Some experiments on the sound of silence in phonetic perception[a]

Michael F. Dorman, Lawrence J. Raphael, and Alvin M. Liberman

Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510

(Received 1 August 1978; revised 5 February 1979)

The results of several experiments demonstrate that silence is an important cue for the perception of stop-consonant and affricate manner. In some circumstances, silence is necessary; in others, it is sufficient. But silence is not the only cue to these manners. There are other cues that are more or less equivalent in their perceptual effects, though they are quite different acoustically. Finally, silence is effective as a cue when it is part of an utterance that is perceived as having been produced by a single male speaker, but not when it separates utterances produced by male and female speakers. These findings are taken to imply that, in these instances, perception is constrained as if by some abstract conception of what vocal tracts do when they make linguistically significant gestures.

PACS numbers: 43.70.Dn, 43.70.Ve

INTRODUCTION

The several experiments to be reported here have in common a concern with silence as one of the cues for the perception of stop consonants. They were designed to illuminate further the processes by which that cue does its perceptual work.

That silence is important for the perception of stops has been established by several studies. Indeed, silence has been found to play a role in perceiving each of the three features—manner, voicing, and place—that a stop consonant comprises. Consider manner. By cutting and splicing magnetic tapes, Bastian, Eimas, and Liberman (1961) showed that the syllable "slit" is heard as "split" when a short interval of silence (about 40 ms) is introduced between the noise at the beginning of the syllable and the vocalic portion. As for voicing, Lisker (1957a) early found that intervocalic stops in trochees were perceived as voiced or voiceless (e.g., "rabid" or "rapid"), depending on the duration of silence between the syllables. Turning finally to place, we take account of the finding by Port (1976) that "rabid" is perceived as "ratted" when the duration of silence between the syllables is reduced.

Our experiments will deal only with the perception of stop-consonant manner. Taken together, and added (when appropriate) to the work of others, they are meant to bear on three related questions: (1) In what circumstances is silence a cue? (2) Does silence have its effect exclusively in the auditory domain, or also at some more abstract (phonetic) remove where perception is constrained as if by knowledge of what a vocal tract does when it makes linguistically significant gestures? (3) If the latter, then whose vocal tract provides the constraint?

I. SILENCE AS A NECESSARY CONDITION BEFORE AND AFTER THE VOWEL; PERCEPTION OF TRANSITION CUES IN SPEECH AND NONSPEECH CONTEXTS

Evidence pointing to the importance of silence as a manner cue came first from experience with syllables in which a stop is (or is not) heard before the vocalic nucleus. Thus, in the early study by Bastian *et al.* (1961), the contrast was between "slit" and "split." Given similar phonetic contexts, the same effect is readily found, so readily indeed that it has become part of the lore of those who experiment with speech, and is taken into account in those formal rules that specify how speech is to be synthesized. In contrast, there is little information about the importance of silence as a manner cue for the perception of stops that follow the vocalic nucleus. We can infer, however, from an early observation by Lisker (1957a) and a more recent study by Abbs (1971) that a silent interval of some length must follow a vowel-stop syllable if the stop is to be perceived.

Our aim is to learn more about these phenomena. To that end, we will first assess the role of silence in the perception of stops (before the vowel) in the syllables [ʃpɛ] and [ʃkɛ] and (after the vowel) in the disyllables [bɛbdɛ], [bɛgdɛ] and [bɛddɛ]. If, as we have reason to expect, silence proves to be important, we will use the results as a basis for further studies that might help us to understand why. Some of those will be reported in this section, others in the sections that follow.

To see what choices we face when we wonder why silence should be a cue for stops, we should first consider the perceptual consequences of altering the acoustic structure of the fricative–vowel syllable shown in Fig. 1: having recorded a naturally produced token of [sa], we find that removing the initial fricative noise will often leave a syllable that sounds like /da/ or [ta]; if

[a] The results of experiments 1, 2a, 2b, and 6 were described in a paper presented at the 89th Meeting of the Acoustical Society of America, Austin, Texas, 1975, and in the Haskins Laboratories (1975). Status Report on Speech Research, SR-42/43, 265–276; experiments 3, 4, and 7 at the 91st Meeting of Acoustical Society of America, Washington, DC, 1976, and (experiments 3 and 4 only) in the Haskins Laboratories (1976) Status Report on Speech Research SR-48, 199–208; and the results of experiments 5a and 5b at the 93rd Meeting of the Acoustical Society of America, State College, PA, 1977.

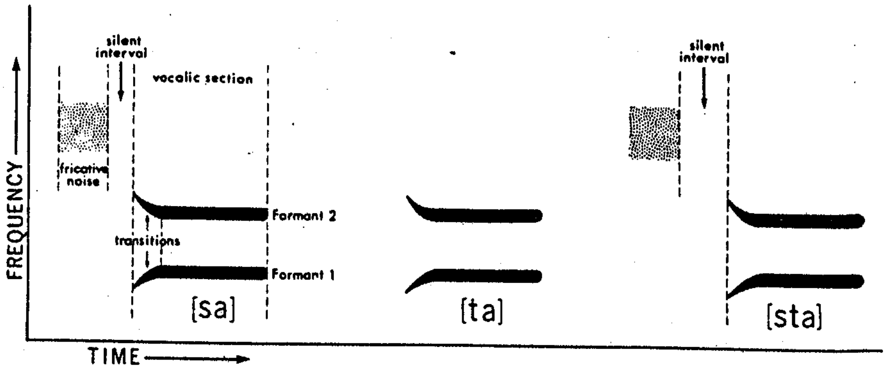

FIG. 1. Schematic representation of stimulus patterns sufficient for the perception of [sa], [ta], and [sta]. Adapted from Liberman and Pisoni, 1977.

we store the noise, but move it backward in time so as to leave a brief (say 50 ms) interval of silence between it and the vocalic portion of the syllable, we produce a syllable that sounds like [sta] (Bastian, 1962). At one level of interpretation there is no mystery in this: the fricative [s] and the stop [t] have similar places of production, hence similar formant transitions. But it is not so clear why silence is necessary in order for the transition cues to give rise to the perception of a stop—that is, why a stop is not heard when fricative noise and formant transitions are separated by only a brief interval.

Broadly speaking, two interpretations are possible. The one we are inclined to favor is that the silence provides information to a (phonetic) perceiving device that is specialized to make appropriate use of it. To see why that is at least plausible, consider that a speaker cannot produce a stop without closing his vocal tract, and that he cannot close his vocal tract without producing a corresponding period of silence. When the listener hears an insufficiently long period of silence between the fricative noise and the vocalic section, it is, by this account, as if he "knew" that a stop should not be perceived because it was not produced.

An alternative interpretation puts the effect of the silence cue squarely in the auditory domain. Thus, we note about the example just offered, that it conforms to the paradigm for auditory forward masking. Conceivably, the fricative noise masks the transition cues that otherwise would be sufficient for the stops; in that case, the role of silence would be to provide time to evade masking. Or, keeping the interpretation still in the auditory domain, we might suppose that the silence collaborates in some kind of perceptual interaction with the transition cues, the result of the interaction being that experience we call a stop.

