Perception & Psychophysics
1984, 36 (6), 523-530

Parameters of spectral/temporal fusion in speech perception

BRUNO H. REPP
Haskins Laboratories, New Haven, Connecticut

and

SHLOMO BENTIN
Haskins Laboratories, New Haven, Connecticut
and Aranne Laboratory of Human Psychophysiology, Hadassah Hospital, Jerusalem, Israel

When the distinctive formant transition of a synthetic syllable is presented to one ear while the remainder (the "base") is presented to the opposite ear, listeners report hearing the original syllable in the ear receiving the base—a phenomenon called "spectral/temporal fusion" by Cutting (1976). We have found that the mere onset (i.e., the first pitch pulse, 10 msec in duration) of an isolated, contralateral third-formant (F3) transition can be sufficient to cue the /da/-/ga/ distinction in this way. We also varied the relative onset times of isolated F3 and base and compared three types of F3 segments (50-msec time-varying, 50-msec constant, 10-msec onset) under both dichotic and diotic presentation. Time-varying F3 segments were superior to constant ones, especially when they lagged behind the base. Diotic performance exceeded dichotic performance, but only when F3 preceded the base, suggesting that upward spread of masking occurred in diotic presentation when F3 coincided with energy in the lower formants. Perhaps most interestingly, subjects' tolerance of temporal asynchrony (roughly ±50 msec) was about the same in dichotic and diotic conditions, suggesting that the temporal integration mechanism that combines phonetic information from the isolated F3 segment and the base operates similarly in both conditions.

It has long been known that perceptual fusion results when the first formant (F1) of a synthetic speech signal is presented to one ear while the higher formants are simultaneously presented to the other ear (Broadbent, 1955; Broadbent & Ladefoged, 1957). In this situation, listeners perceive a single fused stimulus localized toward the side of F1 (see Darwin, Howell, & Brady, 1978). A variant of this paradigm was introduced by Rand (1974), who presented only the time-varying F2 and F3 transitions of CV syllables to one ear while F1 and the steady-state portions of F2 and F3 were presented to the opposite ear. The perceptual fusion that occurs in this situation has been labeled "spectral/temporal fusion" by Cutting (1976).

Spectral/temporal fusion has received considerable attention in recent years. Research on "duplex perception" (Bentin & Mann, 1983; Liberman, 1979; Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Nusbaum, Schwab, & Sawusch, 1983; Repp, Milburn, & Ashkenas, 1983) has focused on the fact that, simultaneously with the speech, the isolated formant transition is perceived as a nonspeech "chirp." Thus the isolated transition contributes to phonetic and nonphonetic percepts at the same time, which has been interpreted as evidence for the simultaneous operation of a speech-specific and a general auditory mode of perception (Liberman, 1982; Liberman et al., 1981; Mann & Liberman, 1983). Recent studies have shown that the speech and nonspeech percepts in this situation are affected in different degrees by manipulations such as masking or attenuation of the distinctive isolated transition (Bentin & Mann, 1983).

In the present studies, we were not directly concerned with duplex perception as such. Rather, we focused on the speech percept only and examined some of the factors that may limit the occurrence of fusion in this special situation. By "fusion," we mean here the contribution of the isolated transition to speech identification. The strict definition of fusion as a single stimulus percept from two separate inputs clearly does not apply in duplex perception. The purpose of Experiment 1 was to determine how long the distinctive isolated formant transition must be to enable listeners to discriminate between two alternative syllables when attending to the ear receiving the nondistinctive base. Experiment 2 is a parametric study of the effects of temporal asynchrony on spectral/temporal fusion, including comparisons of dynamic and static "transitions" and of dichotic versus diotic presentation.

This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories. Shlomo Bentin was supported by a stipend from the Jesselson foundation. We are grateful to Alvin Liberman, Michael Studdert-Kennedy, and the reviewers for their helpful comments. The authors' mailing address is: Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695.

EXPERIMENT 1

All previous studies of spectral/temporal fusion have followed the standard paradigm described above. In each case, a complete formant transition was presented to the ear contralateral to the base, although the duration of the isolated transition varied from 30 to 70 msec across different studies. In the present study, we wished to determine, first, whether the full transition was needed to make the speech distinction, or whether a truncated version or even just the onset of the transition would suffice. Second, we asked whether the presence of the steady-state continuation of the same formant in the base was a necessary condition for spectral/temporal fusion to occur. The second half of the term, "spectral/temporal," suggests that an affirmative answer was assumed by Cutting (1976). To test this inference, we omitted from the base the steady-state resonance following the critical transition, expecting (on the basis of pilot observations) that fusion would nevertheless be obtained. (A direct comparison of conditions with and without this steady-state formant in the base was conducted in Experiment 2.)

