Perception & Psychophysics
1974, Vol. 16 (2), 201-204

Contingent feature analysis in speech perception*

WILLIAM E. COOPER
Massachusetts Institute of Technology, E10-044, Cambridge, Massachusetts 02139

A contingent adaptation effect is reported for speech perception. Experiments were conducted to test the effects of an alternating sequence of two adapting syllables, [da] and [t^hi], on the perception of two series of synthetic speech syllables, [ba]-[p^ha] and [bi]-[p^hi]. Each of the test series consisted of 11 stimuli varying in voice onset time, a cue which distinguishes voiced from voiceless stop consonants in word-initial position. The [da]-[t^hi] adapting sequence produced opposite shifts in the loci of the phonetic boundaries for the two test series. For the [ba]-[p^ha] series, listeners made fewer identification responses to the [b] category after adaptation, while for the [bi]-[p^hi] series, listeners made more responses to the [b] category. The opposing shifts indicate that the perceptual analysis of voicing in stop consonants is carried out with respect to vowel environment.

An important feature of speech sounds is the consonant feature *voicing*, which marks phonemic contrasts in virtually all natural languages (cf. Lisker & Abramson, 1964). In English, voicing serves to minimally distinguish voiced from voiceless stop consonants; i.e., [b] from [p], [d] from [t], and [g] from [k].

Voicing distinctions are signaled in speech production by voice onset time (VOT), a temporal relation between the onset of vocal cord vibration and the release of oral closure.[1] Voiceless stops are normally produced with an onset of vocal cord vibration that lags the release of closure by more than 40 msec, whereas voiced stops are produced with an earlier onset of voicing. Acoustically, VOT is manifest as the onset of first formant energy relative to the onset of higher formant energy in the spectrum, with the higher formants being excited by a noise source rather than a periodic source during the interval when the first formant is absent. With speech synthesis techniques, it is possible to construct a series of consonant-vowel stimuli which vary in VOT and which range from consonants perceived as voiced stops to those perceived as voiceless stops (cf. Lisker & Abramson, 1970).

Unlike some of the feature cues for consonants, VOT does not vary appreciably as a function of the preceding or following vowel, at least according to the original studies on speech production of VOT (cf. Lisker & Abramson, 1967). One might be led to assume, then, that the perception of voicing is primarily accomplished by feature detectors that extract VOT information from consonants irrespective of their vowel environment. The results of the present study indicate that, on the contrary, voicing detection is carried out in a vowel-dependent manner.

Prior experiments have shown that the perception of voicing in stop consonants can be selectively adapted (Eimas & Corbit, 1973; Eimas, Cooper, & Corbit, 1973). Adaptation with a voiced stop ([ba] or [da]) produced a shift in perception such that listeners made fewer [b] or [d] responses than normal when identifying stimuli varying in VOT ([ba]-[p^ha] or [da]-[t^ha]). Repetitive presentation of a voiceless stop ([p^ha] or [t^ha]) produced a shift in the opposite direction, such that listeners made more [b] or [d] responses after adaptation.

In the present study, selective adaptation experiments were conducted in which both the voicing of the consonant and the vowel environment were varied systematically. Using this procedure, it was possible to test for the presence of contingent adaptation effects along the voicing dimension that could not be attributed to adaptation of the consonant feature of voicing alone.

METHOD

Stimuli

The stimuli were five-formant synthetic speech patterns. All stimuli were CV syllables generated on a terminal analog speech synthesizer at the M.I.T. Research Laboratory of Electronics.[2] The test stimuli included 22 syllables, 11 [ba]-[p^ha] stimuli and 11 [bi]-[p^hi] stimuli. In each series, the 11 stimuli varied from one another in VOT. The VOT variations were in equal steps of 5 msec, covering a range of from +5 to +55 msec of VOT (positive VOT values indicate that the onset of voicing occurred subsequent to the release burst). Relative formant amplitudes were set according to the acoustic theory of speech production (Fant, 1956). The overall amplitudes of the stimuli in the [ba]-[p^ha] series were thus approximately 7 dB greater than the amplitudes of stimuli in the [bi]-[p^hi] series. The fundamental frequency contour and all other acoustic parameters not directly associated with the formant frequencies and amplitudes were equivalent for the stimuli in the two series. The duration of each stimulus was 255 msec.

