Perception of the English /s/–/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices[a)]

D. H. Whalen
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

774

(Received 10 October 1990; accepted for publication 12 June 1991)

A series of experiments compared two approaches to fricative identification, spectral template matching and articulatory dynamics. Natural-speech /s/ and /ʃ/ noises from fricative-vowel or vowel-fricative syllables were cross spliced so that "hybrid" noises started out as either /s/ or /ʃ/ and ended up with the other fricative noise in varying proportions. With both initial and final fricatives, listener judgments most often agreed with the longer part of the noise even when spectral templates would predict the other category. Also, the vocalic formant transitions contributed to the judgment. In another experiment, open transcriptions by four expert listeners similarly showed that all the cues were used; there were also some instances of nonspeech percepts that would be predicted by gestural models. One further experiment had subjects identify two fricatives from hybrid noises between two vocalic segments. When the order of the noises differed from the order of the transitions, the perceived ordering of the fricatives was often the reverse of the order of the noise segments. Taken together with previous results, these experiments indicate that listeners take the whole fricative noise, as well as the transitions, into account in fricative identification.

PACS numbers: 43.71.An, 43.71.Es

INTRODUCTION

The multiplicity of cues to speech distinctions continues to be the central problem in speech perception. As speech scientists, we are struck by the extreme difficulty of describing all the relevant aspects of the speech signal. There are at least 16 different acoustic parameters that can be used by perceivers in making the voicing judgment on intervocalic stops (Lisker, 1986), and similar counts could be made for other distinctions. The slow progress in automatic speech recognition can be attributed in large part to the variability in the speech signal (see Vaissiere, 1985, for a review). As language users, though, we are struck by the apparent ease with which we use this multitude of cues. When confronted with a signal rich in cues, or even an impoverished synthetic signal, we do not consciously generate a set of possible strategies for dealing with the acoustics, we simply perceive speech. Listeners can be trained to become more sensitive to fine distinctions (e.g., Pisoni and Lazarus, 1974; Samuel, 1977; Edman, 1979), a result that is usually claimed to argue for the perceptual use of nonspeech aspects of the signal. However, it can also be argued that it is the sensitivity to the speech aspects of the signal that is being improved. Additionally, there are some individual differences in how much weight is given to a particular cue (Best *et al.*, 1981), and some cues do not affect a subject's identifications until after a certain amount of familiarization has taken place (Whalen *et al.*, 1990). Nonetheless, in our day-to-day experience, we do not knowingly grapple with acoustic variability, we simply hear language. These two observations, of ease for humans and difficulties for machines, must be reconciled, and two main ways of doing so have been proposed in the literature.

The first approach to a resolution is to deemphasize the inherent variability in the speech signal by extracting just those acoustic properties that are believed (or proposed) to be necessary to perception. This is the goal of the spectral template matching theory of Stevens and his colleagues (Stevens, 1980, 1985; Stevens and Blumstein, 1978; Blumstein and Stevens, 1979; Stevens *et al.*, 1986; Stevens and Keyser, 1989). According to the spectral template theory, the values of phonological features are claimed to be extractable from single spectral slices or pairs of slices located at regions of large acoustic discontinuities. Over the many years of work in this framework, the original template model has been modified. One major modification in the more recent work has been that two spectral slices are taken rather than one; the two are taken across an acoustic boundary and then compared. For example, the place of articulation for nasal consonants requires comparison between a slice in the murmur with one in the vocalic section (Kurowski and Blumstein, 1987). Similarly, the dental-alveolar distinction is said to depend on the relationship of the burst amplitude to the vowel amplitude (Jongman *et al.*, 1985), and the place of articulation in stops on more than one temporal slice (Kewley-Port, 1983; Lahiri *et al.*, 1984). This shift to comparison across slices was made partly in response to the findings of Blumstein *et al.* (1982) and Walley and Carrell (1983) who found that listeners do not use the spectral shape at onset as the primary cue. A metric for the detection of voicing in fricatives that involves a comparison of slices in the vowel

[a)] Part of this research was presented at the Fall meeting of the Acoustical Society of America, San Diego, CA, November, 1983 [D. H. Whalen, J. Acoust. Soc. Am. Suppl. 1 **74**, S90 (1983)], and at the University of Michigan, January, 1991.

and in the noise has been proposed (Stevens *et al.*, 1987). And in the case of the fricatives /s/ and /ʃ/, the reliance on a single slice in the noise (Stevens, 1980, p. 840) has been replaced with one that involves a comparison between the noise and the $F3$ of the vowel (Stevens, 1985, p. 247). This distinction between /s/ and /ʃ/ will be tested in the present experiments.

A second major modification in the spectral template framework is the extraction of phonetic features rather than phonetic segments. In the earlier versions of the theory (such as Blumstein and Stevens, 1979), there were three templates for the three stop places of articulation in English. Those three places represent the intersection of two binary features in the system of Jakobson *et al.* (1963), and so the templates were not fully consistent with the phonological assumption that features are the objects of perception. Recent work by Blumstein (1986) has examined the possibility of creating templates for features rather than places, but these new feature templates have not yet been tested. Most of the recent work mentioned in the previous paragraph has been feature oriented by virtue of the focus of each study: Each has examined only pairs of segments distinguished by one feature, rather than a series of segments that would be distinguished by a set of features, as would be the case with stop place of articulation. We therefore do not know whether the new templates will attain the 85% level of success that the earlier ones had with initial stops (Blumstein and Stevens, 1979).

Spectral template theory can be contrasted with a second approach to reconciling the ease of human perception and the difficulty of acoustic analysis. The second approach focuses on the identification of the articulatory gestures that underlie speech sound production. The approach is exemplified by perceptual vector analysis (Fowler, 1984; Fowler and Smith, 1986) and the motor theory (Liberman and Mattingly 1985). A similar framework is implicit in the articulatory phonology of Browman and Goldstein (1985, 1989, 1990). The theory holds that a physical system that might look complex from one perspective (e.g., from an analysis of the acoustic events *per se*) can be seen as simple when certain underlying principles are exposed. On this view, the articulatory gestures, which overlap in time, are specified by their acoustic consequences without any intervening level of analysis. The success of such theories depends on the success of the algorithms that model how the listener attributes the acoustic information to a dynamical system of gestures.

