Cognition, 14 (1983) 211–235

Some differences between phonetic and auditory modes of perception

VIRGINIA A. MANN*

ALVIN M. LIBERMAN**

Haskins Laboratories, Inc.

Abstract

When third-formant transitions are appropriately incorporated into an acoustic syllable, they provide critical support for the phonetic percepts we call [d] and [g], but when presented in isolation they are perceived as time-varying 'chirps'. In the present experiment, both modes of perception were made available simultaneously by presenting the third-format transitions to one ear and the remainder of the acoustic syllable to the other. On the speech side of this duplex percept, where the transitions supported the perception of stop-vowel syllables, perception was categorical and influenced by the presence of a preposed [al] or [ar]. On the nonspeech side, where the same transitions were heard as 'chirps', perception was continuous and free of influence from the preposed syllables. As both differences occurred under conditions in which the acoustic input was constant, we should suppose that they reflect the different properties of auditory and phonetic modes of perception.

In the phonetic domain, the relation between acoustic cue and percept has several characteristics that have been taken to imply a special mode of processing (for recent reviews, see: Liberman, 1982; Liberman and Studdert-Kennedy, 1978; Repp, 1982; Studdert-Kennedy, 1980); but see, for example: Kuhl, 1981, Kuhl and Miller, 1975; Miller, 1977. One such characteristic is that frequency-modulated acoustic cues are integrated with other cues into unitary percepts that seemingly lack the qualities we might have been led, on

---

* Also Bryn Mawr College.

** Also University of Connecticut and Yale University.

The research described in this paper was supported by NICHD Grant HD-01994 and BRS Grant RR-05596 to Haskins Laboratories and by Bryn Mawr College. We wish to thank J. Michael Russell and James Madden for their assistance in scoring the data for this experiment and for their participation in pilot studies. Reprint requests should be sent to Virginia A. Mann, Haskins Laboratories, New Haven, CT 06510, U.S.A.

0010–0277/83/$8.00

purely psychoacoustic grounds, to expect. A case in point, and the one with which we will be concerned, is in the perception of the stop consonants [d] and [g]. As has long been known, sufficient cues for the perceived distinction between these phones are transitions—that is, frequency modulations—of the second or third formant. Thus, when appropriate transitions of the third formant—the cue that will be the subject of our investigation—are presented in an otherwise fixed acoustic context, listeners perceive a syllable consisting of [d] or [g], followed by a vowel. Of special interest to us is that one hears in these percepts none of the time-varying quality, a 'chirpiness', for example, or a glissando, that might be thought to correspond to the time-varying nature of the frequency-modulated signal. Indeed, one finds it difficult to characterize the [d] and [g] percepts, and especially the differences between them, in auditory terms of any kind. It is as if the percepts were as abstract as the phonetic segments they represent.

We might nevertheless account for the percepts without reference to specialized processes of a phonetic sort. Thus we might assume, most simply, a low-level process of sensory integration, similar, perhaps, to the integration of intensity and time into the perception of loudness. But such an assumption is ruled out by the finding that listeners do, in fact, hear the to-be-expected chirps and glissandi when the transition cues are removed from the larger context and sounded alone (Mattingly *et al.*, 1971). Still, we might save an auditory account by noting that the transitions are normally presented in a larger acoustic context, and that they are, therefore, subject to the effects of a purely auditory interaction with the remainder of the pattern. On that account, the peculiarly abstract character of the percept would be thought to emerge from the interaction. Nothing we know about auditory perception suggests the existence of such an interaction, but the possibility is not precluded.

There is, in any case, another characteristic of the way formant transitions function when they cue stop consonants: the phonetic percepts they support are appropriate to their role in language, not only in their abstractness, but also in the extent to which they are categorical. Given transitions that change in relatively small physical steps, from one appropriate for [d] to one appropriate for [g], the percept changes, not in correspondingly small steps, but suddenly (Liberman *et al.*, 1957; Mattingly *et al.*, 1971; Repp, in press; Studdert-Kennedy *et al.*, 1970). This nearly categorical shift marks a sharp boundary between the two phones [d] and [g]; it is commonly reflected and measured as a relative increase in discriminability of the stimuli at the category boundary. But such tendencies toward categorical perception do occur in nonspeech perception as well (see, for example: Burns and Ward, 1978; Locke and Kellar, 1973; Miller *et al.*, 1976; Parks *et al.*, 1969; Siegel and

Siegel, 1977), so the question is not whether it is unique to the perception of stop consonants (and other phonetic segments), but, more properly, whether the categorical boundary between the phonetic segments is of an auditory sort. We have reason to believe it is not, for when the same formant transitions are presented in isolation (and perceived as nonspeech chirps), the obtained discrimination function is continuous—that is, it does not display the abrupt peaks and troughs that typify categorical perception. This result has been obtained in adults (Mattingly *et al.*, 1971) and in infants (Eimas, 1974). It follows, then, that if the categorical effect in the full speech context is to be assigned a purely auditory cause, then, as in the previously noted case, it must be referred, *ad hoc*, to some assumed auditory interaction between the transitions and the remainder of the acoustic pattern.

A quite different characteristic of the way formant transitions cue [d] and [g] is that their effects are subject to the influences of phonetic context. Thus, given abutting vowels, the transition must, of course, move into or out of the vocalic nucleus; hence, the boundary between [d] and [g] will occur in transitions that are at different positions on the spectrum for different vocalic contexts (Delattre *et al.*, 1955; Liberman *et al.*, 1954). More relevant to our concern here, however, is the fact that, given a fixed continuum of formant transitions, a shift in the [d-g] boundary can be produced by neighboring consonants. Such effects have been found with preposed fricatives (Mann and Repp, 1981; Repp and Mann, 1981) and across a syllable boundary with preposed [al] or [ar] (Mann, 1980). In both cases, the shift in the position of the boundary was found to be consistent with the way the formant transitions for [d] and [g] are affected in normal speech by coarticulation with fricatives or with liquids. Therefore, the movement of the category boundary is most plausibly to be understood as a perceptual compensation for the effects of coarticulation. As such, it would presumably reflect a phonetic rather than an auditory process. To appeal, instead, to an auditory interaction would require not only that we set aside the coarticulatory facts, together with the reasonable interpretation based on them, but also that we make a seemingly unreasonable assumption about why speech perception finds parallels in speech production—to wit, that speakers adjust the behavior of their articulatory organs so as to produce in every context just those acoustic effects that will fit boundary shifts caused by pre-existing auditory interactions. Such an interpretation becomes, in the end, hopelessly *ad hoc* and, given what we know of constraints on articulation, quite implausible. But, again, it cannot, in principle, be ruled out.

To control for auditory interaction, we should contrive acoustic patterns that can, depending on specifiable circumstances, be perceived either as speech or as nonspeech. Two techniques are available for this purpose, and

both have been used in other studies to gain the control we seek. One employs stripped-down versions of synthetic speech that can be heard as speech or nonspeech, depending on the natural proclivities of the listeners, how long they have been listening, and just what has or has not been suggested to them (Best *et al.*, 1981; Remez *et al.*, 1981). The other method, and the one we will use, takes advantage of a phenomenon in which, with auditory input held constant, the acoustic cue of interest is perceived *simultaneously* as a nonspeech chirp and as critical support for a phonetic segment. This phenomenon, called 'duplex perception,' was first reported by Rand (1974). Recently, it has been further studied in an investigation of the cues for the liquids [l] and [ ] (Isenberg and Liberman, 1978), and it has been used to control for auditory interaction in a study of silence as a cue for stop consonants (Liberman *et al.*, 1981). Here, we will exploit it to provide an appropriate control for auditory interaction in investigations of the third-formant transition as a cue for the perceived distinction between [d] and [g]. In the first of these, we will be concerned to find out whether the integration of such transitions into unitary phonetic categories is to be attributed to processes of a generally auditory sort, or whether it is the result of processes that are distinctively phonetic. The second part of our study is designed to determine if context-conditioned movement of the boundary between the [d] and [g] categories is also to be regarded as a special attribute of phonetic perception.