The materials used were the syllables /da/ and /ga/, synthesized so as to differ only in the F3 transition. Earlier studies have obtained strong spectral/temporal fusion with similar stimuli (Mann & Liberman, 1983; Repp et al., 1983). The experimental manipulation in Experiment 1, then, was to reduce the duration of the isolated F3 transition (appropriate for either /da/ or /ga/) until only its onset (i.e., the first pitch pulse) remained, while a constant two-formant base was presented in synchrony to the opposite ear. Spectral/temporal fusion was assessed in terms of subjects' ability to distinguish /da/ and /ga/ in the ear receiving the base.

Method

Subjects. Twelve subjects (three males, nine females) were tested. They were all Yale undergraduates and were paid for their participation.

Stimuli. The stimuli were three-formant synthetic approximations of the syllables /da/ and /ga/, produced on the parallel software synthesizer at Haskins Laboratories, as illustrated schematically in Figure 1. The first two formants were identical in both syllables, and constituted the "base." The duration of the base was 250 msec with a 50-msec amplitude ramp at onset and a constant fundamental frequency of 100 Hz for the first 100 msec, followed by a linear decrease to 80 Hz at offset. The first formant began at 279 Hz and increased linearly in frequency during the first 50 msec to a steady state of 765 Hz. The second formant began at 1650 Hz and decreased linearly in frequency during the first 50 msec to a steady state of 1230 Hz. The base by itself is perceived as either /da/ or /ga/ or as ambiguous, depending on the listener. The /da/ third-formant transition, originally 50 msec (5 pitch pulses) in duration, began nominally at 2800 Hz and decreased linearly in frequency to 2550 Hz, whereas the /ga/ transition began nominally at 1800 Hz and increased linearly in frequency to 2550 Hz. (These are the "dynamic" transitions in Figure 1; the actual F3 frequencies in the first pitch pulse were 2775 and 1875 Hz, respectively—see caption to Figure 1.) Five transition durations were used, as indicated by the tick marks in Figure 1: 50, 40, 30, 20, and 10 msec (5, 4, 3, 2, and 1 pitch pulses, respectively). Since the frequency trajectory was not changed, the shorter transitions had offset frequencies increasingly closer to the onset frequencies.

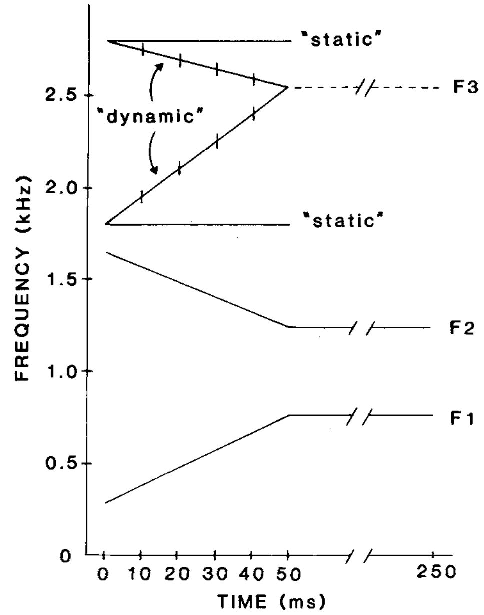

Figure 1. Schematic illustration of the center frequencies of the first three formants in the stimuli of Experiments 1 and 2. All formant transitions are drawn as idealized linear functions connecting the nominal frequencies used in synthesis. The formant frequencies were actually constant within each pitch pulse at values halfway between the nominal onset and offset frequencies for that 10-msec period. The "dynamic" transitions were used in both experiments; the tick marks indicate the shortening manipulation in Experiment 1. The "static" F3 segments were used in Experiment 2 only. The dashed line represents the F3 steady state present *in the base* on half of the trials in Experiment 2.

The stimuli were recorded onto magnetic tape, with the isolated F3 transitions on one channel and the onset-aligned, constant base on the other. There were 240 stimuli altogether: 24 repetitions of the /da/ and /ga/ transitions at each of five durations. The stimuli were arranged in five randomized sequences, with ISIs of 2.5 sec between stimuli and longer intervals between sequences.

Procedure. The tapes were presented at a comfortable intensity over TDH-39 earphones in a quiet room. The base was always in the left ear and the F3 transition was in the right ear. (No pronounced ear asymmetries have been observed in this task.) The subjects were instructed to attend to the left ear and to identify the syllables in writing as beginning with either "d" or "g."

