Perception & Psychophysics
1978, Vol. 23 (4), 331-336

On the relationship between vowel and consonant identification when cued by the same acoustic information

PAUL MERMELSTEIN
Haskins Laboratories, New Haven, Connecticut 06511

When listening to speech, do we recognize syllables or phonemes? Information concerning the organization of the decisions involved in identifying a syllable may be elicited by allowing separate phonetic decisions regarding the vowel and consonant constituents to be controlled by the same acoustic information and by looking for evidence of interaction between these decisions. The duration and first formant frequency of the steady-state vocalic segment in synthesized consonant-vowel-consonant syllables were varied to result in responses of /bɛd/, /bæd/, /bɛt/, and /bæt/. The fact that the duration of the steady-state segment controls both decisions implies that that segment must be included in its entirety in the signal intervals on which the two decisions are based. For most subjects, no further significant interaction between the vocalic and consonantal decision is found beyond the fact that they are both affected by changes in the duration parameter. A model of two separate and independent phonetic decisions based on overlapping ranges of the signal adequately accounts for these data, and no explicit syllable level recognition needs to be introduced.

When listening to speech, do we recognize syllables or phonemes? The question of minimal perceptual units in speech perception has been thoroughly investigated. The dependence of consonantal place of articulation on the succeeding vowel in consonant-vowel syllables (Delattre, Liberman, & Cooper, 1955) argues for the indivisibility of such units. Studies on the effect of the preceding vowel duration on voicing of the final consonant in CVC syllables (Raphael, 1972) lead to similar conclusions. Experiments on backward recognition masking (Massaro, 1974; Pisoni, Note 1) offer further support for the above by revealing that consonant recognition is incomplete until a significant part of the vowel has been heard by the listener. Liberman (1970) has interpreted these findings by viewing the syllables as linguistic units within which phones exist only by way of their constituent features. Features carrying information about any one phoneme are generally temporally distributed, and features signaling separate phonemes overlap significantly. If the interaction between phonetic units exists entirely at the acoustic level, one may establish the criteria for each phonetic decision of a listener in terms of the acoustic variables alone. To succeed in this effort, one requires that the decision processes giving rise to the distinct phonetic components of the syllables be independent. However, if there exists additional interaction between the phonetic processes, then each decision will be determined not only by the current value of the acoustic features but also by all other phonetic decisions based on those features. Strictly hierarchic models of language understanding cannot account for such interaction between separate decisions on the same level. As long as such interactions are weak, models with feedback can be useful explanatory tools, because they allow the output of one phonetic decision to be used as the input to a second one. If strong interactions are found, the usefulness of viewing distinct phonetic decisions based on the same acoustic information as separate must be questioned. Under such conditions, it would be a more parsimonious approach to admit only the existence of higher level units and to deny the relevance of direct perception of phones.

This paper reports on an experiment in which the interaction of vocalic and consonantal decisions was explored in a linguistic environment where the same set of acoustic parameters determined both decisions. We synthesized C_1VC_2 syllables where the spectral and temporal properties of the steady-state vocalic segment controlled the identity of V as well as the voicing feature of C_2. The boundaries in spectral-temporal space follow similar curves for most listeners, despite the fact that the 50% response lines may be significantly displaced from each other. Since the same information, the duration of the steady-state vowel segment, affects both the vocalic and the con-

Anne Fowler's help with the generation of stimulus tapes and collection of the data is much appreciated. Quentin Summerfield provided many helpful comments on this manuscript. This work was supported in part by the National Institute of Child Health and Human Development, Grant HD-01994.

sonantal decisions, it was of interest to explore the interaction, if any, between these two decisions. Results of a pilot experiment indicated that short vowel durations result in a predominance of responses comprising the shorter vowel and an unvoiced stop. At longer durations, the syllable comprising the longer vowel and the voiced stop predominates. This led us to examine intermediate duration values where the information regarding both the vowel and the consonant is ambiguous and to look for interaction between the two decisions.

