a female phonetician, in the sentence environment "Now say . . .," and recorded on audiotape. Her fundamental frequency was approximately 200 Hz at the beginning of each syllable. The tokens were digitized on a PDP-8 minicomputer, using a sampling rate of 10 kHz and a 4.5-kHz lowpass filter. The waveform editing program SPLIT (Myerow & Millward, 1978) was used to edit tokens out of the carrier phrase, and the beginning of each token's burst was carefully determined. Zero crossings before pitch periods were located by examining an oscillographic display. The digitized tokens and the locations of the pitch periods were used to generate the spliced tokens that subjects heard.

Tokens were digitized for 22 continua pairs. The endpoint stimuli from those continua were presented to 13 subjects for identification of the whole syllable (not just the initial segment). Those continua pairs for which more than five of the endpoints were incorrectly identified were eliminated. The errors in identification rarely involved the voicing of the initial segment. Often errors were due to misperceptions of the vowel or final consonants. The remaining 16 continua pairs were used in a pilot experiment. Two of these pairs contributed many more errors in identification of the endpoints than did the other pairs, so these 2 pairs were also eliminated. The remaining continua pairs (whose word endpoints and VOTs are given in Table 1) were used in Experiment 2. It is important to note that the selection of stimuli was on grounds independent of whether they produced the lexical effect. The excluded continua were not eliminated because they failed to show a word bias, but simply because the endpoint stimuli were not intelligible enough. Thus, there is no reason to think that the remaining continua pairs are not a representative sample from the set of intelligible continua pairs for English.

In the pilot experiment, there was considerable variation in the position of phoneme boundaries within the pairs of continua. One source of this variation could be random fluctuations in the particular aspiration and voicing waveforms used. Random variations in the amplitude of segments of aspiration noise or glottal pulses could influence the position of the phoneme boundary. Therefore, it was decided to use exactly the same aspiration and voicing waveforms to construct the stimuli of both continua of each pair. The first pitch pulse more than 100 msec after stimulus onset was found in the voiced token of one of the continua. In both voiced and voiceless tokens, the initial segment of this duration was replaced with a comparable segment from the corresponding member of the other continuum. The tokens produced sounded perfectly natural but produced pairs of continua that used exactly the same aspiration and voicing sequences for the first 100 msec of the stimuli. For half of the continua pairs, the initial segments were taken from the continuum whose voiced end was a word, and vice versa for the other half.

One objection to the use of stimuli produced in this manner can be raised. Although VOT is clearly controlled by this method, other acoustic cues (e.g., formant transition duration, Stevens & Klatt, 1974) that are known to affect the voicing decision for stops are not controlled systematically. This objection is correct but irrelevant to the purposes of the present study. If the goal for the study were to investigate just the acoustic cue VOT (and not other acoustic cues), this would be a serious problem. But in the present study, it is only important to relate the perception of stimuli varying along some acoustic variable to other factors; it is not important that this variable be any particular acoustic cue. As long as subjects' perceptions of the stimuli depend on the acoustic variable (as they evidently did in this experiment), the stimuli will be sufficient for examining the relation between acoustic information and higher order knowledge.

Received January 12, 1979 ■