Results and Discussion

Performance for 50-, 40-, and 30-msec transitions was nearly perfect: 96%, 97%, and 98% correct, respectively. For 20-msec transitions, performance dropped to 91% correct, and for 10-msec transition onsets, to 84% correct. Individual subjects' scores in the last condition ranged from 66% to 96% correct. Thus, although there was some loss in accuracy, even the 10-msec single pitch-pulse transition onsets were sufficient to distinguish /da/ and /ga/ in the opposite ear. Accordingly, time-varying frequency information in F3 does not seem essential either for this particular phonetic distinction or for spectral/temporal fusion to occur.

In addition, it is clear that the absence of the F3 steady state in the base did not prevent fusion. Since temporal continuity in the relevant frequency band thus seems to contribute little (see also Experiment 2), spectral/temporal fusion appears to be just a special case of spectral fusion (Cutting's, 1976, term for the fusion of complete formants presented simultaneously to different ears). The difference lies in that only the former situation gives rise to a duplex percept (syllable and "chirp"); the mechanism that reconstitutes the speech percept from separate components, however, seems to be the same.

It might be argued that the subjects accomplished their task by paying attention to the chirp-like isolated transition and responding "g" when the chirp was low-pitched and "d" when it was high-pitched (see Nusbaum et al., 1983). Even though no catch trials were employed in the present study, this possibility is virtually ruled out by previous evidence that (1) subjects do attend to the ear receiving the base when instructed to do so (Mann & Liberman, 1983; Repp et al., 1983), and (2) they are unable to associate isolated F3 chirps consistently with the response categories "d" and "g" (Repp et al., 1983). Moreover, all listeners agree that the syllables in the ear receiving the base really do sound alternately like /da/ and /ga/. Therefore, the present subjects' responses almost certainly reflect the combination of information from the two ears.

It may be noted that a 10-msec F3 onset is not only devoid of time-varying information but is also nonperiodic, consisting only of a single glottal cycle. By itself, it sounds like a click. Informally, we have confirmed that fusion is also obtained when this 10-msec pitch pulse is replaced with a 10-msec burst of noise with the same spectral envelope, generated by the aperiodic source of the synthesizer. This observation reveals a possible similarity with a phenomenon reported by Pastore, Szczesiul, Rosenblum, and Schmuckler (1982), who found that a burst of filtered white noise changed the perception of a contralateral /pa/ to /ta/. These findings indicate that dichotic integration of phonetic information can occur even if the signal in one ear is periodic and the other is not. It is not clear whether such phenomena should be attributed to general processes of auditory fusion. Rather, they may constitute evidence for a central phonetic decision mechanism that operates on inputs from both ears.

EXPERIMENT 2

To explore in greater detail the parameters of spectral/temporal fusion, we conducted a multifactorial experiment that included four independent variables: (1) a range of onset asynchronies between the isolated F3 segment and the base, (2) dichotic versus diotic presentation, (3) static (constant frequency) versus dynamic (time-varying frequency) F3 segments, and (4) bases with and without a steady-state F3.

Effects of stimulus onset asynchrony (SOA) on spectral/temporal fusion were studied by Cutting (1976) with synthetic two-formant stimuli. The isolated F2 transition was 70 msec in duration. Cutting used transition-base lead and lag times of up to 160 msec, spaced in logarithmic steps, but reported his results averaged over leads and lags, since he found no significant asymmetry. As expected, speech identification performance dropped as SOA increased. However, performance was still slightly above chance even at the longest interval (160 msec), although the statistical significance of this finding was not determined. The longest interval at which performance was substantially above chance was 40 msec.

In a recent study, Bentin and Mann (1983; Experiment 1) used SOAs of up to 100 msec with two-formant syllables similar to Cutting's, although the transitions were only 50 msec in duration. Only lead times were used; that is, the F3 segment always preceded the base. Subjects' performance declined steadily with increasing SOA, but was still above chance at the 100-msec interval. These results are consistent with Cutting's in that they suggest a considerable tolerance of temporal asynchrony in spectral/temporal fusion.

In the present study, we sought to replicate these findings with stimuli distinguished by a difference in the F3 transition. Particular attention was given to possible performance asymmetries between lead and lag times. Cutting's (1976) negative finding notwithstanding, such asymmetries might be predicted on at least two grounds. First, when the F3 segment lags behind the onset of the base and thus coincides with the vowel, it may suffer some contralateral simultaneous masking which is absent when the F3 segment precedes the base. Second, when the F3 segment lags behind, listeners may conceivably be able to classify the base phonetically before processing the F3 segment. Both considerations predict stronger fusion when the F3 segment leads the base than when it lags behind. On the other hand, one might also predict the opposite: It is known that, in auditory perception, the terminal frequency of a tone glide is more salient than its initial frequency (Nabelek, Nabelek, & Hirsh, 1970; Schwab, 1981). If a leading F3 segment is retained in auditory memory before it is integrated with the base, its distinctiveness might be reduced, because full /da/ and /ga/ transitions have the same terminal frequency. This may confer a relative advantage on lagging F3 segments, which need not be stored in auditory memory.