The spectral template theory and the gestural approach differ in the way they conceptualize the perceptual process, but there are similarities as well. Although the spectral templates in Stevens' theory are claimed to be the primary cues in perception (Stevens and Blumstein, 1978, p. 1367), secondary cues are also argued to play a role. For fricatives, these cues look much more like the formant transitions and noise frequencies that have been examined in the traditional phonetics literature (Stevens, 1985). Since the continuously varying acoustic signal is the source for the extracted gestures in a gestural account, there is bound to be much in common between gestural acoustics and the composite of primary and secondary cues of the spectral template account. Current descriptions of the acoustic cues are impoverished in any account: For various contrasts in restricted contexts, we can tell which acoustic changes will make a difference, but there is no metric that can put all those differences into a single analysis.

Still, the two theories do make different predictions about stimuli containing primary spectral cues that, according to the spectral template account, should be unambiguous. A spectral template model would predict that primary cues should be sufficient to determine the percept. An articulatory account of the same stimuli would have to take the entire signal into account. That being the case, the predictions of the gestural account would be less clear, but the resulting percept would have to be gesturally coherent, and this would imply that portions of the signal might be treated as nonspeech.

The present experiments set up a situation, using cross-spliced natural tokens of the English fricatives /s/ and /ʃ/, in which the primary cues (as specified by the spectral template) are unambiguous but potentially not the determining ones. Acoustical measurements of the noise spectra have consistently shown large differences in the frequency of the fricative noises (e.g., Hughes and Halle, 1956; Strevens, 1960), and natural /s/ and /ʃ/ noises override cues in the vocalic formant transitions for artificially cross-spliced syllables (Harris, 1958). However, these transitions, along with vowel quality, affect listener judgments (Mann and Repp, 1980; Whalen, 1981) as does speaker sex (May, 1976). The fact that some vowel information is also perceivable in both initial (Yeni-Komshian and Soli, 1981) and final (Whalen, 1983) noises is irrelevant from a feature-extracting point of view, since those effects are of the sort that templates were designed to obviate. As for the transitions, a more sensitive response measure (reaction time) reveals that they affect the perceptual process even when they do not seem to contribute to phonetic decisions (Whalen 1984; Whalen and Samuel 1985). This fact, though, could be attributed to the artificial task demands, which might emphasize secondary cues that would be ignored in running speech. The present experiments will avoid those demands for a speeded response, and yet create stimuli in which the contribution of the noise and the transition will be different from a proposed spectral template. That is, if the stimuli contain the primary cues, then there should be no effect of secondary cues on the identification of the fricative segment. According to articulatory dynamics models, the effect will depend on the difficulty the perceptual system has in determining what kind of gesture might have created such a signal. In the extreme case, the perceptual system might reject part of the speech signal altogether as being nonspeech.

I. EXPERIMENT 1

In the first experiment, the initial fricative noises from the English words "suit" and "shoot" were combined in regularly varied proportions, starting as /s/ and ending as /ʃ/ or vice versa. In this way, large changes in the spectrum were inserted into a natural amplitude contour. These hybrid noises were followed by the vocalic segments of either sylla-

ble ("[s]uit" and "[sh]oot"). Thus the transitions supported the phonetic category of the initial part of the fricative noise in half the stimuli, but the final (contiguous) part in the other half.

The acoustic discontinuities should give rise to phonetic judgments, on the spectral template account. However, all the stimuli will still have a long enough fricative noise adjacent to the vowel for a comparison across the noise/vocalic segment boundary. Thus, if only one noise is to be reported, it should be the adjacent one, since it would be the one at the fricative/vowel boundary. The transitions should not affect the judgment since the template will give an unambiguous value. Alternatively, the gestural account would predict that the transitions would influence the decisions, since the perceptual system is assumed to use as much of the signal as possible.

A. Method

A male native speaker of American English recorded three tokens of the words "suit" and "shoot." They were low-pass filtered at 10 kHz and digitized via the Haskins PCM system (Whalen *et al.*, 1990) at 20 kHz with high-frequency pre-emphasis. One token of each word was chosen so that the duration of the fricative noise was the same in each word (200 ms). The vocalic segments began with the onset of voicing.

The fricative noise continuum was constructed by replacing the first part of one fricative noise with an equivalent amount of the other noise. Thus the duration of each hybrid noise remained 200 ms. There were nine duration values (from 0 to 200 ms of the initial noise in steps of 25 ms), with either /s/ or /ʃ/ coming first, giving 18 hybrid noises. (Note that the edited noise with 0 ms of one noise at the beginning was in fact the same as the one with 200 ms of the other noise at the beginning. Each version was nonetheless treated as a separate stimulus.) Cuts were made without regard to zero crossings, since this simple method of making the continuum resulted in no artifacts detectable to the experimenter. Each noise was adjoined to each of the vocalic segments, which had transitions appropriate to either /s/ or /ʃ/. The transitions in the vocalic segment thus supported the fricative category of either the initial or the final noise portion. This combination gave the final total of 36 stimuli.[1]

Each stimulus was repeated five times, and the resulting 180 tokens were randomized. (Given the large number of tokens, the number of repetitions was kept at a minimum.) These were presented with an interstimulus-interval of 2.5 s. Subjects were asked to identify the fricative as "s" or "sh" and to write that response down. The procedure was forced choice: If the subjects were unsure, they were told to guess.

The subjects were 32 young adults with no reported hearing problems who were paid for their participation. (The number of subjects was determined by the constraints of a separate experiment which was given in the same session.)

B. Results

Figure 1 presents the percentage of identifications that agreed with the initial portion of the noise. The results are presented in this way so that the differences in cross-over points for the two fricatives can be easily seen. The functions marked with circles represent increasing "s" responses, while the squares indicate increasing "sh" responses. The end-point stimuli (all /s/ and all /ʃ/) confirm that unambiguous /s/ and /ʃ/ noises override the transition information. Also of note is that all the cross-over points are beyond the point where equal portions (100 ms) of the two noises were present, indicating that more weight was given to the final portion of the noise. The initial portion may have been less salient because of its slow rise time, which resulted in a lower amplitude noise at the beginning.

By comparing the solid lines with the dotted lines for each fricative noise, we can see that, when the transitions support the category of the initial noise, less duration of that noise is needed for the subject to report that fricative. That is, the cross-over point for the solid line is to the left of that for the dotted line with each of the two fricative noises. An analysis of variance, based on the individually calculated PROBIT cross-over points, shows the transition effect to be highly reliable [$F(1,31) = 72.57, p < 0.001$].