Some evidence relevant to these interpretations is already available. Harris (1958), for example, found recognition of the [f]–[θ] contrast to be contingent primarily on the formant transitions that follow the fricative noise. This situation could only arise if the formant transitions had different effects in the auditory domain—that is, if they were not masked by the preceding noise. Evidence from dichotic listening supports this conclusion. Thus, Darwin (1971) found a larger right-ear advantage for fricatives synthesized with appropriate formant transitions following the fricative noise, than for fricatives synthesized without formant transitions. In this instance, too, the transitions must have had different auditory representations when they arrived at the central processing mechanisms responsible for the ear advantage.

Another piece of relevant evidence comes from a study of selective adaptation. Following a now standard adaptation procedure, Ganong (1975) first measured the displacement of the [bɛ–dɛ] boundary caused by adaptation with [dɛ]. Fricative noise was then placed in front of the [dɛ], and the (perceived) [sɛ] that resulted was used as the adapting stimulus. The outcome was a shift in the [bɛ–dɛ] boundary as large as that found when the adapting stimulus was [dɛ]. Patterns that contained the noise, but not the formant transitions, did not produce so large a shift. This indicates not only that the transition cues were getting through, but that they were getting through in full strength.

Thus, we are led to believe that the transition cues make a significant perceptual contribution, whether or not they are preceded by a period of silence. On that view, silence is important, not because it provides time to evade masking, or because it collaborates in an auditory interaction, but because it provides information that is essential to determining how the transitions are to be interpreted in phonetic perception.

The experiments in this section are designed to get at that matter via a different—perhaps more direct—route by comparing the effect of the fricative noise on transition cues that are, in one case, in a speech context, and in the other, not. The results will bear, of course, on a masking interpretation, but also on the possibility of auditory interactions, since we will be able to determine whether or not there are qualitative changes in the perception of the nonspeech transition cues depending on the presence or absence of the silence.

A. Experiment 1

Our first experiment was designed (1) to assess the role of silence in the perception of stop manner prevocalically in the syllables [ʃ pɛ] and [ʃ kɛ], and (2) to determine whether the fricative noise of [ʃ] masks or interacts with information carried on the transition cues for the stops when those are isolated from the rest of the syllable and are heard as nonspeech.

1. Method

Two sets of stimuli were made. Members of the one—to be referred to as the "speech" stimuli—were appropriate for determining the effect of silence on the perception of the stop consonants in [ʃpɛ] and [ʃkɛ]. They were made in the following way. First, the syllables [ʃɛ], [gɛ], and [bɛ] were recorded by a male speaker, then digitized and stored, using the Pulse Code Modulation (PCM) system at Haskins Laboratories.[1] Working from high-resolution oscillograms, and taking advantage of computer control. we next separated the fricative noise of the [ʃ] from the vocalic portion of the syllable [ʃɛ], and removed the syllable-initial bursts from the [gɛ] and [bɛ]. To create the experimental stimuli, we prefixed the ʃ noise to what remained of the [bɛ] and [gɛ], leaving silent intervals of 0, 4, 8, 12, 16, 20, 40, 60, 80, and 100 ms between the offset of the fricative noise and the vocalic section appropriate for [gɛ] and [bɛ] [see Fig. 2(a) for a schematic representation of one of the ʃ noise plus [gɛ] stimuli]. Four tokens of each stimulus type were produced. These were randomized and recorded on magnetic tape with a 3-s interval between stimuli.

Members of the other set—to be referred to as the "nonspeech" stimuli—were intended to enable us to measure the extent to which the transition cues that distinguish the stops in [ʃpɛ] and [ʃkɛ] are themselves masked by the ʃ noise. These stimuli were made in the following way. First, the [bɛ] and [gɛ] patterns of the speech set were bandpass filtered between 0.9 and 3.5 kHz, and truncated so as to include only the first 50 ms of the signal. This procedure eliminated the first formant, producing signals that contained only the second- and third-formant transitions. (Listeners could hear these stimuli as "chirps," and we supposed that with only a few minutes of practice they would be able to identify them by pitch as "low" or "high.") Then, to create a test of the identifiability of these transitions for comparison with the condition in which they were the essential cues for place of articulation, we prefixed the ʃ noise, setting the same intervals of silence between it and the chirps that we had used in creating the "speech" stimuli. [See Fig. 2(b) for a schematic representation of the "chirp" stimulus derived from the "speech" stimulus shown in Fig. 2(a).] The resulting signals were randomized and recorded on magnetic tape with a 3-s interval between stimuli.

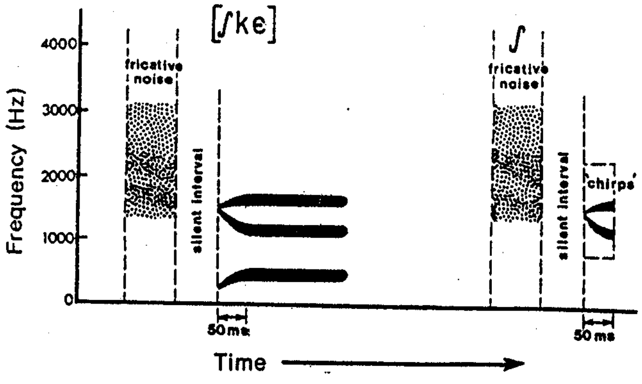

FIG. 2. (a) Schematic representation of one of the speech patterns used in experiment 1. (b) Schematic representation of the corresponding nonspeech ("chirp") pattern.

The subjects were nine volunteers, all undergraduates at Lehman College, who had not previously served in experiments on speech perception. Divided into groups of five and four, they listened in a sound-attenuated room, first to the speech stimuli, and then in a second session, to the "nonspeech" stimuli. In the speech condition, the listeners were told they would hear approximations of the syllables [ʃpɛ], [ʃkɛ], and [ʃɛ], and were asked to indicate on a printed response sheet what they had heard. To provide some "practice," we presented twenty of the stimuli before the experiment proper began; no information was given about the "correctness" of the responses.

In the "nonspeech" condition, the subjects were told they would hear tokens of three stimulus types: ʃ noise alone, ʃ noise followed by a low-pitched chirp (which they were to call "low"), or ʃ noise followed by a high-pitched chirp (which they were to call "high"). They were asked to indicate on their response sheets what they had heard. In this condition, the "practice" consisted of presenting 50 of the stimuli. In order to make sure that the subjects did, in fact, learn to identify the chirps, we provided knowledge of results. To preclude biasing the experimental outcome by experience during the practice sessions, we avoided all short silent intervals—in which the chirps might or might not be heard—presenting only those stimuli in which the noise preceded the chirps by 100 ms. During the experimental session, no information about "correct" responses was given.