A second comparison in Experiment 2 concerned dichotic versus diotic presentation of the stimulus components. Rand (1974) conducted such a comparison for onset-synchronous transition and base and found better speech discrimination in the dichotic condition. He attributed this to simultaneous masking of higher by lower formants in the diotic condition and to release from this form of peripheral upward spread of masking in the dichotic condition. Subsequent studies (e.g., Danaher & Pickett, 1975; Nearey & Levitt, 1974; Nye, Nearey & Rand, 1974) have replicated this difference, although there are also negative findings in the literature (Nusbaum et al., 1983; Repp et al., 1983). This is the first study to vary

SOA in such a comparison. If upward spread of masking operates, then the advantage of dichotic over diotic performance should hold at all lag times, as long as the F3 segment coincides with the base. However, no such difference should exist at lead times, unless there is significant peripheral backward masking of the F3 segment by the base, which seems unlikely.

Another question of interest was whether listeners would be equally tolerant of stimulus onset asynchronies in diotic and in dichotic presentation. Presented monotically or diotically, onset-synchronous transition and base constitute, of course, an intact syllable. It has not been attempted previously to advance or delay the isolated transition with respect to the base when both occur in the same channel. At least one dichotic fusion phenomenon (the influence of a contralateral white noise burst on the perceived place of articulation of a stop consonant) does not seem to occur when the stimulus components are presented diotically (Pastore et al., 1982). We considered it possible that fusion of transition and base in the diotic condition might be restricted to short SOAs, where there is physical overlap, whereas in the dichotic condition subjects might be less sensitive to temporal asynchronies.

A third comparison of interest concerned the nature of the F3 segment conveying the distinctive information. Three kinds of F3 segments were compared: (1) standard 50-msec time-varying ("dynamic") F3 transitions, (2) short 10-msec onsets (as in Experiment 1), and (3) 50-msec constant ("static") F3 segments, which were obtained by extending the transition onset frequencies, as illustrated in Figure 1. The static F3 segments were of special interest: First, would they be sufficient to cue the /da/-/ga/ distinction? (The effectiveness of the short F3 segments in Experiment 1 suggests a positive answer.) Second, would they be as effective as dynamic F3 segments, or does the dynamic information convey additional phonetic distinctiveness? Third, the static F3 segments for /da/ and /ga/ have distinctive terminal (as well as initial) frequencies, which may be an advantage at F3 lead times. Up to a lead time of 40 msec, the distinctive end of a static F3 segment actually still overlaps with the onset of the base. As a result, performance at short lead times may be better for static than for dynamic F3 segments, unless the distinctive phonetic information derives strictly from F3 onset and physical overlap is irrelevant. Comparisons with the short F3 segment should also be enlightening in that regard, although the short duration of this stimulus entails a loss in energy and a consequent decrement in discriminability.

In addition to these three major factors (SOA, mode of presentation, and type of F3 segment), the experiment also included a comparison of bases with and without an F3 steady state. Since Experiment 1 had shown strong fusion in the absence of an F3 steady state, little effect of this last factor was expected.

Method

Subjects. Twelve paid volunteers participated, six men and six women. Five of them had been subjects in Experiment 1. Of the other seven, two had to be replaced because of exceedingly poor performance.

Stimuli. The basic stimuli were the same as in Experiment 1. In addition to the base used there, a second base was used which included a steady-state F3 at 2550 Hz, starting 50 msec after the onset of F1 and F2, at the same time as the steady states of these formants. (The vowel had very nearly the same quality with and without F3.) There were three kinds of F3 segments: The dynamic (50-msec) and short (10-msec) versions corresponded to the extremes of transition duration used in Experiment 1; the static (50-msec) F3 segments were synthesized at constant frequencies corresponding to the nominal onset frequencies of the dynamic segments (see Figure 1).

Three stimulus tapes were recorded, each corresponding to a different type of F3 segment. Each tape contained 10 blocks of 22 stimuli, each block being a randomization of the two F3 segments for /da/ and /ga/ recorded on one track, at 11 different SOAs in relation to the base on the other track. The 11 SOAs were : −100, −70, −40, −20, −10, 0, 10, 20, 40, 70, and 100 msec; a negative SOA means that the F3 segment led the base. In addition, odd-numbered blocks contained the base without F3, and even-numbered blocks contained the base with a steady-state F3. The ISI was 2 sec, and there were 6 sec between blocks.

Design and Procedure. Each of the three stimulus tapes was presented in two conditions: dichotic and diotic. All six conditions were presented in a single session. The order of conditions was strictly counterbalanced across subjects, with the constraint that all diotic conditions either preceded or followed all dichotic conditions.