There is also a difference between the two fricatives. While the cross-over points were just past the half-way point when /s/ was the initial noise, the crossover was much later for /ʃ/-initial stimuli [$F(1,31) = 60.77, p < 0.001$]. In fact, when /s/ transitions were present, a mere 25 ms of /s/ noise at the end of the frication was sufficient to give /s/ percepts half the time as can be seen at the square with the dotted line at 175 ms in Fig. 1.

Finally, the analysis of variance revealed an interaction between the effect of the transitions on the cross-over points and the difference between the two fricatives [$F(1,31) = 9.08, p < 0.01$]. The influence of the transition is smaller with the /ʃ/-initial stimuli. A separate analysis shows that the effect is still significant, however [$F(1,31) = 16.97, p < 0.001$]. The interaction could be due to properties of either the noise or the transitions, and the present results do not distinguish between these two possibilities.

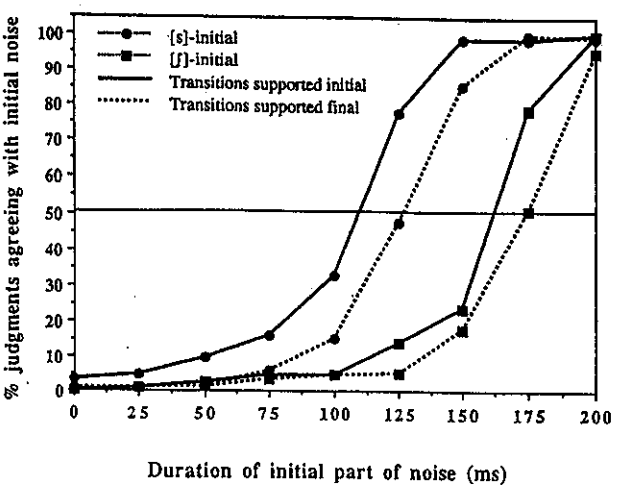

FIG. 1. Fricative judgments for initial fricative noises in initial position, experiment 1.

C. Discussion

Cues from both parts of the noise and from the transitions were used by listeners in making their category judgments, a result at variance with the template matching model. The judgment was not determined by the first slice (the first 10 to 30 ms of the noise), nor by the first slice of the noise after a major acoustic change, nor by the slice adjacent to the vocalic segment, nor by the difference in amplitude in the $F3$ region at the release of the consonant. Although by Stevens' (1985) template all of the stimuli should have been heard as /s/, subjects reliably heard /ʃ/s as well. Even if we assume that a modified template would correctly categorize these stimuli, the portion of the acoustic signal at the boundary between the noise and the transition did not determine the category. Instead, the reported fricative tended to be consistent with the longer of the two noise portions, though with the later part having more weight, and with some influence of the transitions. The added weight of the later part may have been due its temporal position, to the greater amplitude of the second half of the fricative noise, or to an increase in importance for the portion of the signal contiguous with the transitions.

The effect of the transitions may indicate that fricative gestures that are imputed to both the noise and the vocalic segment are better supported than ones that treat the acoustic segments separately. There are two ways of describing the effect of transitions on the identification of the hybrid noises. (1) The fricative supported by the initial part of the noise was chosen when its duration was shorter if the transitions supported that percept. (2) By the same token, shorter *final* portions decided the judgment if the transitions supported the category of the final fricative. This leads us to two, partially consistent explanations. On the one hand, transitions could have cue value even if they are not contiguous with the appropriate fricative (as in the first description of the effect above). On the other hand, it could be that transitions are ineffective unless they are contiguous with the appropriate noise (as in the second description). Either description could be given in gestural terms, since effects across intervening segments have been found (e.g., Fowler, 1981; Whalen, 1990).

Finally, there was a difference in the amount of noise needed at the beginning depending on which fricative noise came first: /s/ noise began determining the judgment just after the mid-way point, while more /ʃ/ noise was needed for an /ʃ/ judgment. It seems that since the following vowel was /u/, listeners were willing to tolerate a great deal of low-frequency noise as attributable to the upcoming vowel's resonances (cf. Soli 1981). This resonance happens to be at the lower edge of the frequency of the fricative pole for /ʃ/. Thus listeners may have been led to attribute some of the /ʃ/ noise to strong vowel coarticulation.

II. EXPERIMENT 2

The second experiment examines the behavior of discontinuities introduced into final fricatives. While the template model does not explicitly discuss syllable-final fricatives, the predicted acoustic difference was in fact found. The stop templates have been applied to final stops as well as initial ones (Blumstein and Stevens, 1979), so it was assumed that the criteria for making the /s/–/ʃ/ distinction would apply to final fricatives. The words chosen were "mess" and "mesh." Discontinuities were introduced in the same manner as in the first experiment.

In experiment 1, the latter part of the fricative noise was given more weight in deciding the fricative category. This may have been due to the greater amplitude of the last half. However, the latter part was also contiguous with the vocalic segment, which could equally well account for its dominance. If the kinds of dynamic gestures that could account for most of the acoustic signal have an easier time attributing coherence to contiguous segments, then the *first* part of the fricative noise should have a greater effect in the final fricatives. Thus experiment 2 was designed to see if similar results would be obtained when the hybrid noises were put into syllable-final position, in which case the initial portion of the noise would be contiguous with the vocalic formant transitions.

A. Method

The same talker as in experiment 1 recorded tokens of "mess" and "mesh," which were digitized as before. The fricative noises were removed, leaving vocalic segments (of equal duration) with transitions appropriate to /s/ or to /ʃ/. These then had the 18 hybrid noises of experiment 1 added at the end, again resulting in 36 stimuli. There was no peak in the /s/ noise that was closer than 600 Hz to the vowel's $F3$ (possibly because the noises were not, in fact, produced in this environment), but that one was 28 dB down from the vowel's $F3$. The closest peak in the /ʃ/ noise was at the same intensity as the vowel's $F3$. For the mismatched transitions, both noises were in the /s/ category, with the /s/ noise being 22 dB down from the /ʃ/ transitions' $F3$, and the /ʃ/ noise being 6 dB from the /s/ transition's $F3$. Thus, like the original initial position, the match to Stevens' template was good for /s/. In this case, though, the /ʃ/ noise was in the ambiguous region when it occurred with the /ʃ/ transitions. Additionally, since the noises had been produced in initial position, they sounded somewhat less natural in final position. Nonetheless, the syllables were coherent and easy to perceive as "mess" or "mesh."

Subjects were asked to identify one fricative, and to write down "s" or "sh," guessing if necessary. Five repetitions of each stimulus were randomized, with an interstimulus interval of 2.5 s.