In both "speech" and "nonspeech" conditions the stimuli were reproduced via a Revox 1240 tape recorder and AR-4x loudspeaker.

2. Results and discussion

The results for the speech condition are shown in Fig. 3. Since the identification functions for [ʃpɛ] and [ʃkɛ] were found on preliminary examination to have similar shapes, we have averaged them; this facilitates comparison with the identification function for [ʃɛ]. We see that when the silent interval was less than 20 ms, listeners reported hearing [ʃɛ]—that is to say, they did not hear a stop. The stops were identified with

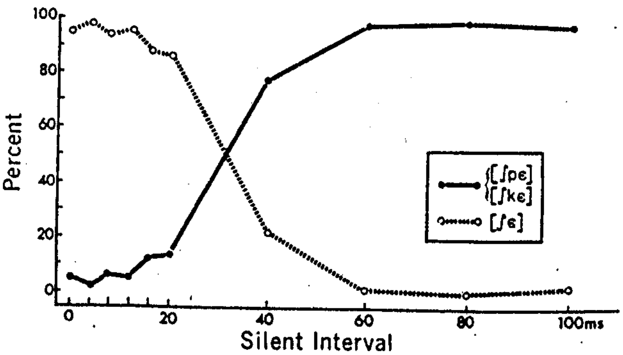

FIG. 3. Silence as a necessary condition for stop manner; identification of stimulus patterns as [ʃpɛ]–[ʃkɛ] or [ʃɛ].

75% accuracy only when the silent interval exceeded about 40 ms. Thus, we find silence to be an important condition for the perception of stops in fricative–stop–vowel syllables.

The identification functions shown in Fig. 3 were derived from the responses of seven of the nine subjects. The two other subjects identified the ʃ noise plus [gɛ] stimuli in the same manner as the group of seven, but made a total of only one [ʃɛ] response to the ʃ noise plus [bɛ] stimuli. To account for that we should consider that in the case of [ʃpɛ] the places of articulation signaled by the fricative noise and the vocalic transitions were quite different, the former being palatal and the latter bilabial. In our own listening to these patterns, it seemed that when there was little silence between ʃ noise and [bɛ], we heard [ʃɛ], but with a nonspeech chirp—as if the transitions could not be integrated into the phonetic percept but were audible nevertheless. It is possible that our subjects, hearing the same chirp, elected to call these stimuli [ʃpɛ]. In the case of ʃ noise plus [gɛ] the disparity in place of articulation was not so great, and it is perhaps for that reason that when the ʃ noise was moved close to the [gɛ] we, and all our subjects, heard only [ʃɛ]. Indeed, the disparity in place of articulation can be reduced even further, as it is, for example, in the case of s-noise plus [ta] that we described in the introduction. There, the places of articulation for the fricative and stop are exactly the same, and the [sa] that results from putting the fricative noise close to the vocalic section is virtually indistinguishable from one that is produced by a human speaker who articulates in a perfectly normal way.

We should emphasize that the interval of silence necessary for stop perception in fricative–stop–vowel syllables is not invariant. Indeed, from the early work of Bastian (1962) and from recent work by Bailey, Summerfield, and Dorman (in preparation) and by Summerfield and Bailey (1977), we know that the interval varies according to how several other cues are set. These include, at the least, the duration of the fricative noise, the rate of fricative noise offset, the rise time of the amplitude envelope of the vocalic portion of the syllable, and the starting frequency of the first-formant transition. (We discuss the importance of such relations among cues more fully in Sec. II.)

We should also emphasize that we do not mean to imply that listeners cannot discriminate between a naturally produced [ʃɛ] and one composed of ʃ noise followed at a brief interval by [gɛ] (or [bɛ]). As we pointed out above, in these cases a listener may hear a normal [ʃɛ] or [ʃɛ] with a nonspeech chirp in it. Now we should add that for some articulations of [gɛ] a fricative noise placed just in front will cause a listener to perceive [ʃjɛ] (Liberman and Pisoni, 1977). The point we wish to make is that listeners do not in such cases commonly report a stop.

Redirecting our attention to experiment 1, we see in Fig. 4 that the results of the nonspeech condition are quite different from those of the speech condition. The isolated formant transitions taken from [bɛ] and [gɛ] were clearly audible—indeed, highly identifiable—as chirps at all intervals of silence, even zero. That outcome is wholly consistent with the evidence presented at the introduction to this section in that transition cues that follow fricative noise are nonetheless effective as auditory events, whether separated from the noise or not. As for the possibility that the transition cues somehow interact with silence, there had previously been no data that were directly relevant. Now we see in the results of our experiment a suggestion that such auditory interaction does not occur: Our subjects not only heard the nonspeech transitions (no matter how close they were to the fricative noise), but they correctly identified them as well; moreover, our own listening made it plain that, more generally, the fricative noise did not appreciably affect the perception of the nonspeech transitions in any qualitative way.

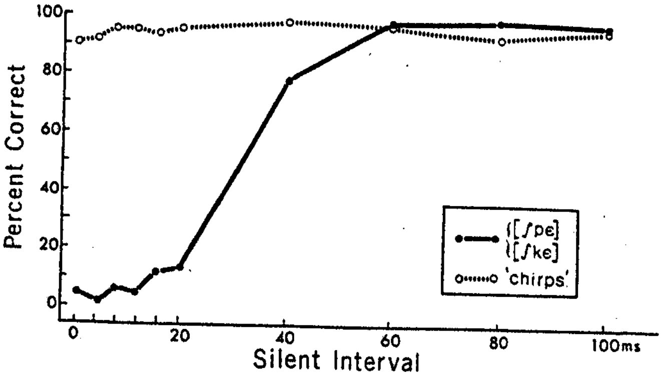

FIG. 4. Percent correct identification of the transition cues in the speech ([ʃpɛ]–[ʃkɛ]) and nonspeech (chirps) contexts.

B. Experiments 2a and 2b

In the previous experiment we found silence to be a necessary condition for the perception of stops in prevocalic position. The experiments reported here were designed to find out if silence is also a necessary condition for the perception of stops in postvocalic position. There were two such experiments, divided according to purpose and the nature of the stimuli.

In one experiment (2a), the stimuli were the synthetic disyllables [bɛb dɛ] and [bɛg dɛ], so made as to provide variation in the interval of silence between the first and second syllables. Given the hypothesis that underlies all the experiments of this paper, we should expect that a relatively long silence would be essential if the listener is to perceive both the syllable-final [b] and [g] and the syllable-initial [d], since a speaker must close his vocal tract for a longer period to say [bɛb dɛ] or [bɛg dɛ] then to say [bɛ dɛ], [bɛ bɛ], or [bɛ gɛ]. Pilot work revealed that with reductions in the duration of the silent interval, it was the syllable-final stops [b] and [g] that disappeared; the syllable-initial [d] could be heard even at very short intervals of silence. This may be owing, in part, to the fact that, in production, the [d], and especially the flapped [d], requires very little closure (Port, 1976), and in part, perhaps, to the fact that unreleased syllable-final stops tend to be relatively unintelligible at best. At all events, it is the syllable-final stops that are, in the kinds of patterns we used, the more sensitive to variations in the duration of intersyllabic silence.