A brief familiarization sequence with dynamic F3 segments at SOA=0 was presented at the beginning of the session. This sequence included 10 stimuli in which /da/ and /ga/ alternated, followed by a random arrangement of 20 stimuli. The sequence was first presented diotically and then dichotically. The subjects tried to identify the syllables and were given feedback after the sequence. If more than a few errors were committed, the sequence was presented a second time.

The subjects were run individually under the same conditions as in Experiment 1. The tape-recorder channels were calibrated for equal intensity of a repeated vowel. Diotic presentation was achieved by mixing the two channels together and feeding the result to both earphone channels. No intensity adjustment was made; because of the relative weakness of the F3 segment, the increase in the total amplitude of the mixed syllables over the isolated base was minimal. In the dichotic conditions, the F3 segment was presented to the right ear for half of the subjects and to the left ear for the other half.

The structure of the stimuli and of the test tapes was explained to the subjects in advance. They were asked *not* to rely on the high or low pitch of the F3 segment and to focus attention on the speech percept only. A forced choice between "d" and "g" responses was required for each stimulus.

Results

The main results are shown in Figure 2, where the percentage of correct consonant identifications is plotted as a function of SOA (abscissa), type of F3 segment (separate functions), and presentation condition (separate panels). A five-way repeated-measures analysis of variance was conducted which included, in addition to the three factors just mentioned, type of base and high/low F3 as factors; that is, the statistical analysis was conducted on "g" responses (or equivalently, "d" responses), not on percent correct. In this analysis, all effects with respect to percent correct are interactions involving the high/low F3 factor.

The first result evident from Figure 2 is that SOA had a clear effect: Performance decreased as SOA increased

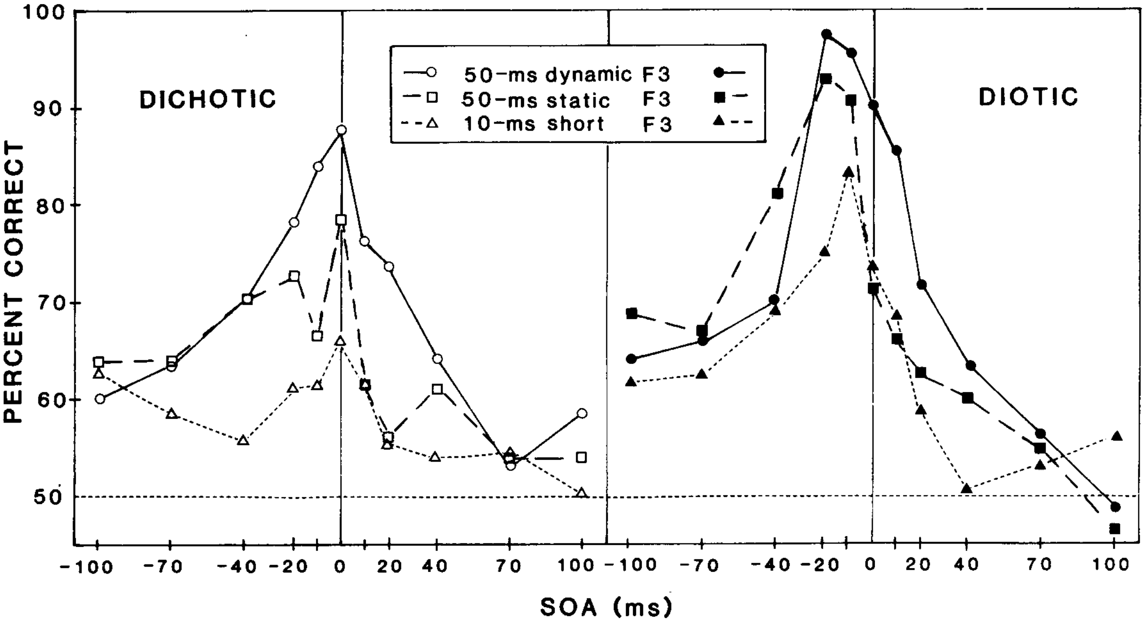

Figure 2. Percent correct as a function of SOA, separately for dichotic and diotic conditions, with type of F3 segment as parameter.

in either direction [F(10,110) = 32.07, p < .0001]. A second clear effect is that of type of F3 segment: Performance was generally best for the dynamic F3 segments and poorest for the short F3 segments [F(2,22) = 11.02, p < .0005]. Performance for the short F3 segments at SOA=0 in the dichotic condition was a good deal worse than in Experiment 1, for reasons that are not obvious. The third main effect evident from the figure is that, unexpectedly, performance in the diotic condition was higher than in the dichotic condition [F(1,11) = 7.06, p < .03].