The boundaries for vowel perception in spectral-temporal space are of interest themselves, aside from any interactions between separate phonetic decisions. They allow a reexamination of previous suggestions based on speech-production data, namely that the listener takes into account the dynamic aspects of the speakers' production to interpret the spectral information (Lindblom, 1963). Active correction processes have been suggested (Halle & Stevens, 1962) as possibly mediating the task in an attempt to account for perceptual invariance of vowels in the face of significant spectral-temporal variations. Our results, insofar as durational variations of isolated CVCs allow one to make inferences concerning durational variations that may arise due to stress and rate, do not support the existence of such correction processes.

THEORY

The interdependence of the perceptual boundary between vowels along the spectral and temporal dimensions is of interest because of the light it may shed on the possible awareness of a listener regarding the speaker's production constraints. Lindblom (1963) has analyzed spectrographic data on the shift in vowel formants of speakers as a function of vowel length. He reported an "undershoot of the vowel target" at short durations. His interpretation of the data reads as follows: "The talker does not adjust control of his vocal tract at fast rates to compensate for its response delay. His strategy of encoding is clearly not intended for a listener who demands absolute acoustic invariance in the realization of phonemes but it presupposes that the listener is able to correct for coarticulation effects." Such an interpretation is in line with the analysis-by-synthesis theory of speech perception (Halle & Stevens, 1962). The first formant boundary between vowels can serve as a reliable indicator of any such correction. Existence of such active correction processes would predict an upward shift in first-formant frequency at the boundary with increasing vowel duration. A shorter vowel, on the assumption of insufficient time to reach the intended articulation, would be judged phonemically identical to a more open longer vowel that possesses a higher first-formant frequency.

That vowel duration affects the judgment of voicing of the postvocalic stop is well known (Denes, 1955; Raphael, 1972). The dependence of the same voicing feature on the spectral configuration of the vowel has received less attention. The more open the vowel, the greater change in F_1 is incurred in moving to the appropriate articulatory configuration from a stop configuration (Kent & Moll, 1969). Since longer vowel duration favors a voiced-consonantal percept, one expects to find more rapid spectral changes preceding an unvoiced stop. In prevocalic position, a more rapid spectral change has, in fact, been found by Stevens and Klatt (1974) to favor the perception of a voiceless stop. Subsequently, Summerfield and Haggard (1977) found that it was the higher F_1 onset frequency that acted as a positive cue for the perception of voiceless prevocalic stops. We may suspect, therefore, that if the postvocalic vowel-consonant transition is fixed in duration, but its starting F_1 frequency is increased, the likelihood of hearing a voiceless stop will also be increased. In turn, this could lead to an increase in the vowel-duration boundary that separates voiceless and voiced stops as F_1 of the vowel is increased.

The ability to control distinct phonemic decisions by varying the stimuli along the same acoustic continuum offers potentially interesting opportunities to explore the nature and interdependence of these decisions. As long as the distinct phonemic decisions are cued by a multiplicity of acoustic cues, any interaction found between the decisions may be due to the interaction of the acoustic cues themselves. One could argue that the listener extracts continuous, but complex, higher-order features from the simultaneous values of the distinct acoustic cues and arrives at phonemic decisions based on the values of these higher-order features. When the separate decisions are based on the same acoustic information and are found not to be independent, we are led to conclude that the outcome of one phonetic decision influences the outcome of a second such decision. If, on the contrary, these decisions appear to be independent, we are free to conclude that separate phonetic decision processes adequately model the decoding of simple CVC stimuli.

Since we are concerned with the phonetic processing of syllable-sized stimuli, any interaction we find as a result of temporal variation of individual segments may provide evidence concerning an auditory segmentation of the stimulus prior to arriving at phonemic decisions. We may postulate a process that segments the stimulus in time somewhere in the vocalic region and assigns the vowel and consonantal categories on the basis of the duration of the resultant segments. Such a segmentation hypothesis would predict an interaction that increases the likelihood of a vowel decision appropriate to a longer vowel segment cooccurring with a consonantal deci-

sion appropriate to a shorter consonantal segment. Thus, if increasing the overall temporal duration increases the likelihood of hearing V_1 rather than V_2 and that of C_1 rather than C_2, the joint probabilities would be expected to follow the relation