In addition to the 22 test stimuli, two adapting stimuli were synthesized, [da] and [t^hi]. The adapting syllables contained VOT values of +5 and +55 msec, respectively, and were 255 msec in duration. The fundamental frequency contour, relative formant amplitudes, steady state formants, and bandwidths of each adapting stimulus corresponded to the parameter values of the test stimuli having the same vowel.

*This research was supported by NIH Training Grant NIH-5-T01-GM01064-12 to the Department of Psychology, M.I.T., and by the Research Laboratory of Electronics, M.I.T. I am grateful to Dennis H. Klatt for his instructions on speech synthesis procedures and for criticisms of the manuscript. Kenneth N. Stevens and Merrill F. Garrett also provided helpful comments.

Table 1
Individual and Mean Phonetic Boundary Loci for Each Test Condition (in Milliseconds of VOT)

	[ba]-[p^{h}a] Series		[bi]-[p^{h}i] Series	
		With		With
	No	[da]-[t^{h}i]	No	[da]-[t^{h}i]
S	Adapt	Adapt	Adapt	Adapt
A.C.	27.4	20.7	28.0	32.3
J.C.	26.1	27.9	31.1	35.4
L.C.	20.1	23.3	21.7	26.7
L.Ch.	23.5	18.8	28.5	25.2
P.F.	22.0	7.4	38.9	36.2
S.G.	31.0	33.8	31.2	30.4
B.H.	20.2	19.3	38.3	32.8
H.H.	30.6	25.5	36.3	36.2
P.H.	29.4	26.1	34.4	35.0
J.I.	26.2	17.3	21.2	31.3
J.K.	36.7	33.6	36.1	39.1
T.M.	27.0	25.8	31.9	37.3
C.S.	29.8	21.2	37.9	40.2
T.S.	16.4	16.0	30.0	36.0
Mean	26.2	22.6	31.8	33.9

Subjects

Fourteen M.I.T. undergraduates served as naive listeners in the experiment. All were native speakers of English. None had prior phonetic training or any known hearing impairment.

Procedure

To obtain a measure of perception in the unadapted state, Ss were first instructed to identify stimuli from each of the two test series. The stimuli of each series were presented in random order with an interstimulus interval of 2.5 sec. The listeners first identified the stimuli of the [ba]-[p^{h}a] series and then the stimuli of the [bi]-[p^{h}i] series. Ss listened binaurally over Telex headphones to the output of an Ampex PR-10 tape recorder. They were instructed to write "B" or "P" as their response, and to make the response as soon as possible after hearing each stimulus item. For each S, a total of 20 responses was obtained to each of the 22 test stimuli. All baseline testing was conducted during a single hour-long session.

Starting with the day after baseline testing, Ss were presented two adaptation tests, one test per day on consecutive days. In each test, adaptation consisted of listening to repetitions of an alternating sequence of the two syllables [da] and [t^{h}i], each syllable being presented three times in succession before a changeover to the alternate syllable. The interstimulus interval for repetitions of the adapting stimuli was 350 msec. After each minute of adaptation, Ss were presented four randomly selected test stimuli, each stimulus to be identified as either B or P. An interstimulus interval of 1.5 sec intervened between the last repetition of the adapting sequence and the first stimulus to be identified; adjacent stimuli to be identified were separated by 2.5 sec of silence. After the fourth stimulus was presented for identification, 5 sec of silence intervened before the onset of the next adaptation trial. All timing intervals were specified by computer, and the entire adaptation sessions were recorded onto tape.

In one adaptation test, the stimuli to be identified consisted of the [ba]-[p^{h}a] series, and in the other adaptation test, the [bi]-[p^{h}i] series. Seven Ss were presented the adaptation test with the [ba]-[p^{h}a] stimuli first; the other seven were presented the two adaptation tests in reverse order. Each of the adaptation tests lasted about 1 h. For each S, 10 responses were obtained to each of the 22 test stimuli during the adaptation tests.