Because of the general convergence of scores at the extremes of the SOA range, interactions with SOA also reflect main effects, at least in part. These interactions were highly significant for both type of F3 segment [F(20,220) = 5.51, p < .0001] and presentation condition [F(10,110) = 8.22, p < .0001]. Despite this latter interaction, listeners' tolerance of SOAs seemed similar in the two presentation conditions. No other effects on percent correct were significant.

Some more detailed differences in Figure 2 are not directly captured by the statistical analysis but deserve attention. First, in the dichotic condition, performance was generally best at SOA=0, as expected, but in the diotic condition, optimal performance was at short negative SOAs. Second, the effect of SOA was generally asymmetric, though more so in the diotic than in the dichotic condition: performance was generally better when the F3 segment led the base than when it lagged behind. This was especially true for the longest intervals used: At −70 and −100 msec of SOA, performance was clearly above chance (p < .05 for 11 of 12 conditions by sign test), whereas scores were near chance at 70 and 100 msec of SOA (p < .05 for only 1 of 12 conditions). Indeed, the absence of any decline in performance between −70 and −100 msec of SOA suggests an asymptote that may reflect an effect other than spectral/temporal fusion, such as a response bias contingent on the perceived pitch of the F3 segment. Third, it may be noted that the superiority of dynamic over static F3 segments did not hold at lead times of −40 msec or more, and that the superiority of static over short F3 segments was much more pronounced at negative than at positive SOAs.

One consequence of the differential asymmetry of the effect of SOA in the dichotic and diotic conditions is that diotic performance exceeded dichotic performance primarily at short F3 segment lead times. This is especially clear from Figure 3, where the difference between diotic and dichotic scores is plotted. It is also evident that this difference is similar for all three types of F3 segments. (The relevant interaction was not significant.)

The statistical analysis revealed several additional effects which related specifically to the percentage of "g" (or "d") responses, rather than to percent correct. Figure 4 shows the percentage of "g" responses as a function of SOA, high/low F3, and type of base; the scores are averaged over the three types of F3 segment and the two presentation conditions. Naturally, there were more "g" responses to stimuli including the low F3 than to stimuli including the high F3 [F(1,11) = 166.84, p < .0001]. It is also evident that the effect of the low F3 segment, which increased "g" responses when effective, was larger than that of the high F3, which decreased "g" responses, so that the total number of "g" responses varied significantly with SOA [F(10,110) = 5.31, p < .0001]. Of course, the interaction of high/low F3 and SOA was highly significant; it corresponds to the main effect of SOA on percent correct, reported above. It may also be noted that the asymmetry around SOA=0 at short

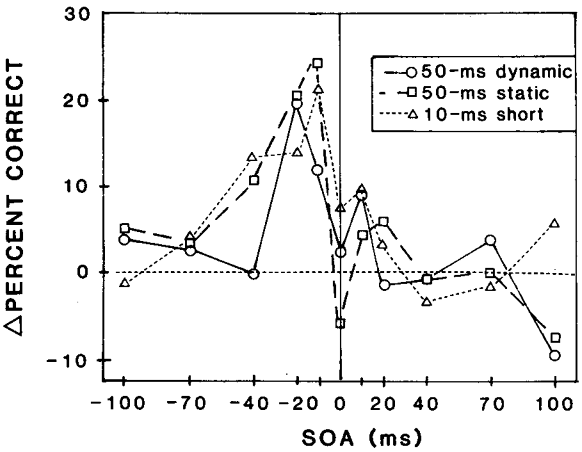

Figure 3. Difference between diotic and dichotic scores (Figure 2) as a function of SOA.

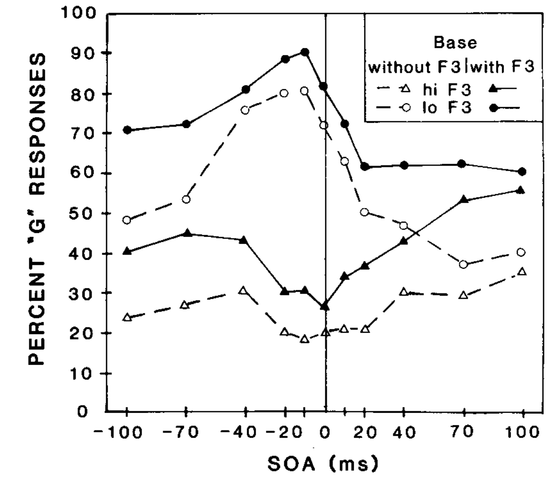

Figure 4. Percent "g" responses as a function of SOA, with high/low F3 and type of base (with or without F3 steady state) as parameters.