Experiment I

Our aim in the first experiment was to measure discriminability of third-formant transitions on both sides of a duplex percept—that is, when, on the 'speech' side, the transitions provide crucial support for the perceived difference between [da] and [ga], and when, on the 'nonspeech' side, they are heard as unspeechlike 'chirps'. The stimulus patterns were three-formant synthetic syllables in which the third formant varied in nine steps, from a setting appropriate for [da] to one appropriate for [ga].

To produce duplex perception of these third-formant transitions, we separated them from the (fixed) remainder of the pattern—which we will, for convenience, call the 'base'—and presented the separated constituents dichotically. Thus, the transitions, which in isolation sound like chirps, and the base, which in isolation sounds like a syllable (most commonly, [da]), are free to mix and hence to interact in the listener's nervous system. The usual result is two percepts, present simultaneously. On one side of this duplexity is a syllable, [da] or [ga], which is perceptibly different from the base but

very similar, perhaps identical, to what is heard when the two constituents (transition and base) are mixed electronically and presented in the normal manner (Liberman *et al.*, 1981; Repp, 1982). On the other side is a nonspeech 'chirp' that seems identical to what is heard when the transition is presented in isolation.

Given systematic variation in the formant transitions, we can measure discriminability, hence tendencies toward categoricalness, of the resulting speech and nonspeech components of the duplex percept. To the extent that there is categorical discrimination of the formant transitions heard on the speech side of the duplex percept, the discrimination function should have marked peaks and troughs which accord with predictions derived from phonetic labeling responses (Liberman *et al.*, 1957). To the extent that the phonetic categories themselves have a purely auditory basis, the discrimination function for the same formant transitions when heard on the nonspeech side of the duplex percept should also have marked peaks and troughs and, like the function for discrimination of speech percepts, should meet with predictions derived from phonetic labeling.

Method

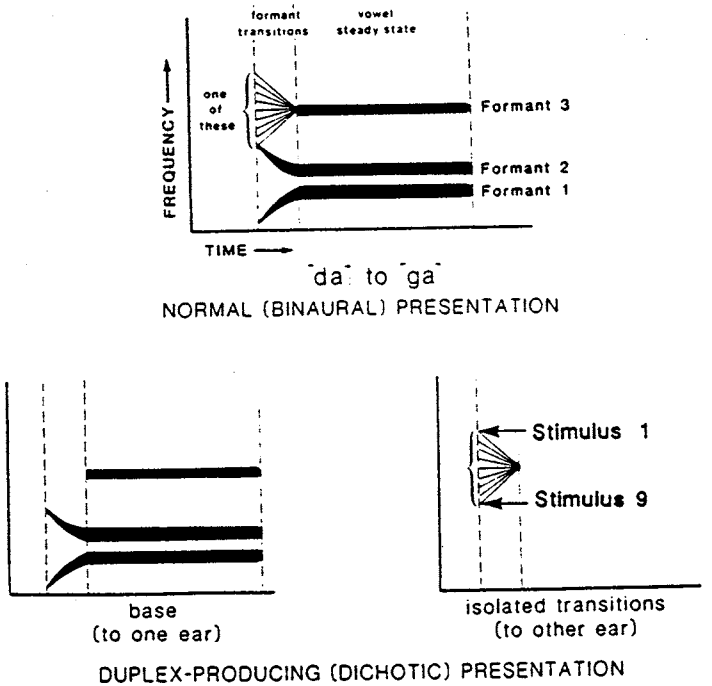
Materials

Stimulus continuum

At the top of Figure 1 is a schematic representation of the stimulus patterns. These patterns, very similar to those used by Mann (1980) in the study referred to in the Introduction, were designed to be synthetic approximations to the syllables [da] and [ga]. They were produced on the parallel resonance synthesizer at Haskins Laboratories. The lower half of Figure 1 shows how the stimuli were divided into the two constituents—the fixed 'base' and the variable 'isolated transitions'—that will, when presented dichotically, produce the duplex percept. The base is 250 msec in total duration, with a 50-msec ramp in overall intensity at onset and offset, and a fundamental frequency that falls linearly from 110 to 80 Hz. The first- and second-formant transitions are 50 msec in duration and step-wise linear in 5-msec steps; they begin at 279 and 1764 Hz, arriving finally at steady-state values of 765 and 1230 Hz, with bandwidths of 60 and 80 Hz, respectively. The third formant of the base begins 50 msec later than the others and maintains a steady state at 2527 Hz with a bandwidth of 120 Hz. In accordance with natural speech, this third formant is slightly less intense than the other two.

The continuum of nine formant transitions was synthesized separately from

Figure 1. *Schematic representation of the patterns used to produce the duplex percepts, including the constant base portion and the continuum of nine formant transitions.*



the base. Each transition is 50 msec in duration and step-wise linear in 5-msec steps; fundamental frequency and amplitude contour are as in the first 50 msec of the base stimulus, the offset frequency is the steady-state third-formant frequency of the base, and the bandwidth is 120 Hz. Onset frequency systematically varies across the continuum in eight equal steps, from 3196 Hz in Stimulus 1 to 1853 Hz in Stimulus 9. As can be seen in Figure 1, the first four transitions have falling slopes, the fifth is flat, and the final four are rising. The slopes of the four rising transitions are equal in value to the slopes of the transitions that fall. For convenience, we will refer to the transitions hereafter by number, as shown in Figure 1, from most falling to most rising.

Test tapes

The base stimulus and the continuum of transitions were digitized at 10,000 Hz prior to being recorded onto magnetic tape for the purpose of testing. As

was appropriate for dichotic presentation (and duplex perception), the base was recorded onto one track, the isolated transitions onto the other.

A (duplex perception) labeling tape was constructed for use in the initial screening of subjects and for determining how the subjects identified the stimuli. This tape comprised a practice sequence consisting of four repetitions of the base in conjunction with each of the two endpoint transitions, followed by a test sequence with four sets of 27 stimuli each. Across these sets, the nine transitions occurred twelve times each in a randomized order. The inter-stimulus interval was 3 sec; the inter-set interval was 6 sec.

Our measure of discrimination performance was obtained by the method known as AXB. (A and B are the two stimuli to be discriminated; X is one or the other. The subject's task is to decide if X is less like A or less like B.) We chose to present stimuli at three-step intervals along the continuum of formant transitions, because pilot work (Mann *et al.*, 1981) had suggested that for most subjects a separation of that size puts discrimination of the chirps and the speech in a sensitive region—that is, it keeps discrimination from falling to the floor or rising to the ceiling. This step size also provided a sensitive measure of the context-induced shifts in phonetic category boundary which were to be the concern of our second experiment.