The subjects were 21 young adults with no reported hearing problems who were paid for their participation. (The number of subjects was again determined by the constraints of a separate experiment which was given in the same session.)

B. Results

Figure 2 displays the results in a plot similar to Fig. 1. The cross-over points fall closer to the midway point than in experiment 1. As in the first experiment, the transitions had an effect, as seen in the displacement of the dashed lines

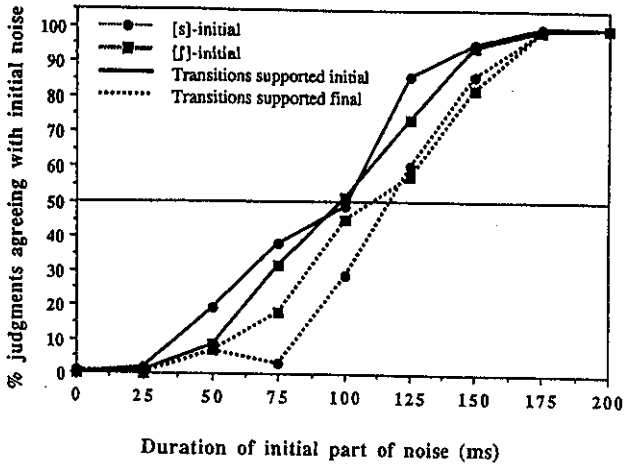

FIG. 2. Fricative judgments for initial fricative noises in final position, experiment 2.

relative to the solid lines. An analysis of variance, based on individually computed cross-over points, indicated that this effect was significant [$F(1,20) = 43.18, p < 0.001$].

The two fricatives in this instance behaved similarly [$F(1,20) < 1$, n.s.]. There was an interaction between the fricative category and the transition effect [$F(1,20) = 17.40, p < 0.001$]; the transition effect was larger for /s/-initial noises, as in experiment 1. A separate analysis of just the /ʃ/-initial noises, however, showed that those transitions did influence the judgment of the fricative [$F(1,20) = 15.68, p < 0.001$].

C. Discussion

The fricatives in syllable-final position were perceived in relationship to both the noise and the transitions, rather than the spectral template. In this set of stimuli, the /s/ noises met Stevens' (1985) template for /s/, while the /ʃ/ noise was either at a level that would be classified as ambiguous (with /ʃ/ transitions) or in the /s/ region (with /s/ transitions). That being the case, the template model would predict that the /s/ noises should have been determined by the noise at the boundary while the /ʃ/ noises should have been determined by secondary cues. The only secondary cue mentioned specifically is the set of formant transitions in the vocalic segment (Stevens, 1985). Here, though, as before, it took some longer stretch of the noise before a fricative judgment was consistent with that noise. Displacing the fricative noises from "suit" and "shoot" to final position did reduce the importance of the later portion of the noise. Since the cross-over points are earlier than in experiment 1, there is some gain in importance for the noise portion contiguous with the vocalic segment. In final as well as initial position, the fricative judgment is a decision based on noise duration, noise position, and the cue value of the transitions. In gestural terms, it is the fricative gesture that receives the most support that determines the percept.

The asymmetry between the two fricatives in the first experiment was attributed to the effects of the resonance for the vowel /u/. The fact that these two fricative noises have the same cross-over points when in the context of the vowel /ɛ/ supports this interpretation based on acoustic analysis. When the vowel does not result in a confusable vowel resonance within the noise, the two orderings of the fricative noise pieces have the same cross-over points. However, some effect of syllabic position may be at work as well.

III. EXPERIMENT 3

The noises in experiment 2 were in an unnatural position—they were produced syllable initially and spliced to be syllable final. To see if noises produced in syllable-final position give similar results, and to make sure that the results were not dependent on the one pair of noises used so far, a replication of experiment 2 was conducted. This replication used the original fricative noises from the tokens of "mess" and "mesh" used in experiment 2. These stimuli more fully met Stevens' template for the /s/–/ʃ/ distinction.

A. Method

The tokens of "mess" and "mesh" used in experiment 2 had been chosen so that the fricative noises were of equal duration, in this instance, 280 ms. These were edited as before, to vary in eight steps. Thus the step size was longer than in experiment 1 (35 ms rather than 25 ms). The amplitude of the noise peak near the frequency of the vowel's $F3$ relative to the vowel's $F3$ amplitude was + 5 dB for /ʃ/ and − 9 dB for /s/, which satisfies the acoustic criteria in Stevens (1985). With mismatched transitions, both are potentially ambiguous, as − 1 dB for /ʃ/ and − 3 for /s/. Each of the hybrid noises was combined with each of the vocalic segments, and five repetitions of each were presented in random order for identification of the fricative as "s" or "sh."

The subjects of experiment 2 participated in experiment 3 in the same session.

B. Results

Figure 3 shows the results plotted as in Figs. 1 and 2. In this experiment, the cross-over points were somewhat before

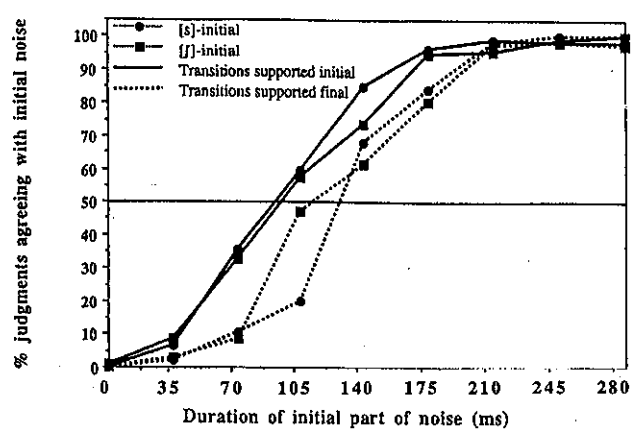

FIG. 3. Fricative judgments for final fricative noises in final position, experiment 3.

the halfway point of the continuum. The amplitude contour of the final noises was less asymmetrical than the initial noises used in the first two experiments, so it could be that there was less of a reason for subjects to make their judgments based on the second half of the noise. However, the members of this continuum were of longer duration than those of the first continuum. If the cross-over points are computed as absolute number of milliseconds of noise at which the judgment shifts, rather than as a place on the continuum, the present results are statistically indistinguishable from those of experiment 2 (in an analysis of variance for the two experiments combined [$F(1,20) < 1.0$, n.s.]).