As in the experiments with prevocalic stops, we though it useful to provide data relevant to the possibility that the outcome is to be accounted for in terms of masking—backward masking in the case of the postvocalic stops—or auditory interaction. To that end, we determined whether silence is also necessary for the perception of the formant transitions that are sufficient to distinguish the syllable-final stops when those transitions are presented in isolation, and sound like chirps.

In the other experiment (2b), the stimuli were natural speech, not synthetic, and they included not only [bɛb dɛ] and [bɛg dɛ] but also the geminate condition [bɛd dɛ].[2] The use of natural speech will permit a comparison with the results obtained when the stimuli were synthetic. The point of testing the geminate condition is that, in production, the articulatory closure for the geminate stops is longer than that for single stops, and a study by Pickett and Decker (1960) leads us to suspect that the amount of silence necessary for perception may also be longer. A comparison of the two cases of syllable-final stops seemed, therefore, to be in order.

1. Method

To produce stimuli for experiment 2a—the one with synthetic stimuli—we used the Haskins Laboratories parallel-resonance synthesizer to generate two-formant patterns appropriate for the disyllables [bɛb dɛ] and [bɛg dɛ]. A schematic representation of [bɛb dɛ] is shown in Fig. 5. That disyllable differed from the other one [bɛg dɛ] in the second-formant transition, the sole cue in these patterns for the perceived distinction between the syllable-final stops: for [b] the transition is falling, as shown in the figure, while for [g] it is rising. We then introduced periods of silence between the second syllable [dɛ] and the first syllable [bɛb] or [bɛg]. These periods ranged from 0 to 150 ms in steps of 10 ms. Four tokens of each stimulus were generated. To produce a test sequence appropriate for presentation to our subjects, we put these stimuli into a random sequence with a 3-s interval between successive stimuli. That test sequence was used in what will be referred to as the "speech" condition.

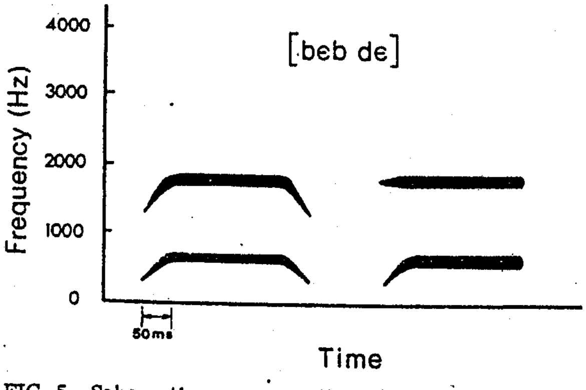

FIG. 5. Schematic representation of one of the stimulus patterns for experiment 2a.

To produce the corresponding stimuli for the "nonspeech" condition, we simply isolated the second-formant transitions that alone distinguished the [bɛb] and [bɛg] patterns of the "speech" stimuli (falling for [b], rising for [g]), and then produced stimuli that were otherwise identical with those of the "speech" condition—that is, we placed after the isolated transitions the same synthetic [dɛ] that had been used in the "speech" condition, and introduced between it and the transitions the same intervals of silence.

The subjects for experiment 2a were six undergraduates at Lehman College who had previously participated in experiments on speech perception. They were tested individually. Test order ("speech" versus "nonspeech") was counterbalanced across subjects. In the "speech" condition, the subjects were asked to respond [bɛb dɛ], [bɛg dɛ], or [bɛ dɛ], and to write their responses. To familiarize the subjects with the stimuli, we had them listen to twenty of the patterns before the experiment began. The stimuli were reproduced on a Revox 1240 tape recorder via TDH 39 headphones.

In the "nonspeech" condition, the subjects were told they would hear a high-pitched chirp followed by [dɛ], a low-pitched chirp followed by [dɛ], or [dɛ] alone. They were asked to respond accordingly. To teach the subjects to identify the chirps, and to make sure they could reliably do so, we first presented 50 [b] and [g] chirps in random order with feedback of results. Then we presented, also in random order, twenty-five [b] and [g] chirps followed in each case, after a 120 ms interval, by [dɛ]. Again, subjects were told the correct answers after they had made their responses. The point of using only the 120-ms interval was to avoid biassing the results by providing "correct" responses in those cases where the [dɛ] syllable was sufficiently close that "masking" might conceivably have occurred. Finally, the test proper was begun.

The procedures for experiment 2b—the one that included the geminate case and was done with natural speech—were as follows. Having recorded a male speaker saying [bɛb], [bɛd], [bɛg], and [dɛ], we used the editing facilities provided by the Haskins Laboratories PCM systems to truncate closure voicing following the syllable-final transitions to 15 ms. To each of the syllables [bɛb], [bɛd], and [bɛg], we then appended the syllable [dɛ], separating it from [bɛb], [bɛd], or [bɛg] by periods of silence that ranged from 0 to 90 ms in steps of 10 ms. Three tokens of each stimulus were generated. These were randomized and recorded onto tape with 4-s interval between stimuli.

The subjects for this experiment were eight volunteers, all undergraduates at Lehman College who had not previously served in speech-perception studies. They were asked to identify each of the stimuli as [bɛb dɛ], [bɛg dɛ], [bɛd dɛ], or [bɛ dɛ] and, in writing their responses, to include the entire syllable. There was a preliminary "practice" session in which the subjects heard and identified twenty stimulus patterns.

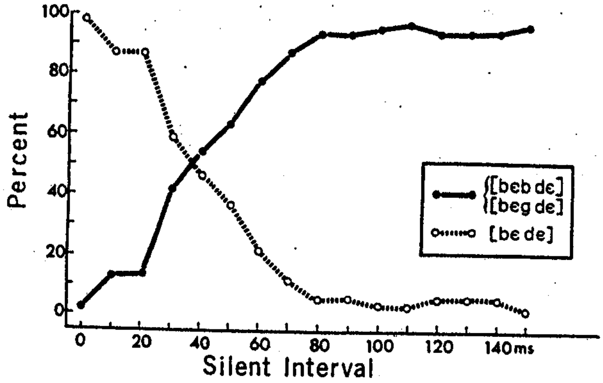

FIG. 6. Silence as a necessary condition for stop manner; identification of stimulus patterns as [bɛb dɛ]–[bɛg dɛ], or [bɛ dɛ].

The signals were produced in the manner described in experiment 1.