The duplex-perception discrimination tape consisted, then, of sets of stimulus triads, one practice set and six test sets. Each such set contained randomized sequences of the six possible three-step combinations of stimuli along the continuum (i.e., by stimulus number; 1 *versus* 4, 2 *versus* 5, 3 *versus* 6, 4 *versus* 7, 5 *versus* 8, and 6 *versus* 9), occurring once each in AAB, ABB, BAA, and BBA triads. Thus, over the course of the test sets, listeners responded to a total of 24 triads for each pair. Within triads, the inter-stimulus interval was 500 msec, the inter-triad interval was 3 sec, and the inter-set interval was 6 sec.

An additional AXB discrimination tape was constructed to be used in pretest screening of the subjects, since pilot work (Mann *et al.*, 1981) had suggested that some subjects encounter specific difficulty in discriminating isolated chirps at three-step intervals along the continuum, and that such subjects also fail to discriminate chirp components of the duplex percept. This same tape served the further purpose of providing a basis for comparison with the nonspeech side of the duplex percept. The stimulus arrangement was analogous to that of the duplex-perception discrimination tape, save that there was no base stimulus for presentation to the other ear, and different randomizations determined the order of triads within each set.

Procedure

Subjects in an initial pool of 14 were pretested in groups of three or four while seated in a quiet room as the stimuli were played over earphones. For convenience, the third-formant transitions were always presented to the right ear and the base stimulus to the left. The purpose of the first pretest was to see if the subjects could discriminate the transitions when they are presented in isolation. To that end, subjects listened to the discrimination tape that contained the isolated transitions and were instructed to respond 'A' or 'B' according to whether the first or the third stimulus of each triad was less like the other two. Completion of the practice and test sets of item triads was followed by a second pretest. This served two purposes. First, it was a screening device by which we could determine whether subjects were consistent in their labeling of the endpoint stimuli of the duplex [da]-[ga] continuum. While the vast majority of subjects give consistent responses to the end points of our continuum when the base and third-formant stimuli are electronically fused, some subjects tend to give inconsistent responses when base and transition are dichotically presented, and we wished to exclude such subjects from our study. The second purpose served by the pretest was to provide a full identification function by which to determine, for those subjects in the main experiment, the extent to which discrimination on the speech side of the duplex percept is categorical. Both purposes of the second pretest were accomplished by having the subjects listen to the practice and test sequences of the duplex labeling tape and respond 'd' or 'g' as appropriate.

The subjects who survived the pretest participated in experiments that provided the results we will present. These experiments were divided into two sessions, one week apart and counterbalanced in order across subjects. In the test sessions, as in the pretest, the third-formant transitions were always presented to the right ear and the base stimulus to the left. In one session, subjects were instructed that the goal was to determine how well speech sounds could be discriminated in the face of some nonspeech distractors. They then listened to the practice and test sets of the duplex-perception AXB discrimination tape, responding on the basis of the perceived similarity in the speech percepts of each stimulus triad. In the other session, the subjects were instructed that the goal was to determine how well nonspeech sounds could be discriminated in the face of speech sounds as distractors. At this time, they also listened to the practice and test sets of the duplex AXB discrimination tape, but responded on the basis of the perceived similarity among chirp percepts. Subjects listened to the same tape in the two sessions, but were kept in ignorance of this fact. They were instructed to listen to the target speech sounds or chirps, according to the session, and to ignore the

'distractor' on the ground that attention to it could only impair their performance on the assigned task.

Subjects

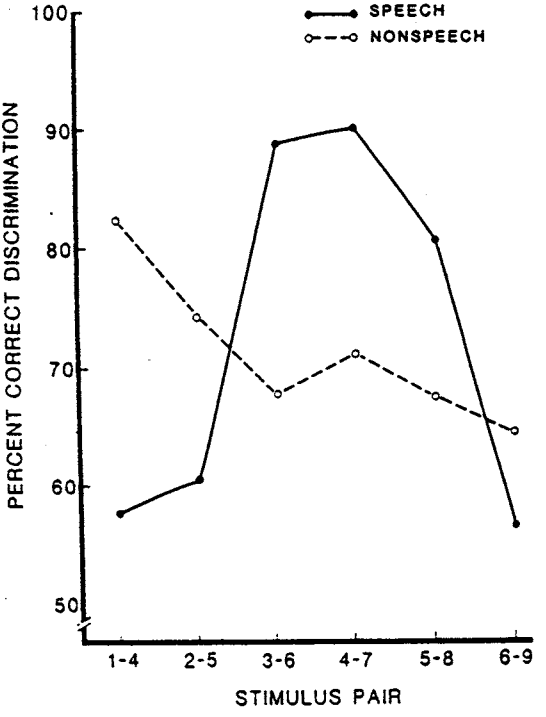
The subjects were paid student volunteers recruited from an introductory psychology course. All were female, and none had extensive experience in listening to synthetic speech. Of an initial pool of fourteen, six subjects were judged on the basis of the pretests to be insufficiently consistent in their responses and were therefore excluded from the experiment proper, two for having been unable to discriminate the isolated transitions at a level above chance, and four for having been inconsistent in the way they labeled the endpoints of the duplex continuum as 'd' (stimulus one) and 'g' (stimulus nine). Thus the final subject group included a total of eight subjects who participated in each of two sessions.

Results

We should first report the phenomenological results of the experiment, which were clear. Given the variable third-formant transitions in one ear and the remaining, fixed part of the acoustic pattern (the base) in the other, the subjects did report duplex percepts: a syllable, [da] or [ga], depending on the transition, and a nonspeech 'chirp'. The chirps on the nonspeech side of the duplexity had a time-varying quality corresponding, apparently, to the time-varying nature of the formant transitions. This is to say, they were not noticeably different from what the subjects perceived when the transitions were presented in isolation. On the speech side, the syllable [da] or [ga] lacked the 'chirpiness' that characterized perception on the nonspeech side, and they were not different from what listeners perceive when transitions and base are mixed electronically and presented in the normal manner. The base, which sounded like [da], was *not* perceived. That is, when the transition was appropriate for [ga], listeners typically perceived [ga], not [ga] and also (or half the time) [da]. Thus, perception was *duplex* not triplex: listeners perceived only speech (the fusion of base and transitions) and nonspeech (the transitions as if in isolation).

Beyond these observations, the data (averaged across the eight subjects) consist of discrimination functions for the speech and chirp components of the duplex percept (Figure 2); a labeling function for the speech component of the duplex percept (Figure 3A), together with the discrimination function (Figure 3B) that is predicted from it on the assumption of categorical percep-