The transitions again had an effect on the cross-over points [$F(1,20) = 75.92$, $p < 0.001$]. There is a nonsignificant tendency [$F(1,20) = 1.61$, n.s.] for the /s/ transitions to have less of an effect, as in experiment 2. There is no difference between /s/- and /ʃ/-initial noises [$F(1,20) < 1.0$, n.s.].

C. Discussion

The stimuli of experiment 3 more fully met Stevens' (1985) spectral template, and yet the results are quite similar to those of experiment 2. That is, the fricative judgment did not depend just on the spectral slices on either side of the boundary. The secondary cues, which would be predicted to be decisive when they did not match the noise they were adjacent to, had their more usual effect, shifting the boundaries but not being the sole determinant of the fricative category.

The results are, for the most part, in accord with those of experiment 2, but the asymmetry between the two fricative noises disappeared with these final noises. These are the ones that had been produced in the context they were presented in, and that in itself might account for the difference. However, the amplitude contours for these two fricative noises were more symmetrical than those for the initial noises. This fact could change both the amount of perceptual weight given to the noises and, possibly, the imputed dynamics of the final fricative, since syllable final gestures have been found differ from initial ones (Lehiste, 1960, p. 42; Krakow, 1989).

IV. EXPERIMENT 4

In the previous experiments, subjects were limited to giving a single fricative response. Since there was an additional point at which the spectrum changed radically, there might have been additional percepts which were not detected in the first experiments. Fricative clusters, for example, might have been heard, but the constraints imposed by the response set would not have allowed them to appear. Other speech categories, and even nonspeech percepts, might have been heard but not reported in the previous experiments. Therefore, the stimuli of experiment 1 were used with expert listeners, who were asked to describe all the sounds they heard with the various stimuli. An open set of speech sounds and the chance to use descriptions of nonspeech percepts were available.

A. Method

The stimuli of experiment 1 were used. Since the number of judgments to be made was rather large, only a subset of the transitions were used. Along the noise continua, the transitions were alternated, so that if the 0-ms stimulus had /ʃ/ transitions, the 25-ms stimulus would have /s/, the 50, /ʃ/, and so on. Thus each subject heard only one set of transitions for each of the hybrid noises. The subjects were unaware of the stimulus manipulation involved.

The stimuli were presented in random order, but each was presented to the subject as often desired, until the final judgment was recorded. They were to report the phonetic structure of the prevocalic consonant(s), along with an informal description of any concurrent nonspeech noises they might hear.

The subjects were four colleagues from Haskins Laboratories who had at least 5 years of experience in acoustic phonetics.

B. Results

Phonetic and nonphonetic judgments are given in Table I. This collapses across transition type, since there were too few judgments to draw firm conclusions about their influence.

As can be seen, there were many cases in which two fricatives were heard. Additionally, there were many cases in which a brief (25 or 50 ms) segment was perceived as a stop at the onset of the noise. The place of this was usually correlated with the noise (/t/ for /s/ and /k/ for /ʃ/). However, at durations of 75–150 ms, the noises were more often heard as fricatives. There were only six instances in which a nonspeech percept was obtained, and only three instances in which another fricative (/x/ or /θ/) was heard.

C. Discussion

The detailed judgments by the expert listeners show that there was a tendency for more than one fricative judgment to be possible for the hybrid stimuli. Even for the noises with equal portions of each fricative, though, half of the judgments contained only one linguistic category. Thus the subjects in the earlier experiments were not too severely constrained by being limited to a simple response.

More important, the character of the noise envelope constrained the judgments. Only when the spectral change occurred in a very small time frame (25 ms) were there many stop percepts. At longer durations, either a simple or complex fricative was heard. This is in spite of the fact that there were equally many auditory discontinuities no matter where the change took place. Thus, for spectral template models, there appears to be a missing piece. Some sort of timekeeper seems to be necessary to interpret the place features that are extracted from the noise. If there is no such function, then the change from stop to fricative percepts as the duration increases is without explanation.

For gestural models, there seems to be the possibility of making phonetic sense of some acoustic signals that have changes too abrupt for a real vocal tract to make. However,

TABLE I. Results for experiment 4, judgments by four expert listeners to the stimuli from experiment 1. When two percepts are separated by a hyphen, it indicates that one category (or other sound) was perceived as preceding the other. When a plus sign occurs, it indicates that the two percepts were simultaneous. A question mark indicates uncertainty about the presence and/or category of a sound.

/s/-initial:	Number of milliseconds of initial noise								
	0	25	50	75	100	125	150	175	200
Subject:									
AML	ʃ	t? − ʃ	ʃ	ʃ?	noise + ʃ	noise + ʃ	p? − s	s	s
SYM	ʃ	t? − ʃ	p − ʃ	č	t − sy	sy	sy	sy	s
SDS	ʃ	t? − ʃ	s − ʃ	s − ʃ	s − ʃ	s + ʃ	s + ʃ	s	s
PSB	ʃ	ʃ	x? − ʃ	ʃ	s	s	s	s	s
/ʃ/-initial:	0	25	50	75	100	125	150	175	200
Subject:									
AML	s	s	p? − s	k? − s	s?	noise + s	ʃ? − s	noise + s	ʃ
SYM	s	noise + s	noise + s	s − θ	s − θ	ʃ	ʃ	ʃ	ʃ
SDS	s	s	s	ʃ − s	ʃ − s	ʃ − s	ʃ − s	ʃ − s	ʃ
PSB	s	k? − s	k? − s	s	s	s	ʃ − s	ʃ	ʃ

for two subjects, there were instances where part of the signal was rejected as being nonspeech. Thus there is some indication that the limits to perceptual plasticity can be reached in these stimuli. We cannot tell whether the stimuli that did not elicit nonspeech percepts were heard as having an unusual articulation, were erroneously heard as having a stable fricative noise spectrum, or simply with nonspeech percepts that were too faint to be reported. We do know that they were able, in many instances, to use the longer of the two noises to completely determine the consonant percept.

V. EXPERIMENT 5

The final experiment was designed to show that the two sets of vocalic formant transitions, which have already been shown to be effective secondary fricative cues, can provide fricative information even when not contiguous with the appropriate noise. If they do so, then they will provide information for noises that do not require secondary cues, and they will do so for spectral slices that are discontinuous to them.