2. Results and discussion

The effect of silence on the perception of syllable-final stops in synthetic [bɛb dɛ] and [bɛg dɛ] (experiment 2a) is shown in Fig. 6. There we have plotted the average [bɛb dɛ] and [bɛg dɛ] responses for comparison with the [bɛ dɛ] responses. (The identification functions for [bɛb dɛ] and [bɛg dɛ] were similar, so we have collapsed them into a single function.) One sees that, over the range 0 to about 30 ms of intersyllabic silence, the predominant response was [bɛ dɛ]—that is, our subjects did not report a syllable-final stop.[3] We should emphasize that, as in the experiment on prevocalic stops, it was not the case that a subject heard a stop but misidentified it; rather, he simply did not hear it. A silent interval of about 58 ms was necessary before the subjects identified the stops with 75% accuracy. Thus, for the perception of stops in postvocalic position, as for those that were prevocalic, silence is important.

It will be remembered that we were also concerned with how the isolated formant transitions of the syllable-final [b] and [g] (nonspeech condition) are affected when the stimulus patterns are otherwise exactly the same as in the speech condition just reported. The results of the nonspeech condition are shown in Fig. 7. We note, first, that no subject used the response "no chirp"—that is, no subject ever failed to hear a chirp, even when there was no silence between the chirp and the syllable. This is dramatically different from the result obtained in the "speech" condition. There, given comparable conditions, our subjects did not hear the corresponding syllable-final stops at all. Looking at the percent correct identification of the chirps, we see that at the shortest intervals of silence identification is less accurate than at the longest intervals. Indeed, this difference in accuracy is significant ($F=2.07, p<0.05$). We should note, however, that even at the brief intervals our listeners averaged about 70% correct. Thus, it does not appear that backward masking can account for the complete absence of the stop percept at brief silent intervals.

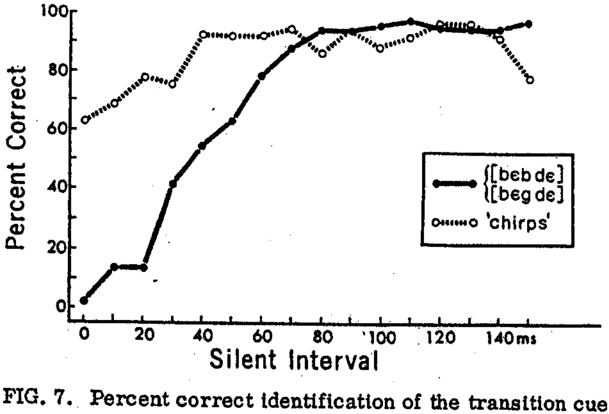

FIG. 7. Percent correct identification of the transition cues in the speech [(bɛb dɛ]–[bɛg dɛ]) and nonspeech (chirps) contexts.

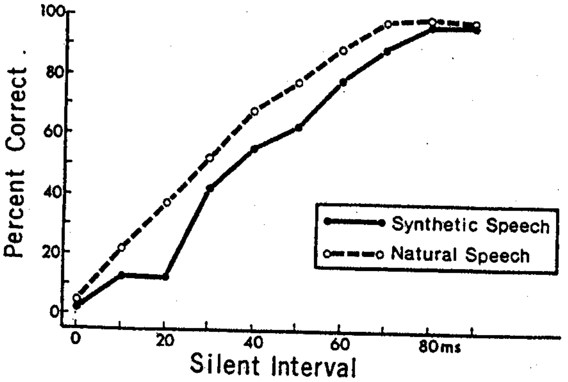

FIG. 8. Identification functions for syllable-final stops in synthetic and natural speech.

We turn now to the results of experiment 2b. It will be recalled that this experiment differed from the previous one in that the geminate condition was included, and natural rather than synthetic speech was used. Let us first compare the results obtained with natural speech and with synthetic speech. For that purpose we will look only at the data pertaining to syllable-final [b] and [g], omitting the geminate condition. These are shown in Fig. 8, together with the comparable data (from Fig. 7) for synthetic speech. The results are quite similar—in both conditions some interval of silence is necessary for listeners to identify a stop. However, the duration of that interval does differ by about 15 ms between the two conditions. We should

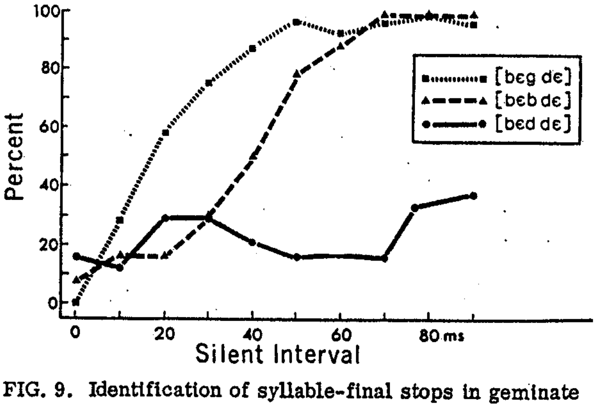

FIG. 9. Identification of syllable-final stops in geminate ([bɛd dɛ]) and nongeminate ([bɛb dɛ] and [bɛg dɛ]) conditions.

suppose that this difference is due to variation between the conditions in the "settings" of the cues (for stop manner) other than silence, e.g., formant transitions.

Turning now to the comparison between geminate and nongeminate stops, we see in Fig. 9 that subjects needed a longer silent interval to identify syllable-final [d] than [b] or [g];[4] even at the longest interval the identification of [d] reached only 38% correct. Further research by Repp (1976) suggests that an interval of approximately 200 ms is necessary for listeners to identify the syllable-final stop in a sequence of identical stops (see also Pickett and Decker, 1960; Fujisaki, Nakamura, and Imoto, 1975). This result is then another piece of evidence that speaks against an explanation of the perceptual disappearance of the syllable-final stops in terms of recognition masking, for one would be hard-pressed to explain why syllable-initial [d] should "backward-mask" syllable-final [d] over a period four times longer than it masks [b], or [g].

More direct evidence that syllable-final transitions are not "backward masked" is also to be found in studies by Repp (1976, 1976). Having presented to listeners VCV's that had been synthesized with and without syllable-final transitions, he found, in the case of stimuli without syllable-final transitions, that the time required to identify the medial consonant increased as a function of the duration of the closure interval; in the case of stimuli with syllable-final transitions, however, the time required was more nearly constant (Repp, 1976). Clearly, then, the syllable-final transitions had a perceptual effect even though they were not heard as discrete phonetic events. This same conclusion can be drawn from another experiment by Repp (1976). In that experiment the syllable [dɛ] was preceded, in the one case, by [ad], in the other case by [ab]. In both cases the listeners perceived [adɛ]. Nevertheless, they discriminated between the stimuli at a level slightly better than chance.

Returning now to our own results, we conclude from experiments 2a and 2b that, just as silence is important for the perception of stops in prevocalic position, so also is it important for the perception of stops in postvocalic position. Moreover, the results are consistent with the evidence presented in the Introduction—namely, that silence is important, not because it provides time to evade masking or because it enters into an auditory interaction, but rather because it provides information about the behavior of a vocal tract.