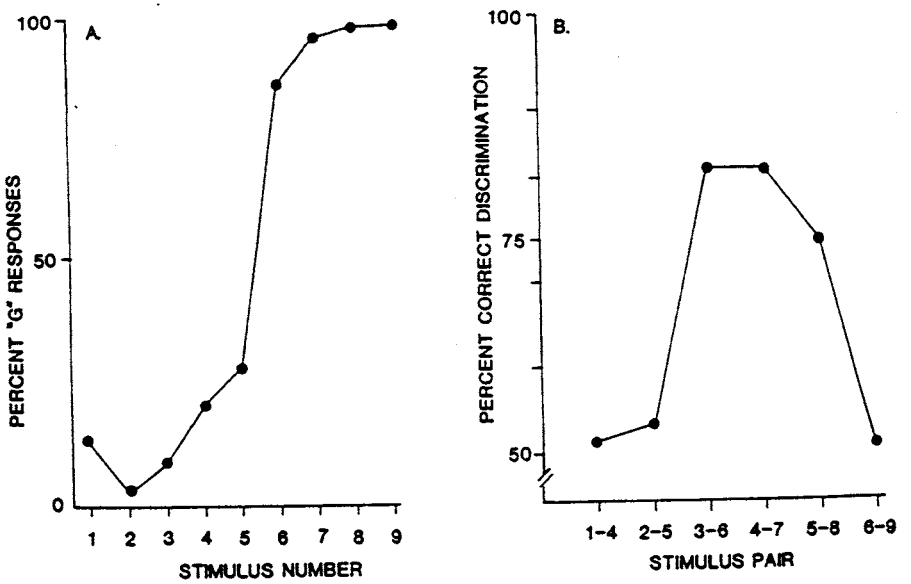
Figure 2. *Discrimination of the third-formant transitions on the speech and nonspeech sides of the duplex percept.*



tion (Liberman *et al.*, 1957); and a discrimination function for chirps presented in isolation (Figure 4). Consider, first, Figure 2, which compares discrimination of the duplex percepts under instructions to concentrate on speech (solid line) with that under instructions to concentrate on chirps (dashed line). Note that, while the overall level of performance on the two tasks is roughly comparable, the shapes of the two functions differ markedly. This is verified statistically by a significant interaction between the nature of the attended percept and the stimulus pair being discriminated: $F(5,35) = 13.9$, $p < 0.001$.

The overall shape of the speech function—its marked peaks and troughs—is consistent with categorical perception. To see how consistent, however, we must compare the speech-discrimination function that was obtained with the one that is predicted on the assumption of perfectly categorical perception. Plainly, the predicted discrimination function, which is in Figure 3B, is quite

Figure 3. (A) *Labeling of speech percepts as* [d] *or* [g]. (B) *Discrimination function predicted from labeling responses, given the assumption of categorical perception.*

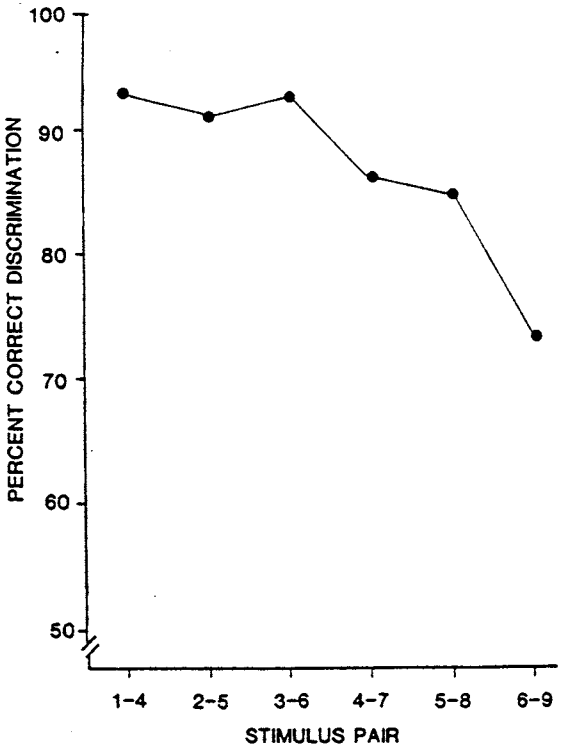


similar to the one we obtained. We conclude, therefore, that when the third-formant transitions were integrated into a phonetic percept, where they provided critical support for the distinction between [da] and [ga], they were perceived quite categorically.

In contrast to the way the transitions were discriminated on the speech side of the duplex percept is the discrimination function obtained with the same transitions on the nonspeech side, where they were perceived as chirps. As shown in Figure 2, the 'chirp' function has no marked peaks or troughs and is similar in shape to the function obtained with isolated transitions in Figure 4, although the absolute level is lower: $F(1,7) = 7.3$, $p < 0.05$. The initial pair of rising chirps (Pair 1-4) is significantly more discriminable than the final pair of falling chirps (Pair 6-9), both for isolated chirps, $t(14) = 4{,}37$, $p < 0.005$ and for the chirp components of the duplex percept, $t(14) = 2.6$, $p < 0.02$.

As noted by Mattingly *et al.* (1971), there are at least two strategies that listeners might use in discriminating the isolated transitions: they could, in effect, judge their slopes or, alternatively, their most apparent pitches. If our

Figure 4. *Discrimination of isolated third-formant transitions.*



subjects had opted for the first strategy, as the subjects in the Mattingly *et al.* study appear to have done, then discrimination would have been best for the transitions that straddle the horizontal transition (Transition 5). But that was not the result. Rather, discrimination became poorer as the transitions changed progressively from most falling to most rising. That result leads us to take into account an observation by Brady *et al.* (1961), who noted that the most apparent pitch of frequency ramps, which resemble isolated transitions, is closer to the frequency of their offsets than their onsets. They also observed, however, that this effect is stronger for rising ramps than for falling ones. Since our transitions have variable onset frequencies but the same offset, we should suppose that if, as in the study by Brady *et al.*, the tendency to judge pitch by the offset increased as the transitions changed from falling to rising, then we should have obtained the decrease in discrimination that our results do, in fact, show. We are inclined to conclude, therefore, that our

subjects were, to a considerable extent, discriminating the transitions on the basis of their most apparent pitches.

Though the overall level of discrimination for the two sides of the duplex percept was roughly equal, as noted earlier, discrimination of the transitions on the speech side was, in its most sensitive region, better than discrimination of the transitions on the nonspeech side. But, surely, we do not therefore conclude that speech discrimination exceeds the resolving power of the system, only that we have no idea how the resolving power is to be measured. Beyond this truism, two observations are pertinent. One is that, as can be seen by comparing Figures 2 and 4, the general level of nonspeech discrimination obtained when the transitions were presented outside the duplex context was somewhat higher than when they were perceived inside it. Perhaps this should be attributed to distractions provided by the circumstance that, in the duplex case, the two percepts, speech and nonspeech, were present at the same time. The other observation is that we should not, in any case, rule out the possibility that the human listener is, in fact, more sensitive to the formant transitions when they support a phonetic percept than when they do not. Indeed, Bentin and Mann (Reference Note 1) have evidence that, in the matter of absolute threshold sensitivity, the speech context does provide the more sensitive measure—that is, the closer approximation to the physiological limit—and for interesting reasons.

In summary, the difference between the two sides of the duplex percept is very great indeed. On the nonspeech side, the formant transitions evoke a percept that has the time-varying, chirpy quality that psychoacoustic considerations should have led us to expect, and the discrimination function is continuous. On the speech side, where the same formant transitions provide critical support for the stops in the syllables [da] and [ga], there is no apparent chirpiness in the percepts, and discrimination is nearly categorical.

Experiment II

The second experiment draws on the fact, noted in the Introduction, that the category boundary along a synthetic [da]-[ga] continuum in which the third-formant onset provides the sufficient cue, can be systematically shifted by the presence of a preposed [al] or [ar] (Mann, 1980). For stimuli preceded by [al], the category boundary shifts towards a higher third-formant onset (more 'g' responses), whereas a preceding [ar] causes a shift in the opposite direction. Both perceptual shifts are consistent with observations about the acoustic consequences of articulatory accommodation to the new contexts: stop consonants that are coarticulated with a preceding liquid apparently assimi-

late toward the place of liquid articulation. That is, stops preceded by [al] tend to contain a higher third-formant onset frequency than those preceded by [ar], suggesting that they receive a more forward place of articulation. On that basis, Mann (1980) supposed that the perceptual context effect of the (preposed) liquids reflects the application to perception of some tacit knowledge about speech production. This in turn implies the existence of some specialized phonetic process.