To show such an effect, the hybrid noises were to be placed between two vocalic segments. A pretest showed that noises consisting of equal portions of the two fricatives easily give rise to two different fricative judgments when the hybrid noise occurs between two vocalic segments. In experiment 5, the ordering of the fricative noises is put in conflict with the order of the formant transitions for half the trials. If the transitions have an effect only on the noise portion they are contiguous with, then the ordering of the transitions should have no effect on the identifications. If the transitions support the category even of noncontiguous noises, then the order reported will occasionally be determined by the order of the transitions.

A. Method

The edited noises consisting of equal portions of the two initial fricative noises were used. These two noises (one with /s/ first, the other with /ʃ/ first) were placed between the two vocalic segments. The transition information in the initial vocalic segment always supported the category opposite to that of the transitions in the final vocalic segment. If the "me[ss]" vocalic segment preceded the noise, the "[sh]oot" vocalic segment followed, and if the "me[sh]" vocalic segment preceded the noise, the "[s]uit" vocalic segment followed. In this way, the transitions either supported the order of the fricatives as specified by the order of the noise portions, or they were opposed. These manipulations resulted in four stimuli. The noises were either in the order /s/–/ʃ/ or the reverse, and the transitions were either in the same order as the noises or the reverse.

Ten repetitions were randomized for presentation with a 2.5-s interstimulus interval. The response set was limited to "mess shoot" and "mesh suit," written as "s-sh" or "sh-s." No other combinations were allowed.

Twelve young adults with no reported hearing problems were paid to participate.

B. Results

The results are presented in Table II. The percentage of responses in which the order of the fricatives was the reverse of the order of the fricative noises was calculated. These responses will simply be called "reversals" in this discussion. There were reversals in response to all stimuli, though there were significantly more when the transitions were in the reverse order (as revealed by an analysis of variance

TABLE II. Percentage of fricative judgments reversed from the order of the fricative noises (experiment 5).

Noise:	Transitions supported order of noises	Transitions contradicted order of noises
/s/-initial:	6.3	41.7
/ʃ/-initial:	7.9	6.7

[$F(1,11) = 18.25$, $p < 0.01$]). The two fricative noises behaved differently, as can be seen both in the main effect of fricative noise [$F(1,11) = 7.22$, $p < 0.05$], and in the interaction of the transition and fricative effects [$F(1,11) = 13.03$, $p < 0.01$]. As is obvious from the table, the majority of reversals came when the /s/ noise was first and the transitions in the final vocalic segment were appropriate to /s/. The three remaining values are statistically indistinguishable.

C. Discussion

The results of experiment 5 indicate that, at least in some cases, vocalic formant transitions provide cue value for the fricative that they were produced with even when they are adjacent to a fricative noise from a different fricative. If the transitions contributed to the percept only when contiguous to an appropriate noise, then the number of reversals reported would have been the same in all conditions. Instead, there was a significant tendency to allow the transitions to override the order information in the noise when they were in conflict.

The reversed transitions were not as successful in overriding the fricative noise order when the /ʃ/ noise came first. One result from experiment 2 helps explain this fact. The stimulus with /ʃ/ noise first had /s/ transitions in the first vocalic segment. Those are just the transitions that were not as effective in shifting the boundary between fricatives in experiment 2. The transitions in "me[ss]" were not as effective a cue when followed by the edited noises from "suit" and "shoot." Whether this due to the unnaturally high onset amplitude of this displaced fricative noise or some other phonetic interaction, it does provide a basis for explaining the asymmetry in the effectiveness of the reversed transitions.

VI. GENERAL DISCUSSION

When a hybrid fricative noise, one made from portions of two different fricative categories, provides a clear instance of a spectral discontinuity, it is not always the case that two fricatives are perceived. Instead, the limitations imposed by the single amplitude contour seem to be just as salient perceptually. Additionally, the formant transitions can give information independently of the fricative noise, as shown by the significant number of reversals of fricative percepts when the order of two noise portions conflicts with the order of the transitions. Thus even when natural spectral shapes (albeit in manipulated positions) are put in positions where brief spectral intervals should, in template matching accounts, provide phonetic information, they often do not. Instead, a more complicated picture emerges, one treating the two portions of hybrid fricative noises via incorporation into one percept, rejection of some of the sound as nonspeech, or generation of multiple percepts.

The present results also pose difficulties for theories of perception which emphasize incorporation of all available information (e.g., Fowler, 1984; Whalen, 1984; Fowler and Smith, 1986; Whalen, 1989). The variability and individual differences in the judgments make for complicated predictions. Thus, for example, it is not unreasonable that a brief portion of noise, one which changes spectrum so rapidly that no vocal tract could have produced it, would be heard as a nonspeech noise. Yet, two of the four subjects in the open response paradigm did not report hearing such noises. Similarly, fricative noises usually maintain a constant spectrum over their entire duration (Behrens and Blumstein, 1988), but a sequence of fricatives can have a shift in the middle (Zue and Shattuck-Hufnagel, 1979). Yet subjects sometimes heard only one fricative, even when the one they heard was supported by the shorter part of the noise. Although individual differences are difficult to incorporate into any theory, the differences found here seem to make use of the several aspects of the signal, confirming the multiplicity of cues rather than the exclusive use of any subset. While this makes an exact description cumbersome, these theories do allow for a description of the relevant factors.

Template matching theories have more difficulty incorporating these results. The most explicit of these models (Stevens, 1985) accounted for the /s/–/ʃ/ distinction in English primarily via the comparison of signal amplitude in two spectral slices around the consonant release. For the fricatives before the vowel /u/ in the present stimuli, the acoustic difference predicted does not hold, though perhaps some modifications to allow for certain vowel differences could be made. This template would have to be modified in any case to encompass cases in which fricative noises are not immediately adjacent to vowels. This includes cases of two adjacent fricatives that change during the noise (e.g., "'s sure hot in here") to the more common case of a fricative before a stop or liquid. While such changes can be envisioned, the main results of the present experiments suggest that even a modified template will be inadequate.

When the noise is not completely from one fricative category but instead changes from one to the other in the noise, the resulting percept is not determined by the time intervals at the release. Neither is the percept directly computable from those intervals around the spectral change in the noise, nor from any combination of those two regions of change. Instead, different portions of the noise are given different perceptual weight depending on the fricative involved, the duration of each portion, and the transitions present in the vocalic segment. Such results would require some contribution of the duration of the various segments, which would require some variety of timekeeper. A timekeeper seems necessary in any event to account for languages with geminate consonants or distinctive vowel length, but it might also be used for the fricative distinction at hand (Behrens and Blumstein, 1988). The perceptual evidence of Jongman (1989) indicates that windows as short as those proposed by Stevens may not be sufficient for the entire set of English fricatives, a finding which might become incorporated into a feature-with-time theory. The means of incorporating the information provided by a timekeeping function with the more direct representations in the spectrum must await further elaboration. From the current work, we can tell at least that the strength of some of the features extracted automatically from the spectral slices would have to depend on the duration of the signal encompassing that slice.