II. SILENCE AS A SUFFICIENT CONDITION BEFORE AND AFTER THE VOWEL; PERCEPTUAL EQUIVALENCE OF SILENCE AND SOUND

In the studies so far described, stops were (or were not) perceived in patterns that contained transition cues appropriate for stop manner. Now we shall turn to cases in which the transition cues are absent, and it is left to the power of the silence cue itself to produce the effect of a stop. We should note that even in the early study by Bastian *et al.* (1961), silence might have borne the entire burden, but we cannot be sure because the procedures of cutting and splicing the magnetic tape may have introduced a transient, which of itself could contribute to the perception of a stop. We should also note that others (Summerfield and Bailey, 1977), working independently of us, have recently demonstrated the power of silence to cue stop manner prevocalically in the context of fricative–vowel versus fricative–stop–vowel, e.g., [si] versus [ski], where the vocalic section alone is, by perceptual test, not sufficient to produce the stop. At all events, we, too, wish to test the silence cue in such circumstances, and to do it for several positions in the syllable: in prevocalic position ("slit" versus "split"); in intervocalic position ("say shop" versus "say chop," the affricate "ch" [tʃ] being taken here as a stop-initiated fricative); and in postvocalic position ("dish" versus "ditch"). The results may throw more light on the role of silence in the perception of stop manner, since in these instances there are no obvious transition cues to be masked. They will also provide the basis for further investigations into the reasons why silence should have a role in stop perception at all.

To see the point of one of these further investigations we should recall that, as we have supposed, the role of silence might be to tell the listener that the speaker either did or did not close his vocal tract appropriately for the production of a stop consonant. But to make that suggestion is to imply that our perception of speech is constrained to some degree by a device that acts as if it knew what vocal tracts can and cannot do when they make linguistically relevant gestures; or, more generally, that there is, in speech, a link between perception and production. Further evidence for such a link comes, for example, from studies that have established an equivalence in phonetic perception between cues that are very different from an acoustic (and presumably auditory) point of view, but which are the correlated results of the same articulatory gesture. One of the earliest of these is of special interest to us because it dealt with silence, albeit as a cue to voicing rather than manner (Lisker, 1957b). The context was that of "rabid" versus "rapid." The results were (1) that variation in the duration of intersyllabic silence was sufficient to cue the voicing distinction between the two words, and (2) that the location of the voicing boundary on the continuum of intersyllabic silence varied as a function of whether the stimuli were synthesized, say, with or without a transition of the first formant at the end of the first syllable. Thus, cues with different acoustic properties were nevertheless found to be equivalent in phonetic perception: Just as stimuli characterized by the presence of a transition of the first formant and a relatively long silent interval were heard as "rapid," so also were stimuli characterized by the absence of a transition of the first formant and a shorter silent interval. We should ask now why silence should give rise to the same phonetic percept as the frequency modulation of the first-formant transition. The answer is surely hard to find so long as we think in terms of what we know, or can surmise, about auditory perception. But in articulation we find the tie that binds: These acoustically dissimilar events are both to be found among the many acoustic consequences of the gesture that converts "rabid" to "rapid." There are other, equally diverse acoustic consequences of the gesture, and these, too, according to the results of the early study and its current extensions (Lisker, 1977) have an equivalence in phonetic perception.

Since articulatory gestures commonly have multiple and diverse acoustic consequences, we should expect to find many cases of such perceptual equivalence among acoustically dissimilar cues. To be sure, there is no problem in finding such cases; they abound, and have been studied for all three phonetic dimensions: manner, voicing, and place. (For a review, see Liberman and Studdert-Kennedy, in press). In the third experiment of this section we examine one additional case. Taking advantage of the fact that the stop gesture which differentiates fricative from affricate in "ditch" versus "dish" generates changes in both the duration of the silent closure interval and changes in the onset and duration of the fricative noise, we examine the perceptual equivalence between silence, on the one hand, and, on the other, the rise time of the friction and also its duration.

A. Experiment 3

Our third experiment was designed to determine whether the perception of "split" could be induced by inserting silence between the fricative noise of [s] and the syllable "lit." Is silence, in this sense, a sufficient condition for the perception of stop manner, and, if so, over what range of durations is silence effective? The second question is interesting because we know that neither a very brief nor a very long closure is appropriate for stop manner. A too-brief closure would presumably indicate that the speaker had not closed his vocal tract long enough to have said "split." A too-long closure, on the other hand, would suggest that he had produced the "s," then waited a while, and finally said "lit." That being so, we would suppose that only a limited range of silent intervals would signal the production of stop manner.

1. Method

A male speaker's recordings of the fricative noise of [s] and the syllable "lit" were digitized and stored in computer memory. (Both segments were produced in isolation.) Having listened carefully to these segments, we judged that the noise of the [s] did not end with a stop, nor did the "lit" begin with a stop. Using the editing facilities provided by the Haskins Laboratories PCM system, we then appended the "s noise" to the "lit," separating these two segments by intervals of silence that ranged from 0 to 100 ms in steps of 15 ms, and from 100 to 650 ms in steps of 50 ms. Three tokens of each stimulus were generated. The resulting stimuli were randomized and recorded on audio tape with a 3-s interval between stimuli. The listeners were instructed to label the stimuli as "slit," "split," or "s" followed by "lit." (The last named category is not "slit," but rather "s" plus "lit," with a clearly perceptible period of silence in between.)

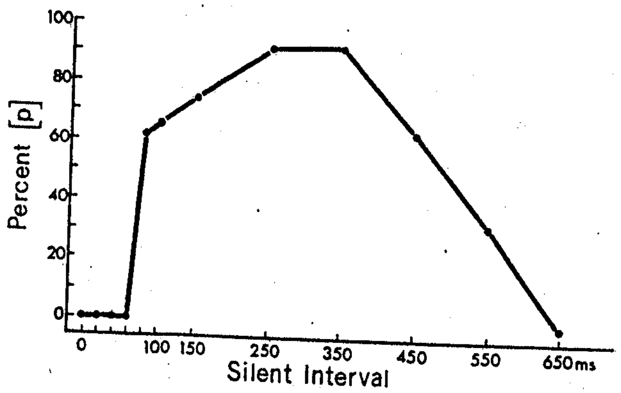

FIG. 10. Silence as a sufficient condition of stop manner; identification of [p] in patterns composed of "s" followed by "lit."

The subjects were ten volunteers, all undergraduates at Lehman College who had not previously served in experiments on speech perception. They were tested in two groups of five, each under conditions similar to those of experiment 1. To familiarize the listeners with the stimuli, we had them listen to the entire stimulus continuum before the test sequence began.

2. Results and discussion

The effect of inserting intervals of silence between the "s-noise" and [lit] is shown in Fig. 10. There we see that at silent intervals of less than 60 ms listeners reported "slit," but at longer intervals—out to about 450 ms—they reported "split." In this case, then, silence is a sufficient condition for stop manner. Notice, however, that at the longest silent interval the stop was not heard; rather, the subjects reported "s-silence-lit." Thus, neither the very brief nor the very long silent intervals produced a stop percept. This outcome accords well with our earlier supposition that only a limited range of silent intervals should signal stop manner.