But, as we pointed out in the Introduction, the possibility of auditory interaction exists, at least in principle. To control for such interaction, we will again take advantage of duplex perception. That will be done by putting the syllables [al] and [ar] in front of the 'base' of the dichotically presented (and duplexly perceived) [da]-[ga] stimuli of Experiment I. We can find out then whether the preposed [al] and [ar] affect perception of the formant transitions on both sides of the duplex percept or, as we suspect, only when they are perceived as speech.

Method

Materials

Stimulus continua

Two continua of disyllables were constructed by putting in front of the synthetic stimuli of Experiment I naturally produced syllables whose fundamental frequency and formant structure approximated those of the synthetic stimuli and thus permitted the disyllable to be perceived as a coherent utterance produced by one and the same vocal tract. An [al-da] to [al-ga] continuum was formed in this way, using the base stimulus from Experiment I and a token of [al] that had been excised from an utterance of [al-da] produced by a male native speaker of English. An [ar-da] to [ar-ga] continuum was constructed by putting in front of the base a token of [ar] excised from an utterance of [ar-da] produced by the same speaker. In each case, a 100-msec silent gap separated the offset of the natural syllable from the onset of the synthetic one. The continuum of formant transitions that cued the [d]–[g] distinction was as in Experiment I.

Test tapes

All stimuli were digitized at 10,000 Hz prior to being recorded onto magnetic tape for the purpose of testing. The arrangements of the stimuli on the magnetic tape was as in Experiment I, except, of course, that the 'base' was preceded by [al] or [ar].

To determine how the subjects would identify the stimuli, and thus provide a basis for predicting what perfectly categorical discrimination functions should look like, we made a dichotic 'labeling' tape, appropriate for duplex perception. It consisted of a practice sequence containing four repetitions of each endpoint transition paired with [al] plus base, and four repetitions of each endpoint transition paired with [ar] plus base, followed by a test sequence containing eight sets of 27 stimuli each. Over the test sets, each of the nine transitions occurred, in random order, a total of twelve times in conjunction with each preposed syllable.

To test discrimination by the method of AXB, another dichotic tape was prepared in which the stimuli were recorded in triads, exactly as in Experiment I, except that the base stimulus in half the triads was preceded by [al] and in half by [ar]. Which syllable ([al] or [ar]) preceded the base was randomized from trial to trial. For both [al] and [ar] conditions, the six pairs of to-be-discriminated transitions were equally represented across the triads, as were the various orders of transitions within each pair. As in Experiment I, listeners gave a total of 24 responses to each pair of transitions as preceded by each of the two syllables.

Procedure

Experiment II was run in two experimental sessions that also included Experiment I. Thus, in one session—the session in which the instruction was to attend to speech percepts—the subjects first heard the labeling tape and then the discrimination tapes for the two experiments. Order was counterbalanced. In the other session, where the instruction was to attend to chirp percepts, they also listened to the two discrimination tapes. Here, too, order was counterbalanced.

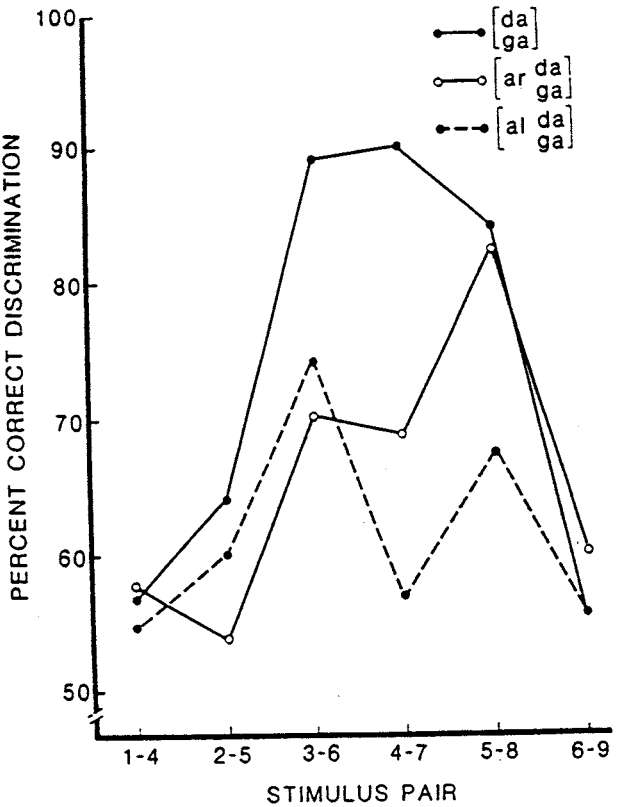
Subjects

The subjects were the same eight young women who participated in Experiment I.

Results

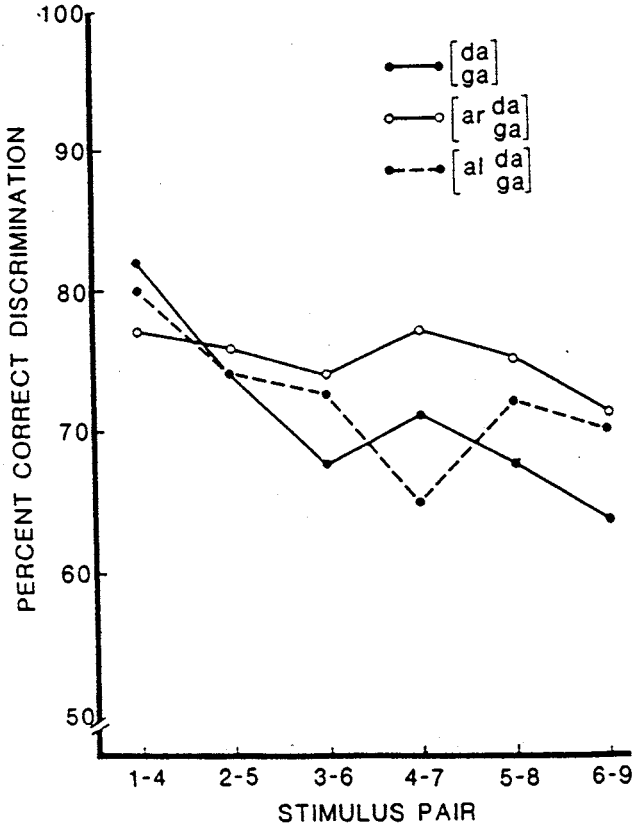
The point of this experiment, it will be remembered, was to test the effects of a preposed [al] or [ar] on the perception of third-formant transitions when, in the one case, they are integrated into a speech percept and when, in the other, they are perceived as nonspeech chirps. To display those effects, we

Figure 5. *The influence of preposed syllables,* [al] *and* [ar], *on discrimination of the transitions on the speech side of the duplex percept. The analogous function obtained without preposed syllables (Experiment I) is reproduced for purposes of comparison.*



have, in Figures 5 and 6, combined the results of Experiments I and II. Discrimination functions for the speech side of the duplex percepts are in Figure 5 and those for the nonspeech side in Figure 6. A glance at these two figures reveals our main finding: context had a strong effect on discrimination of the transitions on the speech side of the percept but not on the nonspeech side. Looking more closely at the speech side in Figure 5, we see that the peak in the function for [da]–[ga] syllables preceded by [ar] (solid lines and open circles) is shifted to the right of that obtained in Experiment I, where there was no preposed [ar] (solid lines, closed circles). On the assumption