$$p(V_1, C_1) < p(V_1) \cdot p(C_1),$$

but

$$p(V_1, C_2) > p(V_1) \cdot p(C_2).$$

EXPERIMENT 1
IDENTIFICATION OF VOWELS AND CONSONANTS

Method

Listeners were asked to identify synthetic speech stimuli as [bɛd], [bæd], [bɛt], or [bæt]. The /ɛ-æ/ vowel contrast was selected because of the strong effect of vowel duration on the identification of those vowels (Stevens, 1959). The /d-t/ consonant contrast was selected to exemplify the known dependence of the perception of voicing of a postvocalic consonant on the vowel duration (Denes, 1955). The initial /b/ was added to all syllables so that they represented common English words.

The stimuli were prepared with the aid of a programmed (software) synthesis program with three adjustable formants. The program allowed the specification of acoustic segments in terms of duration, initial and final formant frequencies, formant frequency contours (linear, parabolic, or cubic functions), voicing amplitude onset or offset, and formant bandwidths. Starting with a series formant synthesis configuration, by addition of the individual formant bandwidth controls and overall-amplitude offset relative to the formant trajectory, the final transition was adjusted to give rise to voiced or voiceless stops depending on the vowel duration. The initial and final formant trajectories were linear with time and lasted 48 msec each. Initial and final formant frequencies corresponding to the stop articulations were set at 100, 1,000, and 2,000 and 100, 2,000, and 3,000 Hz, respectively. The variables controlled were the duration of the central stationary vocalic segment D^V and its first formant frequency F_1^V. The second and third formants of the steady-state segment were set at 1,800 and 2,500 Hz, respectively. Two higher formants were fixed at 3,500 and 4,500 Hz. The formant bandwidths were set at 60, 80, and 100 Hz except for the final transition. There the first formant bandwidth was adjusted to result in a first formant level whose drop-off increased linearly in magnitude to -20 dB over the extent of the transition. As a result of preliminary experiments, the F_1^V range was set to 625-675 Hz with 25-Hz increments. Ten distinct duration values were used for D^V, namely 48, 72, 80, 88, 96, 104, 112, 120, 160, and 240 msec. All duration values were integral multiples of the constant excitation duration of 8 msec, corresponding to a fundamental frequency of 125 Hz.

Stimuli were presented as identical pairs separated by a 0.5-sec pause. Five different randomizations of the set of 30 stimuli resulted in 150 presentations and responses by each listener. Eight listeners were selected, and all had at least some phonetic training so that they would be likely to make a consistent /ɛ/ vs. /æ/ vowel judgment.

Results

The vowel and consonant identification results of each subject were pooled separately and analyzed in the form of probability of identification as functions of duration at specified F_1^V values. Ogives were fitted to the frequency of /æ/ responses collapsed over both consonants and to the frequency of /d/ responses collapsed over both vowels.

The duration boundary values obtained by the above method are shown in Figure 1. The duration boundary for vowel identification is found to decrease with an increase of the F_1^V frequency for all subjects. The slope of the boundary ranges between 1 and 3 msec/Hz. As expected, increasing F_1^V favors the likelihood of /æ/ responses and there exists a tradeoff relation between F_1 value and duration. The likelihood of hearing a relatively shorter but more open vowel as /æ/ can equal that of a relatively longer but less open vowel.

The variation in the duration boundary for consonant-voicing with changes in F_1^V is much less. Nevertheless, four out of the eight subjects showed an increase in the /t-d/ duration boundary for F_1^V values ranging between 625 and 675 Hz, which is significant at the .05 level. For no subject can we find a significant decrease in the same boundary value. Evidently, an increase in F_1^V while the endpoint of the final transition is kept fixed is a weak negative cue for the voicing feature of the postvocalic stop. As F_1^V is increased, the stop is more likely to be heard as voiceless, unless the D^V duration is also increased.