The contribution of the formant transitions is also problematic from a template matching perspective. While the formant transitions play a role in the revised template account (Stevens, 1985, p. 249), they do so only when the primary cue is not present. In the present results, transitions played a role even when the primary cue was present. (Note that even though portions of the noise and portions of the vocalic segment are used in the template, it is only the amplitude of the two $F3$'s that is used, so the template does not include transitions *per se*.) The transitions either shifted the boundaries (experiments 1–3) or significantly often reversed the order of the fricative judgments (experiment 5). Their status as "secondary," then, is questionable. Indeed, one of the main defining aspects of secondary features is that they are learned relatively late in language acquisition (Stevens and Blumstein, 1978, p 1367). Transitions for /s/ and /ʃ/, though, have been shown to be weighted *more* heavily by younger children than by older children and adults (Nittrouer and Studdert-Kennedy, 1987).

It might be possible to recast the role of secondary features to be more like that of enhancing features, that is, that they should play a role in perception in all cases. Enhancing features (Stevens *et al.*, 1986; Stevens and Keyser, 1989) are meant to increase the salience of the primary features. They are claimed to be potentially continuous rather than discrete (Stevens *et al.*, 1986, p. 426), but all of the examples given are in fact discrete features. In addition, as Repp (1986) points out, most of the examples of enhancing features actually seem to enhance the acoustic realization of the primary feature directly, rather than being a separately perceivable feature. For the fricatives, the transitions are a secondary cue, but they would presumably be interpreted much like a stop's transitions, yielding values of the features anterior and coronal. Thus they might represent an alternative to the noises feature values, not necessarily an enhancement. The fricatives include a combination of features that is not present in the English stops. /ʃ/ is − anterior while /s/ is + anterior, though both are + coronal. In the stops, all + coronals are + anterior. This would perhaps not be a problem if the stop features themselves were treated in terms of features, not categories as was originally done (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979). Although it might be possible to divide the templates in Blumstein and Stevens (1979) from three categories to two binary features, one cannot assess the success of such a move without further tests. Elaborated metrics such as that in Lahiri *et al.* (1984) might make the features more available, but the conditions under which those features would combine with the fricative (rather than, say, contribute to a stop percept) remain to be worked out.

Coupled with the evidence that transitions are taken into account even when they do not determine the overt judgments (Whalen, 1984), these considerations make it unlikely that the template proposed (Stevens, 1985) is in fact operative in speech. If the theory is modified so that secondary features are always considered even when the primary features are sufficient, then the original appeal of the theory is lessened, as it would then be more complex. That is, all the aspects of the speech signal would be taken into account, just as they are in gestural models; the role of the original "primary" cues would be decreased substantially (Blumstein *et al.*, 1982, Walley and Carrell, 1983).

In short, fricative judgments depend on the cues that are the ones most often cited: the frequency, amplitude, and duration of the noise, and the location information in the formant transitions. While amplitude differences between brief time intervals on either side of a major spectral change may be consistently observed (though counterexamples were present even in the current natural stimuli), they do not predict the way many stimuli are perceived. The present results do not show listeners generating as many percepts as they ought to based on spectral slices, and the contribution of "secondary" cues occurs even when the primary cues should have sufficed, providing little support for the template matching approach. Instead, human listeners seem to use as much of the information specifying the English fricatives /s/ and /ʃ/ as they can, as shown here when they are presented with acoustic hybrids.

ACKNOWLEDGMENTS

The research and writing of this paper were supported by NIH Grant No. HD-01994 to Haskins Laboratories. Suzanne Boyce provided helpful comments, besides running the subjects. Carol A. Fowler, Michael Studdert-Kennedy, Bruno H. Repp, Richard McGowan, and Andrea Levitt made helpful suggestions. The comments of Ken Stevens and three anonymous reviewers on earlier drafts led to major improvements.

[1] It is interesting to note that the stimuli were all consistent with Stevens' (1985) acoustic criterion for the /s/ category. That is, the amplitude of the fricative noise in the $F3$ region relative to the vowel's $F3$ amplitude was − 16 dB for /s/ and − 7 dB for /ʃ/. For the noises adjacent to the transitions from the other category, the values are − 8 dB for the /s/ noise with /ʃ/ transitions, and − 15 dB for the /ʃ/ noise with /s/ transitions. The amplitude measurements were based on LPC spectra of a 26-ms window just before and just after the end of the fricatives noise. First differencing, as used in Blumstein and Stevens (1979), was omitted, since the high frequencies had already been pre-emphasized in hardware. Since the Nyquist frequency was 10 kHz rather than 5 kHz, a 24-pole algorithm was used here rather than the 14-pole version used by Blumstein and Stevens (1979), giving equivalent spectral resolution in the lower frequency region. Since only those values within 2 dB of equality are claimed to be ambiguous (Stevens, 1985, p. 245), there should have been no reliance on secondary cues. Perhaps this simply indicates that the vowel /u/ does not bear the same relation to the fricative noise as the vowels previously examined by Stevens. If so, then the metric needs some revision. But if not, then both of these tokens should have been unambiguously "suit." As it turns out, the syllables with nonhybrid noises were identified as intended 98% of the time (see also Whalen, 1985).

Behrens, S. J., and Blumstein, S. E. (1988). "Acoustic characteristics of English voiceless fricatives: a descriptive analysis," J. Phon. **16**, 295–298.

Best, C. T., Morrongiello, B., and Robson, R. (1981). "Perceptual equivalence of acoustic cues in speech and nonspeech perception," Percept. Psychophys. **29**, 191–211.

Blumstein, S. E. (1986). "On acoustic invariance in speech," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 178–197.

Blumstein, S. E., Isaacs, E., and Mertus, J. (1982). "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **72**, 43–50.