Figure 6. *The influence of preposed syllables,* [al] *and* [ar], *on discrimination of the transitions on the nonspeech side of the duplex percept. The analogous function obtained without preposed syllables (Experiment I) is reproduced for purposes of comparison.*



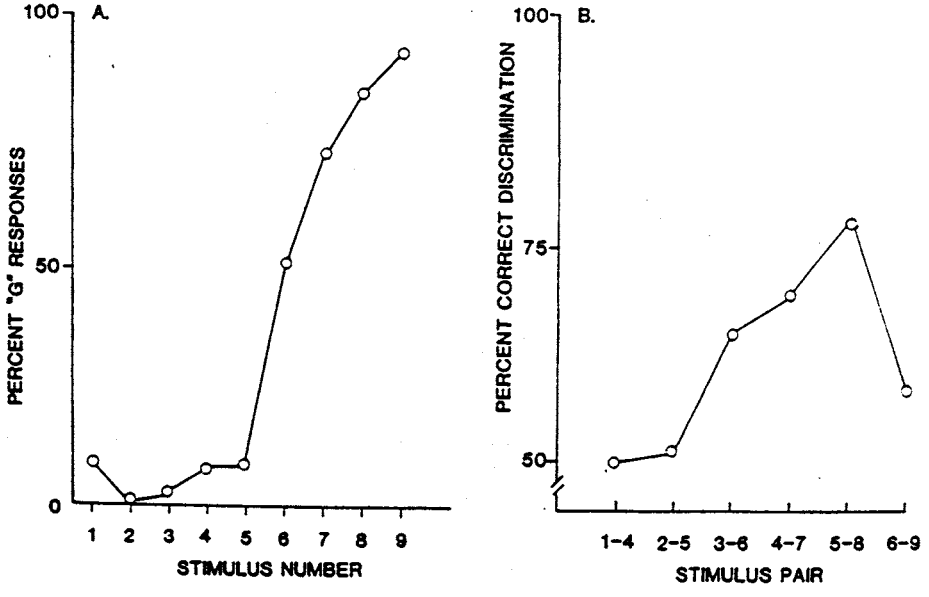
that the location of the discrimination peak reflects the location of the phonetic boundary, an assumption we will justify later, the direction of the shift in the peak is consistent with the earlier results of Mann (1980). Those same earlier results led us to expect a shift in the opposite direction when [al] is preposed. As can be seen in the function described by the dashed lines (filled circles), the nature of the shift due to [al] is somewhat less clear. Possible reasons for this will be discussed later. For the moment, however, the point to be made is that the speech function obtained in this context is, in any case, different from both of the other two.

In contrast to the results obtained on the speech side, the functions of Figure 6 indicate that preposed [al] and [ar] had no effect on discrimination of the transitions when they were perceived, on the nonspeech side, as chirps.

To support the assertions of the preceding paragraphs, we offer the results of a three-way analysis of variance, conducted with the following factors: attended percept (speech or chirps); context (isolated duplex stimuli, stimuli preceded by [al], or stimuli preceded by [ar]); and stimulus pair. Although there was no significant effect of attended percept, suggesting that the average level of performance in our experiments was equivalent for speech and chirps, there was an effect of context: $F(2,14) = 5.38$, $p < 0.025$, and an effect of stimulus pair: $F(5,35) = 5.83$, $p < 0.001$. Most important to our observations about the special influence of context on speech perception are the interactions among the three main factors. First, there was an interaction between attended percept and stimulus pair, revealing that the relative difficulty of discriminating individual pairs depended on whether the instruction was to attend to speech or to the chirps: $F(5,35) = 13.18$, $p < 0.001$. Second, there was an interaction between attended percept and context, revealing that the effect of context was greater for speech percepts than for the chirps: $F(2,14) = 11.59$, $p < 0.001$. Finally, there was an interaction of context and stimulus pair: $F(10,70) = 2.46$, $p < 0.025$, and a three-way interaction: $F(10,70) = 2{,}00$, $p < 0.05$. Separate analyses of variance for the two percepts reveal that, in the case of the speech percepts, the preceding syllables influenced both the level: ($F(2,14) = 12.35$, $p < 0.001$), and also the pattern of speech discrimination across stimulus pairs: ($F(10,70) = 3{,}17$, $p < 0{,}005$). For the nonspeech chirps, on the other hand, an analysis of variance indicates that the preposed syllables had no significant effect on either the level or the pattern of performance.

Having seen that the discrimination functions reflect an effect of context on the speech side of the duplex percept, we should now consider the extent to which those functions are predicted from the phonetic labeling results, given the assumption of categorical perception. Consider, first, the results obtained for stimuli preceded by [ar], as shown in Figure 7A. We see that the [da]–[ga] boundary occurs somewhere between Stimulus 5 and Stimulus 6. Comparison with the boundary obtained for the isolated [da]–[ga] stimuli of Experiment I (Figure 3) shows that, as in the earlier experiment by Mann (1980), the [ar] context moved the boundary toward the [ga] end of the stimulus continuum, thus increasing the number of [da] responses. On the assumption of completely categorical peception (Liberman *et al.*, 1957), we should have expected to obtain the discrimination function shown in Figure 7B. In fact, the discrimination function we did obtain (solid lines and open circles of Figure 5) is quite similar to the expected one. Certainly, the peak

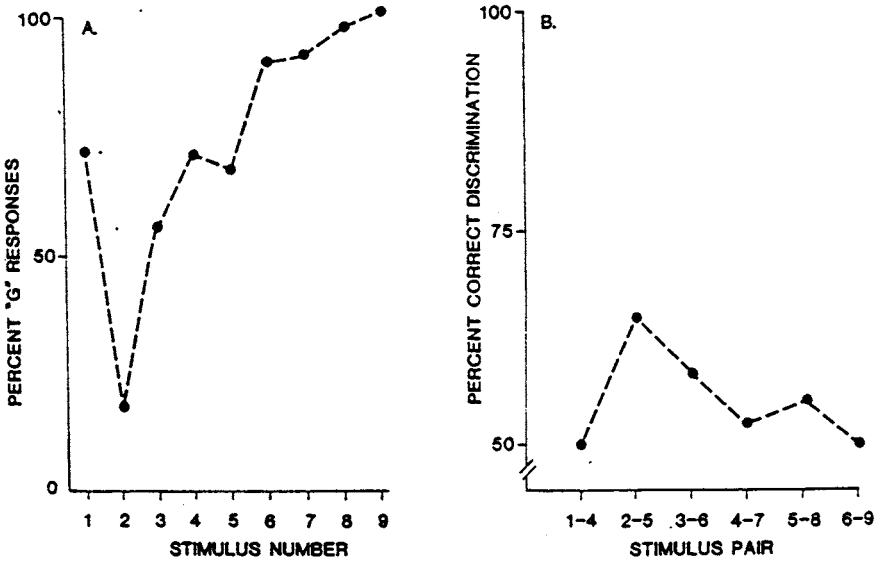
Figure 7. (A) *The influence of a preposed* [ar] *on labeling of speech percepts as* [d] *or* [g]. (B) *Corresponding predicted discrimination function, given the assumption of categorical perception.*



is in the right place and only slightly higher (as it so often is in such situations) than it should have been. Thus, the obtained discrimination function does reflect the phonetic boundary; moreover, it can be seen, by comparison with the result for the isolated syllables, to reflect the context-conditioned shift in that boundary caused by the preposed [ar].