EXPERIMENT 2
INTERACTION BETWEEN VOCALIC AND CONSONANTAL DECISIONS

Method

A second experiment was designed to probe more deeply into the interdependence between vocalic and consonantal decisions. The results shown in Figure 1 reveal a clear vowel-segment dura-

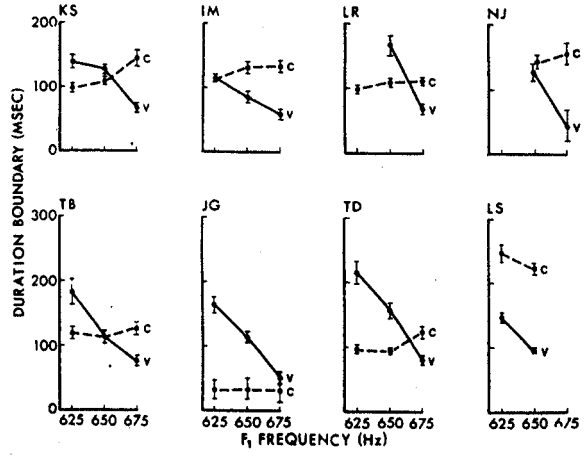

Figure 1. Variation of duration boundaries for vowel /ɛ-æ/ and consonant /t-d/ identification with F_1 of steady-state segments.

tion boundary in the perception of both the vowels and the post-vocalic stops. For most subjects, a value can be found for the frequency of F_1 where the same duration value represents a perceptual boundary between the vowels and the consonants. Both the frequency and the duration values that together correspond to this condition vary from listener to listener. To probe the interdependence of these decisions in the region of maximum ambiguity, a simple adaptive procedure was designed that modified the parameters of each syllable-stimulus presented to the listener based on his previous syllable-identification response. The F_1^v frequency was modified based on the vowel response, the duration of the steady-state vowel segment was modified based on the consonant-voicing response. The magnitude of the change in each parameter was constant, 10 Hz in F_1^v and 8 msec in D^v. Thus, a /bɛt/ response led to a +10 Hz change in F_1^v and a +8 msec change in D^v, while a /bæt/ response was followed by a -10 Hz change in F_1^v and a +8 msec change in D^v. The initial values for the parameters were set at 650 Hz and 104 msec, respectively.

The experiment was carried out on-line at the computer and listeners could ask for the same stimulus to be repeated as many times as they wished. Responses were indicated by keying the last two orthographic characters of the syllable at the terminal keyboard. The experiment continued until 100 responses were collected from each subject.

Eleven subjects, all native speakers of English with no known hearing disabilities, participated in the experiment. Three subjects had previously participated in the off-line identification experiments, eight were new to the syllable identification task. Six subjects were unpaid volunteers from the laboratory staff, five subjects were students at Yale University and were paid $2 for their participation.

Results

The responses of 10 subjects are shown in Table 1. Only responses to stimuli which were not perfectly consistently identified by the listener are included in the table. This restriction has the effect of excluding responses collected initially when the system adapted to the listener's ambiguity region and responses from the extremes of his ambiguity region where interaction due to duration cueing both vowel and consonant responses was more likely. One subjects's results lacked sufficient consistency to allow an ambiguity region to be identified and his responses were eliminated from subsequent consideration.

Table 1
Summary of Responses for Vowel-Consonant Interaction Tests

Subject	æt	ɛt	æd	ɛd	χ^2	Significant at .05
AS	20	20	18	24	.42	No
AF	23	18	24	25	.45	No
GN	23	21	16	25	1.5	No
DK	18	18	16	19	.13	No
TB	22	25	20	21	.03	No
LR	11	28	30	15	−12.4	Yes
DW	12	16	15	16	.18	No
JG	5	23	20	6	−18.9	Yes
AL	29	16	15	33	10.2	Yes
MS	30	15	17	31	9.1	Yes
Total	193	200	191	215	.34	No

Typical results, such as those of subject A.F., are shown in Figure 2. The responses from the outlined ambiguity region were collected and one interaction test was performed on the total responses of each syllable to the stimuli contained in that region. Such pooling of responses can be justified when the region of stimulus parameters is sufficiently small and insufficient responses are available to any one stimulus to obtain reliable separate interaction estimates. A one degree of freedom chi-square test at the .05 level of significance yielded no interaction for six subjects and oppositely directed interaction results for two groups of two subjects. The lack of significant interaction leads to a model of phonetic processing where the syllable response is made up of inde-