RESULTS AND DISCUSSION

The results appear in Table 1, where phonetic boundary loci are computed by a least-mean squares analysis (Ferguson, 1959) for each listener. The baseline results were similar overall to those obtained by Lisker and Abramson (1970) (see Fig. 1). Importantly, the phonetic boundary loci of the two series in the baseline test differed significantly from each other ($p < .01$).[3] This difference provided strong initial evidence that VOT is perceived in a vowel-dependent manner. The fact that the VOT boundary locus for the [bi]-[p^{h}i] series was greater than that for the [ba]-[p^{h}a] series is correlated with an interesting recent finding of Klatt (1973). In contrast to Lisker and Abramson (1967), Klatt has found that VOT values for the production of voiceless stops are significantly greater when these stops are followed by high as opposed to low vowels (e.g., [p^{h}i] vs [p^{h}a^y]).

The vowel-dependent nature of VOT perception was also demonstrated in the two adaptation tests of the present study. Adaptation with the [da]-[t^{h}i] alternating sequence produced opposite shifts in the

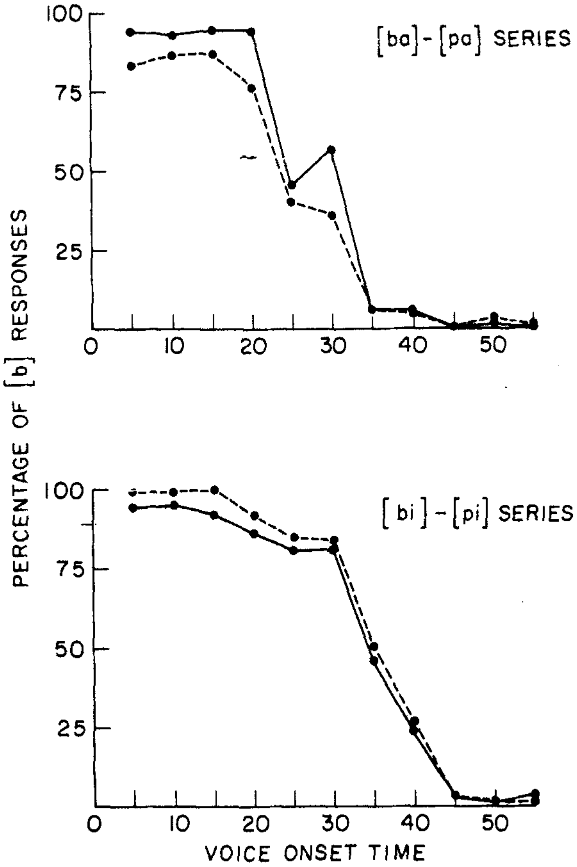

Fig. 1. The percentage of "B" identification responses for the group of 14 Ss. Solid lines denote the identification functions obtained in the unadapted state. The dotted lines denote the identification functions obtained after adaptation to the alternating [da]-[t^{h}i] sequence.

identification functions of the [ba]-[p^{h}a] and [bi]-[p^{h}i] series. The results for the [ba]-[p^{h}a] series indicate the presence of a small, but significant, shift in the locus of the phonetic boundary toward the [b] category, as compared with the baseline performance for the same series ($p < .01$). Ss thus made fewer [b] responses to the [ba]-[p^{h}a] stimuli after adaptation with the [da]-[t^{h}i] alternating sequence. The results for the [bi]-[p^{h}i] series show an opposite shift in the locus of the phonetic boundary after adaptation with the [da]-[t^{h}i] sequence. The shift in this case was in the direction toward the [p] category ($p < .05$). The difference between the two shifts was highly significant ($p < .005$), the net shift for 12 of the 14 listeners being in the same direction.[4]

To account for the obtained shifts, it is proposed that the identification of stimuli in each of the two VOT series was primarily governed by that member of the adapting sequence which contained the same vowel as the test stimuli. In effect, perception of the [ba]-[p^{h}a] stimuli was primarily influenced by the adapting stimulus [da], while the perception of the [bi]-[p^{h}i] stimuli was more highly influenced by the adapting stimulus [t^{h}i]. The differential effectiveness of the two adapting stimuli cannot be accounted for by an adaptation effect presumed to operate on the perception of the consonant feature of *voicing* alone. The effects observed here must rather by attributed to adaptation that operates on *voicing* perception in a vowel-contingent manner.