SOAs, deriving mainly from the diotic condition (see Figure 2), was pronounced only for low-F3 stimuli; the effect of SOA for high-F3 stimuli was more nearly symmetric. The asymmetry at long SOAs was equally present for both types of stimuli, however.

An unexpected result evident in Figure 4 is that, overall, more "g" responses were given when the base contained a steady-state F3 [F(1,11) = 17.13, p < .002]. The presence of a steady-state F3 apparently enhanced the spread of energy following the release, which is characteristic of velar consonants preceding back vowels. This difference was more pronounced at long than at short SOAs—F(10,110) = 7.16, p < .001, for the interaction —which confirms that the effect originated in the base. However, the effect also interacted with type of F3 segment [F(2,22) = 10.18, p < .007], being strongest with the short F3 segments and weakest with the dynamic F3 segments. Thus, the most effective F3 segments also were able to overcome most effectively the bias inherent in the base itself. A triple interaction between type of presentation, SOA, and high/low F3 was also obtained [F(10,110) = 3.44, p < .0006], suggesting that the bias was overcome more effectively by the F3 segments in the diotic condition. The differential SOA asymmetry in the two presentation conditions may also have contributed to this interaction.

Three additional significant interactions in the analysis of variance (between mode of presentation and high/low F3, between type of F3 segment and high/low F3, and between mode of presentation, type of F3 segment, and SOA) essentially parallel effects on percent correct described earlier and therefore need not be discussed any further.

DISCUSSION

Experiment 2, in conjunction with Experiment 1, investigated three factors that were expected to play a role in spectral/temporal fusion of speech stimuli: (1) structural properties of the isolated formant transition and of the base, (2) temporal asynchrony between the transition and the base, and (3) dichotic versus diotic presentation.

It is now clear that the isolated transition need not actually be a transition for fusion to occur. A steady-state formant with the same onset frequency, or even only the first pitch pulse of the transition can be sufficient, although the dynamic frequency transition does seem to convey additional information. Moreover, the base need not contain any continuation of the isolated F3 segment in the form of a steady-state F3. Experiment 2 has also shown that these same stimulus conditions enable listeners to discriminate /da/ and /ga/ in diotic presentation, when (at SOA=0) the stimulus components are physically intergrated and the F3 segment is not perceived as a separate nonspeech stimulus. What is different about the dichotic situation is the presence of the added nonspeech percept: Segregation by input channel is effective at an auditory level of perception but apparently leaves phonetic perception unaffected, at least in the present paradigm.

This conclusion is also supported by the finding that the range of SOAs over which above-chance speech discrimination was obtained was very similar in dichotic and diotic presentation. Thus, even when the isolated F3 segment preceded the base on the *same* channel, it was nevertheless (partially) integrated with the base into a phonetic percept. Thus, the expectation that listeners would be less tolerant of SOAs in diotic presentation was not borne out, and the present results in fact suggest that spectral/temporal fusion is not specific to dichotic presentation at all. Nor is duplex perception: the F3 segment preceding the base on the same channel is perceived as a nonspeech event—a case of monaural duplex perception. We conclude that perceptual integration in phonetic perception operates regardless of mode of stimulus presentation, and apparently regardless of whether the stimulus appears unitary or segregated at an auditory level of perception. Although there are some obvious limits to this dissociation, it nevertheless strengthens further the traditional distinction between speech and nonspeech modes of perception.

There were two kinds of asymmetries with respect to the effects of SOA. One of them was equally present in dichotic and diotic presentation: speech discrimination was above chance at long negative SOAs but dropped to chance at long positive SOAs. No such asymmetry was noted by Cutting (1976); however, the above-chance scores at long negative SOAs replicate the findings of Bentin and Mann (1983). Some of this asymmetry may be due to (central) masking of lagging F3 segments by the overlapping vowel; however, it seems that the above-chance performance with leading F3 segments is the finding in need of explanation. Only speculation is possible at this time. One possibility is that leading F3 segments are preserved in a (central) auditory memory and subsequently integrated with the base, whereas lagging F3 segments somehow cannot take advantage of auditory memory for the acoustically more complex base. Alternatively, identification of the F3 segment as "high" or "low" may have exerted a bias on speech identification, which was more pronounced when the F3 segment led than when it lagged the base. This explanation seems plausible, especially since the subjects were told about the correspondence of F3 segment pitch and phonetic category. Although they were also told to pay attention to the speech percept only, a certain amount of involuntary bias may have been introduced by leading F3 segments. This bias was equally present in diotic and dichotic presentation. Assuming, therefore, that the above-chance performance at long negative SOAs was not due to spectral/temporal fusion proper, the range of SOAs over which this type of fusion operates seems rather limited—roughly, ±50 msec.