B. Experiment 4

To this point we have investigated silence as a condition for the perception of stop manner. Now we turn to silence as a condition for affricate manner. To see why, consider that just as a speaker must close his vocal tract to produce the stop that distinguishes, for example, [sta] from [sa], so also must he close his vocal tract to produce the stop-initiated fricative (i.e., affricate) that distinguishes, for example, the phrase "say chop" from "say shop." There is evidence, moreover, that the silence associated with vocal-tract closure is a cue for the affricate-fricative contrast in intervocalic position. This evidence comes from early experiments with synthetic speech (Kuypers, 1955, Truby, 1955). The purpose of the experiment to be described here is to replicate and expand these early findings. Specifically, we aim to determine whether silence can be a sufficient condition for the fricative-affricate contrast in the naturally produced utterances "say shop" and "say chop."

1. Method

A male speaker's recording of "please say shop" was digitized and stored in computer memory. Using the editing facilities provided by the Haskins Laboratories PCM system, we removed the initial 15 ms of ʃ noise from "shop." The signal that remained still sounded to us like "shop."

We should note parenthetically that in situations of this kind, where there are presumably a number of different cues for the same distinction, it often happens that relatively extreme "settings" of one of the cues will cause the other cues to be "overridden" in perception. For example, in this case, we have reason to believe that the duration and onset of the frication noise, as well as silence, are cues to the affricate-fricative distinction (see Gerstman, 1957). Very long fricative noise, especially when combined with slow onset, may so bias perception toward the fricative that no amount of the silence cue can be effective.

To generate our experimental stimuli we inserted intervals of silence between the offset of "please say" and the onset of "shop." These intervals covered the range 0 to 400 ms. The steps were 10 ms each from 0 to 100 ms and 50 ms each from 100 to 400 ms. Four tokens of each stimulus were generated. The resulting stimuli were randomized and recorded on audio tape with a 4-s interval between stimuli.

The subjects were ten volunteers, all undergraduates at Lehman College who had not previously participated in experiments on speech perception. They were tested *en masse* under listening conditions similar to those of experiment 1. The subjects were told they would hear either "please say shop" or "please say chop," and were instructed to write either "shop" or "chop" on their response sheets. To familiarize them with the experimental stimuli, we played twenty of the stimuli before the test sequence began.

2. Results and discussion

The effect of varying the duration of the silent interval between "please say" and "shop" is shown in Fig. 11. We see that "chop" responses begin to appear when the silent interval exceeds about 30 ms; by 70 ms they account for 75% of the responses. Thus, we conclude that silence can be a sufficient cue for distinguishing the affricate [tʃ] from the fricative [ʃ]. We should remark that, according to preliminary research we have done, the contrast between the voiced counterparts of those phones (i.e., [dʒ] and [ʒ]) can also be cued by silence.

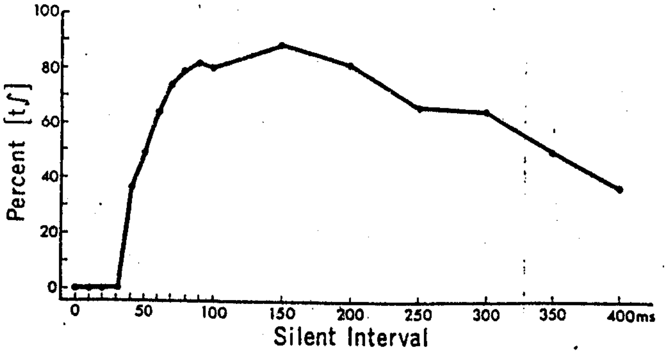

FIG. 11. Silence as a sufficient condition for affricate manner; identification of [tʃ] in patterns composed of "please say" followed by "shop."

Redirecting our attention to the data for the voiceless forms shown in Fig. 11, we see that at the very long intervals of silence there is a tendency for our listeners' perceptions to revert to the fricative [ʃ]. This tendency is similar to that we saw in the case of silence as a cue for stop manner in the contrast "split" versus "slit" (cf. Fig. 10), but it is not so marked. In that connection we note that the longest silent interval for the present experiment with "shop" and "chop" was 400 ms, whereas for the earlier experiment with "slit" and "split" it was 650 ms. When we examine the identification functions for "slit" versus "split," we see that at 400 ms our listeners' responses had only just begun to revert to "s-silence-lit." Presumably, then, in the present experiment, the "chop" responses would have reverted more nearly to "shop" had we carried the silent interval to greater lengths.

Having seen that we convert the utterance "please say shop" into "please say chop" by appropriately increasing the silent interval between "say" and "shop," we should wonder whether we can start with the utterance "please say chop" and convert it to "please say shop" by shortening the silence. The results from preliminary research suggests that this can, indeed, be done, though just how convincingly depends upon the "intensity" of the affricate articulation in "chop" (Raphael and Dorman, 1977). Of course this is analogous to the results obtained in experiments 1 and 2, where too little silence caused stops not to be heard.

C. Experiment 5a and 5b

Having found silence to be sufficient for the perception of affricate manner in syllable-initial position ("shop" versus "chop"), we now wish to determine whether it can be sufficient in syllable-final position, as in "dish" versus "ditch." We also wish in these experiments to examine the effects of two other cues for affricate manner—namely, the duration and rise time of the fricative noise (see Gerstman, 1957)—and to study such relations as there may be between these two cues, on the one hand, and silence on the other.

1. Method

To provide a basis for the stimuli of experiments 5a and 5b, we twice recorded a male speaker saying "put it in the dish." These recordings were digitized and then stored in computer memory. To produce the experimental variation of primary interest we used the PCM editing system to introduce varying durations of silence between the end of voicing associated with the vowel [I] and the beginning of the noise of [ʃ]. These durations ranged from 20 to 150 ms in steps of 10 ms. To enable us to study the effects of the silence cue in combination with the cues of duration and rise time of the fricative noise, we introduced the silence cue into two series of stimuli. In one (experiment 5a) we combined the silence cue with each of two durations of fricative noise, 320 ms and 160 ms, using for this purpose one of the two recordings referred to above. We produced the two durations of noise in the following way. For one we simply used the noise of the original utterance, which was 320 ms in duration. To produce the other, which was 160 ms in duration, we removed 160 ms of noise from the center and then rejoined the cut ends. That operation obviously does not affect the onset or offset characteristics of the noise.

In the other series we combined the silence cue with each of two different conditions of noise rise time, using for this purpose the second of the recordings referred to above. We produced the two rise times in the following way. For one, we simply used the rise time of the original utterance, which was 35 ms. For the other, we reduced the rise time to 5 ms by removing the first 30 ms of the noise. To compensate in the simplest possible way for the resulting reduction in overall duration of the noise, we added 30 ms of noise to the center. (Given that the rise time was not instantaneous, this operation does not ensure that the durations of the stimuli with the two conditions of rise time were psychologically equal. We should note, however, that they were more nearly so than they would have been if the 30-ms insertion had not been made.)