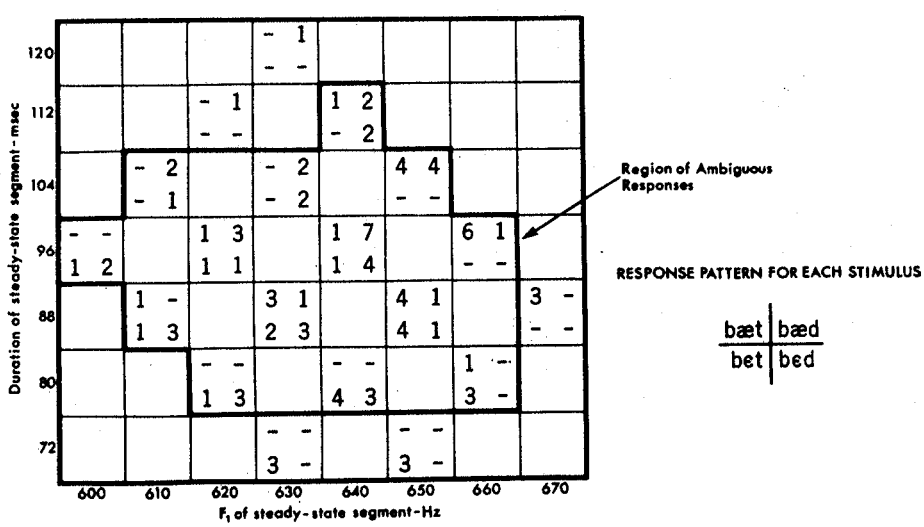

Figure 2. Response pattern for subject A.F. in interaction test.

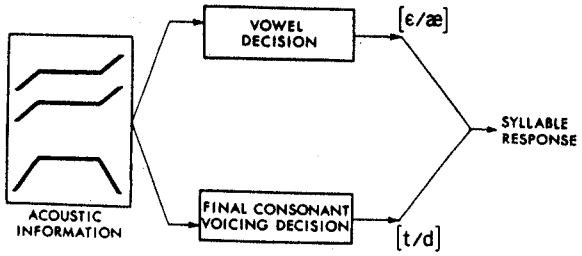

Figure 3. Model of vowel and consonant decisions operating on same acoustic signal to yield a syllable response.

pendent vowel and consonant decision processes operating on overlapping acoustic information. Figure 3 illustrates such a model.

Differences between the response patterns of individual subjects are large. Yet interaction pooled among all subjects is not significant. We have no information, as yet, on the sources that give rise to the divergence of results between individual subjects. Linguistic background, phonetic training, degree of attention paid to the experiment could be just some of the factors.

CONCLUSIONS

Vowel and consonant identification in synthesized syllables can both be controlled by variations in the same acoustic parameter, namely the duration of the steady-state vowel segment. The standard errors for the equal identification probability values are not significantly different in the two cases. It is therefore highly unlikely that, as suggested by Pisoni (Note 2), different processes are invoked for duration discrimination for the purposes of identification of vowels and consonants. When linguistic categories are imposed on both the vowel and consonant continua, identification as a function of duration is comparable.

Our results generally support a view of syllable perception where independent phonetic decision processes operate on overlapping segments of the acoustic signal. These results are in agreement with those of Huggins (1968), who also failed to find perceptual compensation in timing between vowel and consonant in the same syllable. Temporal segmentation of the syllable segment into vowel and consonant parts would predict a negative interaction between the probability of hearing a longer vowel and the probability of hearing a voiced stop. For only 2 of 10 subjects is such correlation found at a significance level exceeding .05. For the other 8 subjects, no such interaction is found, so a segmentation model can be rejected as not generally appropriate for the interpretation of the results obtained above.