The other asymmetry is the unexpected finding of optimal diotic performance at short negative SOAs. This was also the region where diotic performance exceeded dichotic performance. The following explanation may be proposed: Diotic integration of the stimulus components may have been uniformly superior to dichotic integration, but at positive SOAs diotic performance may have been lowered due to peripheral masking of the F3 segment by the lower formants contained in the base. Rand (1974) and many subsequent studies have suggested that dichotic segregation of a higher formant from F1 results in a release from upward spread of masking, which thus is largely a peripheral (channel-specific) effect. In fact, it was surprising that the present data did not show an absolute advantage for dichotic presentation at SOA=0 and at positive SOAs. The upward spread of masking explanation may account for another feature of the present data that seems difficult to explain in other terms: Apparently, the asymmetry in the diotic SOA effect was entirely due to the low F3; stimuli with a high F3 showed no such asymmetry. The reason for this may be that the high F3 evaded masking by the F1 and F2 transitions. The present data thus seem consistent with earlier findings on upward spread of masking, if the assumption is granted that dichotic fusion was not quite as strong as in some of the earlier studies.

An alternative possibility that comes to mind is that an F3 segment protruding from the base (at short negative SOAs) may have been perceived as a release burst. This would explain why speech identification was more accurate at short F3 lead times than at lag times, but it would not be clear why this asymmetry was present only in the diotic condition and only for the high-pitched F3. Nor did the 50-msec F3 segments sound like noisy release bursts; they had a distinct tonal quality. Thus, without additional assumptions yet to be spelled out, this interpretation cannot account for the data.

In summary, the present findings reveal dichotic spectral/temporal fusion to be a phenomenon that is neither specifically dichotic nor specifically temporal. The fact that a temporally or spatially segregated formant segment is audible as a separate nonspeech sound is not surprising; that such an auditorily segregated stimulus component still contributes to an integrated phonetic percept, however, is an observation that deserves continued attention. Although Pastore, Schmuckler, Rosenblum, and Szczesiul (1983) have reported a somewhat analogous phenomenon with musical stimuli, it is still possible to entertain the hypothesis that the fusion effect studied here reflects the operation of a central integrative mechanism specialized for phonetic perception. This hypothesis needs to be tested further with nonspeech analogs of speech stimuli used in studies of spectral/temporal fusion.

REFERENCES

Bentin, S., & Mann, V. A. (1983). Selective effects of masking on speech and nonspeech in the duplex perception paradigm. *Haskins Laboratories Status Report on Speech Research*, **SR-76**, 65-85.

Broadbent, D. E. (1955). A note on binaural fusion. *Quarterly Journal of Experimental Psychology*, **7**, 46-47.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, **29**, 708-710.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**, 114-140.

Danaher, E. M., & Pickett, J. M. (1975). Some masking effects produced by low-frequency vowel formants in persons with sensorineural hearing loss. *Journal of Speech and Hearing Research*, **18**, 261-271.

Darwin, C. J., Howell, P., & Brady, S. A. (1978). Laterality and localization: A right ear advantage for speech heard on the left. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.

Liberman, A. M. (1979). Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the IXth International Congress of Phonetic Sciences* (Vol. 2). Copenhagen: University of Copenhagen.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist*, **37**, 148-167.

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, **30**, 133-143.

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, **14**, 211-235.

Nabelek, I. V., Nabelek, A. K., & Hirsh, I. J. (1970). Pitch of tone bursts of changing frequency. *Journal of the Acoustical Society of America*, **48**, 536-553.

Nearey, T. M., & Levitt, A. G. (1974). Evidence for spectral fusion in dichotic release from upward spread of masking. *Haskins Laboratories Status Report on Speech Research*, **SR-39/40**, 81-89.

Nusbaum, H. C., Schwab, E. C., & Sawusch, J. R. (1983). The role of "chirp" identification in duplex perception. *Perception & Psychophysics*, **33**, 323-332.

Nye, P. W., Nearey, T. M., & Rand, T. C. (1974). Dichotic release from masking: Further results from studies with synthetic speech stimuli. *Haskins Laboratories Status Report on Speech Research*, **SR-37/38**, 123-137.

Pastore, R. E., Schmuckler, M. A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception & Psychophysics*, **33**, 469-474.

Pastore, R. E., Szczesiul, R., Rosenblum, L. D., & Schmuckler, M. A. (1982). When is a [p] a [t], and when is it not. *Journal of the Acoustical Society of America*, **72** (Supplement No. 1), S16. (Abstract)

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. *Perception & Psychophysics*, **33**, 333-337.

Schwab, E. C. (1981). *Auditory and phonetic processing for tone analogs of speech*. Unpublished doctoral dissertation, SUNY at Buffalo.

(Manuscript received August 24, 1984;
revision accepted for publication November 12, 1984.)