As for the labeling function obtained with the preposed [al], seen in Figure 8A, we note, first, a large inversion in the responses to Stimulus 1. Putting that aside for the moment, we see that, by comparison with the labeling data for the isolated syllables (Figure 3), the [da]–[ga] boundary with preposed [al] is shifted strongly toward [da], producing, thus, an increase in the number of [ga] responses. This, too, is consistent with the earlier finding by Mann. However, the most extreme falling transition of her earlier study did not evoke the large number of [ga] responses that its counterpart (Stimulus 1) did in the present one. Of course, the conditions of the two experiments were not identical. In the present experiment, but not in the earlier one, the judgments were made on the speech side of a duplex percept. Another difference between the experiments, and a second likely cause of the difference in result,

Figure 8. (A) *The influence of a preposed* [al] *on labeling of speech percepts as* [d] *or* [g]. (B) *Corresponding predicted discrimination function, given the assumption of categorical perception.*



is that the stimuli were not exactly the same. Perhaps, then, the most extreme falling transition of this experiment went beyond the limit for [da]. At all events, we should note that in the other two labeling functions obtained in this experiment ([da]–[ga] in isolation, as in Figure 3, and [da]–[ga] with [ar] preposed, as in Figure 7) there is also a tendency for the responses to the extreme falling transition of Stimulus 1 to show some inversion toward [ga]. Perhaps the inversion in the [al] context is simply an exaggeration of that tendency, and, as such, a further reflection of the strong bias toward [ga] produced by the preposed [al].

In any case, the labeling results for the [al] context yield the predicted discrimination function seen in Figure 8B. There is only a low peak, but its position reflects a shift in the phonetic boundary opposite to that which was produced by the preposed [ar]. Looking now at the obtained discrimination function in Figure 5, we see a moderately good fit to the one that was predicted. We conclude, then, that in the [al] context, as in the [ar] context, the discrimination function reasonably reflects the phonetic boundary and the effect that context has on it.

In striking contrast to the effects of phonetic context on the speech side of

the duplex percept is the absence of such effects on the nonspeech side. As shown in Figure 6, and as previously noted, the discrimination functions for the transitions perceived as chirps are much the same when [ar] or [al] is preposed as when, in Experiment I, they were not. Moreover, the shape of the functions reflects perception that is more nearly continuous than categorical. The slopes indicate that, as in the case of the isolated patterns of Experiment I, discrimination of falling transitions *versus* less falling ones was, other things equal, better than rising *versus* less rising: $t(14) = 2.75$, $p < 0.02$ for stimuli preceded by [al], and $t(14) = 2.7$, $p < 0.02$ for those preceded by [ar].

Discussion

Our concern has been to account for two effects previously observed in the perception of formant transitions as cues for stop consonants: tendencies toward categorical perception and shifts in the positions of category boundaries with phonetic context. Categorical perception, which we will consider first, has two manifestations, at least in the case of speech perception. The one, and the one to which attention has hitherto been directed almost exclusively, is the discontinuity in perception that defines a boundary on some physical continuum. The other is in the phenomenal nature of the perceived category, which is more appropriate to a linguistic object than to an auditory one (Liberman, 1982). In speech perception, these two manifestations presumably reflect the same underlying process, but they are separable, at least in principle, and we should take a moment to say how.

Given that the formant transitions are modulations in frequency, they might be perceived, correspondingly, as modulations in pitch. If so, perception could be nonetheless categorical. Thus, given a continuum of transitions, the listener might perceive them discontinuously—for example, as rising or falling pitches. Such automatic sorting of auditory percepts would, of course, be of use to listeners since it would relieve them of having deliberately to make the categorical assignments that the phonetic and phonological structure of the language require. But if, as in this example, perception of the transition cues, and all the other cues for the same phone, retained their auditory character, then perception of speech would be like perception of Morse code or some other arbitrary acoustic cipher. In that case, a listener would perceive rising or falling pitches, together with the auditory correlates of the many other acoustic cues, and have then to 'interpret' the resulting melange as a unitary phone. Presumably, the process of interpretation would, in time, become automatic, as, indeed, it does with people skilled at Morse, but the purely auditory character of the percept would continue to intrude.

This would be the more distressing because the auditory percept has little or nothing to do with the linguistic function of the phonetic unit it conveys.

To draw an analogy from visual perception of depth, consider how confusing it would be if, in the use of the retinal disparity cue, we were aware, not just of the distal depth, but also of the proximal disparity (doubling of images) which provided the relevant information. Fortunately, processing is accomplished in this case by a specialized module that uses the proximal disparity to yield, in consciousness, only perception of the distal depth relationships among visual surfaces.

We would argue, then, that a similar module operates in speech perception to yield, in consciousness, only the distal phonetic object, free of the chirps or glissandos we would otherwise hear. This would, as we have indicated, be especially appropriate for the purposes of language, given that everything that we need to know about a stop consonant, for example, has been provided when any particular token has been identified as this stop consonant and not that one. In that sense, a stop consonant represents nothing but the categorical and abstract segment the speaker intended. Hence, awareness of the auditory attributes of its various acoustic cues would, like awareness of proximal retinal disparity, be irrelevant at best, and, at worst, seriously distracting.

As pointed out in the Introduction, listeners are, indeed, quite aware of the auditory attributes of the transitions when they are presented in isolation, in which case they sound like chirps, but not when, as part of a larger acoustic pattern, they support perception of stop consonants. This difference, as was also pointed out, occurs in conjunction with a difference in categorical perception in the more usual sense: discrimination of the transitions is continuous or categorical, depending on whether they are perceived in isolation, as chirps, or, together with the rest of the acoustic pattern, as stop-vowel syllables. As we have indicated, we find it plausible to suppose that incorporation of the transitions into stop percepts, and, in particular, the contrast this presents to their perception as chirps, reflects a specialized phonetic process, well-adapted to providing just the abstract categories the larger language system uses. But it is at least conceivable, if implausible, that ordinary auditory perception is at work—that in this, and in all the many similar cases where there exist parallels between speech perception and speech production, the articulators are so controlled as to produce exactly those combinations of cues that fit into independently existing interactions of an auditory sort.

The second effect that concerns us, namely, that the positions of the category boundaries shift with phonetic context, has been taken as a reflection of the context-conditioned variation in the acoustic signal that results from the way it is produced. Specifically, the variations in the signal are the con-

sequence of the coarticulatory arrangements that make it possible for speakers to fold phonetic segments into larger units—syllables, for example—and thus produce the segments much faster than they otherwise could. (To do otherwise, in this case, would entail making each segment a syllable—that is, to spell.) But listening to speech would be awkward if all the auditory consequences of these context-conditioned variations were prominent in consciousness. Given, in the cases we are concerned with, that the perceptual compensation is made automatically—that is, that the category boundaries shift appropriately—we assume that in this instance, too, we are seeing the effect of a highly adaptive and distinctively phonetic process. But, again, one might suppose, however implausibly, that the effect is simply auditory—that in this, and in every other such case, coarticulation occurs, not to make it easier to speak, but only to accommodate the sounds of speech to the characteristics of the auditory system, and especially to auditory interactions.