Draper and Haggard (1974), on the basis of similar interaction tests between the place and voicing features of the same consonant, suggest a feedback model between the various feature decisions required to establish consonant identity. In contrast, our results indicate an absence of significant interactions for most subjects between vowel and postvocalic voicing decisions. We must therefore conclude that in our experiments the decisions are carried out independently and without significant participation of feedback processes.

Our results offer no support to an interpretation that a listener makes allowance for the temporal production constraints of the speaker as suggested by Lindblom (1963). Lindblom's suggestion would predict that increased duration acts as a negative cue for the more open vowel. If the listener assumes that the speaker's temporal constraints prevented him from reaching his intended target articulation, he should interpret a decrease in duration as a positive cue for the more open vowel. For all subjects, the duration boundary for vowel identification showed contrary results. The spectral-temporal interaction is such that both longer duration and increased F_1 act as positive cues for the more open vowel. Unpublished experiments by this author reveal similar results for these vowels when they are presented, not in the context of a CVC syllable, but in isolation.

The predictions based on temporal production constraints pertain more directly to talking rate than to duration differences. Lindblom (1963) used rate and stress to control duration in the data he analyzed. Now, Fujisaki, Nakamura, & Imoto (1975) show that talking rate influences the identification boundary for Japanese vowels when that boundary is based on duration. Admittedly, length and spectral differences may be more independent for Japanese vowels than for English vowels. Nevertheless, they find that listeners adjust their perceptual boundary in accordance with the talking rate, i.e., the duration boundary is shorter at higher talking rates than at lower rates. Our results show that the duration boundary is also shorter at higher F_1 values. Therefore, our results would lead one to expect that at faster talking rates, which result in shorter durations, the F_1 vowel identification boundary would be higher than at slower rates. This is contrary to Lindblom's prediction. One might possibly argue that by presenting stimuli regularly spaced in time, our experimental conditions did not allow the listener to establish differential expectations concerning the synthesized vowel durations. However, transformation of increased duration from a positive cue for open vowels in isolation to a negative cue when these vowels are embedded in connected speech appears an unlikely possibility.

REFERENCE NOTES

1. Pisoni, D. B. *Perceptual processing time for consonants and vowels.* Haskins Laboratories Status Reports on Speech Research SR-31/32, 1972, 83-92.

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

REFERENCES

DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 769-773.

DENES, P. Effects of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 1955, **27**, 761-768.

DRAPER, G., & HAGGARD, M. Facts and artifacts in feature independence. *Proceedings of Speech Communication Seminar*, Stockholm, 1974, 197-219.

FUJISAKI, H., NAKAMURA, K., & IMOTO, T. Auditory perception of duration of speech and nonspeech stimuli. In G. Fant & M. A. A. Tatham (Eds.), *Perception of speech*. London: Academic Press, 1975.

HALLE, M., & STEVENS, K. N. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, IT-8, 1962, 155-159.

HUGGINS, A. W. F. The perception of timing in natural speech I; compensation within the syllable. *Language and Speech*, 1968, **11**, 1-11.

KENT, R. D., & MOLL, K. L. Vocal-tract characteristics of the stop cognates. *Journal of the Acoustical Society of America*, 1969, **46**, 1549-1555.

LIBERMAN, A. M. The grammars of speech and language. *Cognitive Psychology*, 1970, **1**, 301-323.

LINDBLOM, B. E. F. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, **38**, 1773-1781.

MASSARO, D. W. Perceptual units in speech perception. *Journal of Experimental Psychology*, 1974, **102**, 199-208.

RAPHAEL, L. J. Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. *Journal of the Acoustical Society of America*, 1972, **51**, 1296-1303.

STEVENS, K. N. Effect of duration upon vowel identification. *Journal of the Acoustical Society of America*, 1959, **31**, 109(A).

STEVENS, K. N., & KLATT, D. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, **55**, 653-659.

SUMMERFIELD, Q., & HAGGARD, M. On the dissociation of spectral and temporal cues to the voicing decision in initial stop consonants. *Journal of the Acoustical Society of America*, 1977, **62**, 435-448.

(Received for publication July 1, 1977;
revision accepted January 11, 1978.)