Further support for this vowel-contingent interpretation was provided in a second experiment. Five listeners, none of whom had served in the main experiment, were presented the [ba]-[p^{h}a] and [bi]-[p^{h}i] series for identification in the unadapted state and after adaptation to an alternating sequence of the two syllables [di] and [t^{h}a]. As predicted by the vowel-contingent hypothesis, it was found that after adaptation with the alternating [di]-[t^{h}a] sequence, listeners assigned fewer responses to the [b] category of the [bi]-[p^{h}i] series but assigned more responses to the [b] category of the [ba]-[p^{h}a] series (mean boundary shift = 3.4 msec VOT). Here, as in the main experiment, the perception of each test series was primarily governed by the adapting stimulus which contained the same vowel as the test stimuli.

It was considered possible that the contingent effects observed in this study share some properties with the contingent aftereffects demonstrated in human vision (cf. McCollough, 1965; Harris & Gibson, 1968; Fidell, 1970; Held & Shattuck, 1971; Mayhew & Anstis, 1972).[5] A similarity apparent from the data of the first two experiments was the relatively small magnitude of the contingent effect compared with simple (i.e., single variable) adaptation effects (cf. McCollough, 1965). A further test was conducted to determine whether the present effect exhibited an unusually long time course of recovery, similar to that observed with the contingent aftereffects in vision which in some cases last for as long as 6 weeks (cf. Mayhew & Anstis, 1972). The experiment involved three listeners (A.C., J.I., and J.K.), each of whom had demonstrated relatively large boundary shifts after adaptation in the main experiment. The contingent adaptation effect on both the [ba]-[p^{h}a] and [bi]-[p^{h}i] series was replicated for each of the three Ss (mean boundary shift = 4.7 msec VOT). The listeners were presented another baseline test 5 days after adaptation. For each S, approximately complete recovery from adaptation was obtained. The contingent adaptation effect for speech does not, then, appear to exhibit an extremely long time course, unlike some contingent aftereffects in vision.

To summarize, the results of the main experiment indicated that adaptation to an alternating sequence of [da] and [t^{h}i] produced opposing shifts in the perception of two CV test series varying in VOT. After adaptation, fewer [b] responses were made to stimuli of the [ba]-[p^{h}a] series, whereas more [b] responses were made to stimuli of the [bi]-[p^{h}i] series.

The presence of this contingent adaptation effect, as well as the complementary effect obtained with the [di]-[t^{h}a] adapting sequence, indicates that the analysi of the feature *voicing* in consonants is carried out, a least in part, by detection channels that are vowel-dependent. While the present evidence supports this interpretation, other evidence from our laboratory would appear to support the existence of vowel-independent channels for consonant feature analysis as well (cf. Ades[6]; Cooper, in press). Using the selective adaptation technique, both Ades and I have found that partial adaptation can be produced for consonants varying in the feature *place of articulation* when the adapting syllable contains a different vowel from the vowel of the test series. Such data may well indicate the presence of vowel-independent channels; another way to account for these data, however, would be to postulate the existence of vowel-dependent channels only (at the stage of analysis adapted here), each channel responding to a consonant feature best in a particular vowel environment but responding in other vowel regions as well. Tuning in this case would need to be broad enough so that adaptation to C_1V_1 would adapt channels that also detect, to a lesser extent, the features of C_1V_2. Whether such vowel-dependent channels exist can be studied further with the contingent adaptation technique.