The purpose of the experiments reported here was to exploit the phenomenon of duplex perception to provide data relevant to deciding between these phonetic and auditory interpretations of stop consonant categories and their movement with context. The results were quite clear. Given an isolated third-formant transition appropriate for the stop in [da] or [ga] to one ear, and the remainder of the acoustic syllable to the other, listeners perceived the transitions in two phenomenally different ways: as nonspeech chirps, just like those they perceived when the transitions were presented in isolation, and as critical support for the stops in syllables [da] and [ga], in which case the percept was just like the one that was evoked when the transitions were electronically mixed with the rest of the acoustic pattern and presented in the normal manner. The remainder of the acoustic syllable, which in isolation sounds like speech, was not also perceived, which is to say that the percept was duplex, not triplex. On the nonspeech side of the duplexity, the chirp percept conformed reasonably to what psychoacoustic considerations might have led us to expect. Moreover, perception of these chirps was continuous, and there was no measurable effect of phonetic context. On the speech side, there was a phonetic percept—a stop consonant—not readily describable in auditory terms. In addition, perception was strongly categorical and the category boundary moved in expected ways as a function of phonetic context.

We should emphasize that the two classes of percept were evoked by transitions that were always paired, albeit in the other ear, with the remainder of the acoustic syllable. Thus, the two constituents of the dichotically presented pair, having been mixed in the nervous system, were free to interact or not. If, in that circumstance, we were to attribute the results on the speech side of the percept to interactions of an auditory kind, what would we say then about the results on the other side? How would we, on such an auditory

account, explain why the dichotic constituents interact to produce a normal [da] or [ga], but also fail to interact, not for both constituents, but only for one—the isolated transitions? Why, that is, was there perception of the isolated transition as such, but no comparable 'isolated' perception of the stimulus to the other ear, the 'base' that, by itself, sounds like speech? To account for the fact that the percept was, in this way, only duplex, we should suppose that there are two modes of processing at work in the perception of the transitions, and that, happily from our point of view, the peculiar conditions of the dichotic presentation make the results of both modes available to consciousness. In the one mode, which is auditory, are the processes that underlie perception of the transitions as nonspeech chirps. In the other, which is phonetic, the transitions are incorporated into the speechlike pattern that was presented to the other ear, where they serve the singularly linguistic purpose of distinguishing the abstract categories [da] and [ga].

References

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

Brady, P.T., House, A.S., and Stevens, K.N. (1961) Perception of sounds characterized by a rapidly changing resonant frequency. *J. acoust. Soc. Amer., 33*. 1357–1362.

Burns, E.M., and Ward, W.D. (1978) Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *J. acoust. Soc. Amer., 63*, 456–468.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. *J. acoust. Soc. Amer., 27*, 769–773.

Eimas, P.D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. *Percep. Psychophys., 16*, 513–521.

Isenberg, D., and Liberman, A.M. (1978) Speech and non-speech percepts from the same sound. *J. acoust. Soc. Amer., 64*, Suppl. No. 1, S20. (Abstract)

Kuhl, P.K. (1981) Discrimination of speech by nonhuman animals: Basic auditory sensitivities conductive to the perception of speech-sound categories. *J. acoust. Soc. Amer., 70*, 340–349.

Kuhl, P.K., and Miller, J.D. (1975) Speech perception by the chinchilla: Vocied–voiceless distinction in alveolar plosive consonants. *Sci. 190*, 69–72.

Liberman, A.M. (1982) On finding that speech is special. *Amer. Psychol. 37*, 148–167.

Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Mono. 68*, 1–13.

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957) The discrimination of speech sounds within and across phoneme boundaries. *J. Exper. Psychol. 53*, 358–368.

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

Liberman, A.M., and Studdert-Kennedy, M. (1978) Phonetic perception. In R. Held, H.W. Leibowitz, and H.-L. Teuber (Eds.), *Handbook of Sensory Physiology, Vol. VIII: Perception*. New York, Springer-Verlag, 143–178.

Locke, S., and Kellar, L. (1973) Categorical perception in a non-linguistic mode. *Cortex, 9*, 353–367.

Mann, V.A. (1980) Influence of preceding liquid on stop-consonant perception. *Percep. Psychophys. 28*, 407–412.

Mann, V.A., Madden, J., Russell, J.M., and Liberman, A.M. (1981) Further investigation into the influence of preceding liquids on stop consonant perception. *J. acoust. Soc. Amer., 69*, S91. (Abstract)

Mann, V.A., and Repp, B.H. (1981) Influence of preceding fricative on stop consonant perception. *J. acoust. Soc. Amer., 69*, 548–558.

Mattingly, I.G., Liberman, A.M., Syrdal, A.M., and Halwes, T. (1971) Discrimination in speech and nonspeech modes. *Cog. Psychol. 2*, 131–157.

Miller, J.D. (1977) Perception of speech sounds in animals: Evidence for speech processing by mammalian auditory mechanisms. In T.H. Bullock (Ed.), *Recognition of Complex Acoustic Signals*. Berlin, Abakon Verlagsgesellschaft.

Miller, J.D., Wier, C.C., Pastore, R., Kelly, W.J., and Dooling, R.J. (1976) Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *J. acoust. Soc. Amer., 60*, 410–417.

Parks, T., Wall, C., and Bastian, J. (1969) Category and intracategory discrimination for one visual continuum. *J. exper. Psychol., 31*, 241–245.

Rand, T.C. (1974) Dichotic release from masking for speech. *J. acoust. Soc. Amer., 55*, 678–680.

Remez, R.E. Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981) Speech perception without traditional speech cues. *Sci., 212*, 947–950.

Repp, B.H. (1982) Phonetic trading relationships and context effects: New experimental evidence for a speech mode of perception. *Psychol. Bull., 92*, 81–110.

Repp, B.H. (in press) Categorical perception: Issues, methods, findings. In N.J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 9). New York, Academic Press.

Repp, B.H., and Mann, V.A. (1981) Perceptual assessment of fricative-stop coarticulations. *J. acoust. Soc. Amer., 69*, 1154–1163.

Repp, B.H., Milburn, C., and Ashkenas, J. (1983) Duplex perception: Confirmation of fusion. *Percep. Psychophys., 33*, 333–337.

Siegel, J.A., and Siegel, W. (1977) Categorical perception of tonal intervals: Musicians can't tell *sharp* from *flat*. *Percep. Psychophys., 21*, 399–407.

Studdert-Kennedy, M. (1980) Speech perception. *Lang. Sp., 23*, 45–66.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970) Motor theory of speech perception: A reply to Lane's critical review. *Psychol. Rev., 77*, 234–249.

Reference Note

1. Bentin, S., and Mann, V.A. Speech and nonspeech perception in duplex: Prevalence of the phonetic mode. Manuscript in preparation.

Résumé

Correctement incorporées, les transitions du troisième formant fournissent un support critique pour les percepts phonétiques [d] et [g]. Présentés isolément ces formants sont perçus comme des bruits (chirps) variant dans le temps. Dans l'expérience suivante, les deux formes de perception sont disponibles simultanément par la présentation des transitions du troisième formant à l'une des oreilles et par la présentation du reste de la syllabe acoustique à l'autre. Sur le versant parole de ce percept présenté en 'duplex', les transitions permettent la perception de syllabes plosives-voyelles. Cette perception est catégorielle et est influencée par la préposition des syllabes [al] ou [ar]. Dans la condition 'non parole', ces mêmes transitions sont entendues comme des 'chirps', la perception est continue et les syllabes préposées ne l'influencent pas. Les différences se produisant alors que l'input acoustique est constant, pourraient refléter les propriétés différentes des modes auditifs et perceptifs de perception.