Blumstein, S. E., and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. 66, 1001–1017.
Browman, C. P., and Goldstein, L. (1985). "Dynamic modeling of phonetic structure," in *Phonetic Linguistics*, edited by V. A. Fromkin (Academic, New York), pp. 35–53.
Browman, C. P., and Goldstein, L. (1989). "Articulatory gestures as phonological units," Phonology 6, 201–251.
Browman, C. P., and Goldstein, L. (1990). "Gestural specification using dynamically-defined articulatory structures," J. Phon. 18, 299–320.
Edman, T. R. (1979). "Discrimination of intraphonemic differences along two place of articulation continua," in *Speech Communication Papers*, edited by J. J. Wolf and D. H. Klatt (Acoustical Society of America, New York), pp. 455–458.
Fowler, C. A. (1981). "Production and perception of coarticulation among stressed and unstressed vowels," J. Speech Hear. Res. 46, 127–149.
Fowler, C. A. (1984). "Segmentation of coarticulated speech in perception," Percept. Psychophys. 36, 359–368.
Fowler, C. A., and Smith, M. R. (1986). "Speech perception as vector analysis: An approach to the problem of invariance and segmentation," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 123–136.
Harris, K. S. (1958). "Cues for the discrimination of American English fricatives in spoken syllables," Lang. Speech 1, 1–7.
Hughes, G., and Halle, M. (1956). "Spectral properties of fricative consonants," J. Acoust. Soc. Am. 28, 303–310.
Jakobson, R., Fant, G., and Halle M. (1963). *Preliminaries to Speech Analysis* (MIT, Cambridge, MA).
Jongman, A. (1989). "Duration of frication noise required for identification of English fricatives," J. Acoust. Soc. Am. 85, 1718–1725.
Jongman, A., Blumstein, S. E., and Lahiri, A. (1985). "Acoustic properties for dental and alveolar stop consonants: a cross-language study," J. Phon. 13, 235–251.
Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. 73, 322–335.
Krakow, R. A. (1989). "The effects of syllable structure and stress on movements of the velum and lower lip," Unpublished doctoral dissertation, Yale University.
Kurowski, K., and Blumstein, S. E. (1987). "Acoustic properties for place of articulation in nasal consonants," J. Acoust. Soc. Am. 81, 1917–1927.
Lahiri, A., Gewirth, L., and Blumstein, S. E. (1984). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. 76, 391–404.
Lehiste, I. (1960). "An acoustic–phonetic study of internal open juncture," Phonetica (Supplement) 5.
Liberman, A. M., and Mattingly, I. G. (1985). "The motor theory of speech perception revised," Cognition 21, 1–36.
Lisker, L. (1986). " 'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Lang. Speech 29, 3–11.
Mann, V. A., and Repp, B. H. (1980). "Influence of vocalic context on perception of the /ʃ/–/s/ distinction," Percept. Psychophys. 28, 213–228.
May, J. (1976). "Vocal tract normalization for /s/ and /ʃ/," Haskins Lab. Status Rep. Speech Res. SR-48, 67–73.
Nittrouer, S., and Studdert-Kennedy, M. (1987). "The role of coarticulatory effects in the perception of fricatives by children and adults," J. Speech Hear. Res. 30, 319–329.
Pisoni, D. B., and Lazarus, J. H. (1974). "Categorical and noncategorical modes of speech perception along the voicing continuum," J. Acoust. Soc. Am. 55, 328–333.
Repp, B. H. (1986). "Comment [on K. N. Stevens, S. J. Keyser, and H. Kawasaki (1986)]," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 449–455.
Samuel, A. G. (1977). "The effect of discrimination training on speech perception: Noncategorical perception," Percept. Psychophys. 22, 321–330.
Soli, S. D. (1981). "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation," J. Acoust. Soc. Am. 70, 976–984.
Stevens, K. N. (1980). "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. 68, 836–842.
Stevens, K. N. (1985). "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics*, edited by V. A. Fromkin (Academic, Orlando, FL), pp. 243–255.
Stevens, K. N., and Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. 64, 1358–1368.
Stevens, K. N., Blumstein, S. E., and Glicksman, L. B. (1987). "Voicing distinction for fricatives: Acoustic theory and measurements," J. Acoust. Soc. Am. Suppl. 1 82, S16.
Stevens, K. N., and Keyser, S. J. (1989). "Primary features and their enhancement in consonants," Language 65, 81–106.
Stevens, K. N., Keyser, S. J., and Kawasaki, H. (1986). "Toward a phonetic and phonological theory of redundant features," in *Invariance and Variability in Speech Processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 426–449.
Strevens, P. (1960). "Spectra of fricative noise in human speech," Lang. Speech 3, 32–49.
Vaissiere, J. (1985). "Speech recognition: A tutorial," in *Computer Speech Processing*, edited by F. Fallside and W. A. Woods (Prentice Hall, Englewood Cliffs, NJ), pp. 191–242.
Walley, A. C., and Carrell, T. D. (1983). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. 73, 1011–1022.
Whalen, D. H. (1981). "Effects of vocalic formant transitions and vowel quality on the English [s]–[š] boundary," J. Acoust. Soc. Am. 69, 275–282.
Whalen, D. H. (1983). "Vowel information in post-vocalic fricative noises," Lang. Speech 26, 91–100.
Whalen, D. H. (1984). "Subcategorical phonetic mismatches slow phonetic judgments," Percept. Psychophys. 35, 49–64.
Whalen, D. H. (1985). "Effects of subcategorical mismatches on shadowing," J. Acoust. Soc. Am. Suppl. 1 77, S9.
Whalen, D. H. (1989). "Vowel and consonant judgments are not independent when cued by the same information," Percept. Psychophys. 46, 284–292.
Whalen, D. H. (1990). "Coarticulation is largely planned," J. Phon. 18, 3–35.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1990). "Gradient effects of fundamental frequency on stop consonant voicing judgments," Phonetica 47, 36–49.
Whalen, D. H., and Samuel, A. G. (1985). "Phonetic information is integrated across intervening nonlinguistic sounds," Percept. Psychophys. 37, 579–587.
Whalen, D. H., Wiley, E. R., Rubin, P. E., and Cooper, F. S. (1990). "The Haskins Laboratories' pulse code modulation (PCM) system," Behav. Res. Methods Instrum. Comput. 22, 550–559.
Yeni-Komshian, G. H., and Soli, S. G. (1981). "Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation," J. Acoust. Soc. Am. 70, 966–975.
Zue, V. W., and Shattuck-Hufnagel, S. (1979). "The palatalization of alveolar fricatives in American English," in *Proceedings of the 9th International Congress of Phonetic Sciences, Volume 1*, edited by E. Fischer-Jørgensen (Jensen, Copenhagen, Denmark), p. 215.