REFERENCES

Cooper, W. E. Adaptation of phonetic feature analyzers for place of articulation. Journal of the Acoustical Society of America, in press.

Cooper, W. E., & Blumstein, S. E. A 'labial' feature analyzer in speech perception. Perception & Psychophysics, 1974, 15, 591-600.

Eimas, P. D., Cooper, W. E., & Corbit, J. D. Some properties of linguistic feature detectors. Perception & Psychophysics, 1973, 13, 247-252.

Eimas, P. D., & Corbit, J. D. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 1973, 4, 99-109.

Fant, C. G. M. On the predictability of formant levels and spectrum envelopes from formant frequencies. In *For Roman Jakobson*. The Hague: Mouton, 1956. Pp. 109-120.

Fidell, L. S. Orientation specificity in chromatic adaptation of human "edge-detectors." Perception & Psychophysics, 1970, 8, 235-236.

Haggard, M., Ambler, S., & Callow, M. Pitch as a voicing cue. Journal of the Acoustical Society of America, 1970, 47, 613-617.

Harris, C. S., & Gibson, A. R. Is orientation-specific color adaptation in human vision due to edge-detectors, after-images, or "dipoles"? Science, 1968, 162, 1506-1507.

Held, R., & Shattuck, S. R. Color- and edge-sensitive channels in the human visual system: Tuning for orientation. Science, 1971, 174, 314-316.

Klatt, D. H. Voice-onset time, frication and aspiration in word-initial consonant clusters. Quarterly Progress Report of the M.I.T. Research Laboratory of Electronics, 1973, 109, 124-136.

Lisker, L., & Abramson, A. S. A cross-language study of voicing in initial stops: Acoustic measurements. Word, 1964, 20, 384-422.

Lisker, L., & Abramson, A. S. Some effecs of context on voice onset time in English stops. Language & Speech, 1967, 10, 1-28.

Lisker, L., & Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967*. Prague: Academia, 1970. Pp. 563-567.

Mayhew, J. E. W., & Anstis, S. M. Movement aftereffects contingent on color, intensity, and pattern. Perception & Psychophysics, 1972, 12, 77-85.

McCollough, C. Color adaptation of edge-detectors in the human visual system. Science, 1965, 149, 1115-1116.

Stevens, K. N., & Klatt, D. H. The role of formant transitions in the voiced-voiceless distinction for stops. M.I.T. Research Laboratory of Electronics Quarterly Progress Report, 1971, 101, 188-196.

NOTES

[1] VOT is not the only cue relevant to voicing distinctions among word-initial stop consonants (cf. Haggard, Ambler, & Callow, 1970; Stevens & Klatt, 1971). One additional cue to such distinctions, namely the presence or absence of significant formant transitions after the onset of voicing (Stevens & Klatt, 1971) has recently been shown to be effective in producing selective adaptation effects along the [ba]-[p^{h}a] test series used in the present experiments (Cooper, W. E. Selective adaptation for acoustic cues of voicing in initial stops. Forthcoming). Given this finding, it is probably better to view the present adaptation effects as operating on an analyzer that extracts phonetic information of *voicing* rather than VOT information per se.

[2] Klatt, D. H. An acoustical theory of terminal analog speech synthesis and a control strategy for the replication of a natural utterance. Forthcoming.

[3] Two-tailed t test of significance for correlated observations. All subsequent tests of significance were one-tailed t tests for correlated observations.

[4] As in another study of speech adaptation (Cooper & Blumstein, 1974), the failure to obtain shifts in the expected direction for some Ss was highly correlated with a lack of monotonicity in the identification functions obtained in the unadapted state; this was true for Ss S.G. and B.H.

[5] The search for formal similarities between these effects is carried out in the spirit of obtaining more information about speech adaptation. I do not subscribe to the view that a demonstration of similar adaptation effects in vision and speech necessarily indicates the presence of any strong similarity between the mechanisms underlying such effects.

[6] Ades, A. E. A study of acoustic invariance by selective adaptation. Forthcoming.

(Received for publication December 3, 1973;
revision received March 14, 1974.)