The subjects for experiment 5a were ten undergraduate volunteers from Arizona State University who had not previously participated in research on speech perception. They were tested *en masse* in a large sound-attentuated room. The experimental stimuli were reproduced on a Magnecord 1032 tape recorder via a CEI 41-2 loudspeaker. The subjects for experiment 5b were 12 undergraduate volunteers from Lehman College who had not previously participated in research on speech perception. They were tested in groups of four under the conditions described for experiment 1. The subjects in both experiments were given the same instructions. They were told that they would hear either "put it in the dish" or "put it in the ditch" and were instructed to write either "sh" or "ch" on their response sheets. To familiarize the subjects with the experimental stimuli, we had them listen to twenty stimuli before we started the test sequence.

2. *Results and discussion*

We see the results of experiment 5a in Fig. 12. It is apparent that silence is sufficient in this case to cue the distinction between fricative and affricate postvocalically. At the short intervals of silence the stimuli in both conditions of fricative noise duration were heard as "dish," while at the longest intervals of silence they were heard as "ditch."

It is also apparent that there is a relation between the duration of silence and the duration of fricative noise. Thus, if we look at the silent interval necessary for 50% "ditch" responses, we see that it is approximately 75 ms when the noise is long (320 ms), but only 55 ms when the noise is short (160 ms). The difference in silent interval is significant ($T=0$, $p<0.005$). That is to say that 14 ms of silence (the difference between 89 and 75 ms) is equivalent in these phonetic perceptions to 160 ms of noise.

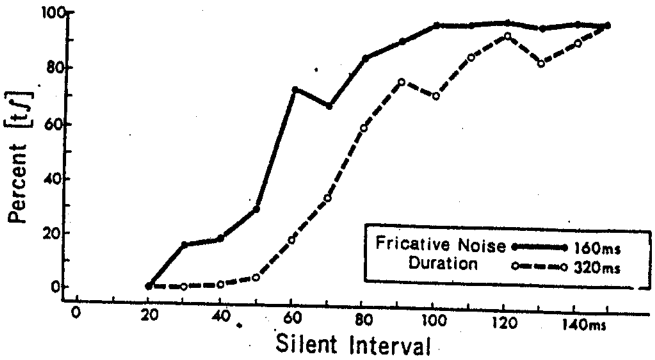

FIG. 12. The relation between silence and sound; identification of [tʃ] for two conditions of fricative-noise duration.

In Fig. 13 we see the results of experiment 5b. Since listeners report "dish" at the shortest intervals of silence and "ditch" at the longest intervals, we see, once again, that silence is sufficient to distinguish between fricative and affricate. And here, too, we see a relation between two acoustic cues to the same distinction: silence and rise time of the fricative noise. The boundary between fricative and affricate is at about 57 ms of silence when the rise time is slow (35 ms), but at 37 ms when the rise time is rapid (0 ms). This difference is significant ($T=1$, $p<0.005$).

We should note that relations of the kind described here can limit the effectiveness of silence as a cue. At one extreme we might have such a long duration of noise, and thus a strong bias toward a fricative, that no amount of silence would be sufficient to overcome it. At the other extreme we might have such a short duration and rise time of the noise, and thus so strong a bias toward the affricate, that even durations of silence near 0 ms would not alter the perception of the affricate. This is consistent with the caveat we mentioned in our earlier discussion. It would apply also in the case of "slit" and "split" to the trading relation between temporal (silence) and spectral cues that have been reported by other investigators (Erickson, Fitch, Halwes, and Liberman, 1977; Liberman and Pisoni, 1977).

Returning now to the main findings of our experi-

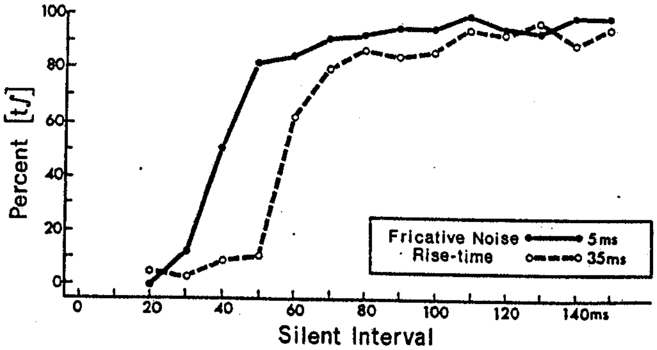

FIG. 13. The relation between silence and sound; identification of [tʃ] for two conditions of fricative-noise rise time.

ment, we should note that the relations among the effects of the several cues are, in principle, like those that have been reported for numerous others (for a review, see Liberman and Studdert-Kennedy, in press). In all cases, cues that are quite different from an acoustic point of view, nevertheless give rise to the same phonetic percept. It is consistent with our hypothesis to suppose that the perceptual equivalence of these cues is owing to the fact that they are the common products of the same linguistically significant gesture.

III. HOW THE EFFECTIVENESS OF SILENCE DEPENDS ON WHETHER IT COMES FROM ONE VOCAL TRACT OR TWO: AN ECOLOGICAL FACTOR IN PHONETIC PERCEPTION

Having suggested that silence is important in stop perception because it provides information about the behavior of a vocal tract, we should now ask: whose vocal tract? We think it could hardly be that of the listener, nor of the speaker, nor, indeed, of any particular person. Rather, it must be some more abstract conception of the behavior of vocal tracts in general. At all events, it is possible to find out; we need only take advantage of certain facts about the ecology of speech.

Consider two of the examples we developed in the earlier parts of our paper. First there was the case of [bɛb dɛ] and [bɛg dɛ], where it was found that a syllable-final stop was not perceived when there was an insufficiently long period of silence between the syllables. We assumed that this was so because the relatively short silence informed the listener that the speaker must not have closed his vocal tract long enough to have produced a syllable-final stop. But what one speaker cannot do, two speakers can: Given collaboration between two speakers—or, indeed, given the accidents of speech when several are talking—the utterance [bɛb dɛ], for example, can be produced with no silence at all between the syllables. Therefore with two speakers, (or more generally two sources of speech) the presence or absence of silence should have no phonetic significance.

Similar considerations apply to our finding that the phrase "please say shop" was heard as "please say chop" when silence was inserted between "say" and "shop." By our account, the silence told the listener that the speaker had closed his vocal tract in a manner appropriate to the production of an affricate; hence, the perception of an affricate. But here, too, the presence or absence of silence provides information only when there is but one speaker, for two can produce "please say" and "chop" with no silence at all between the words "say" and "chop."

Thus, silence does, or does not, provide useful phonetic information depending on whether (and how) the utterance was produced by one speaker or by two. The aim of the experiments to be reported here is to determine if listeners behave accordingly.

A. Experiment 6

The purpose of this experiment was to discover whether the effect of intersyllabic silence on the perception of syllable-final stops in the disyllables [bab da] and [bag da] is different when the syllables are produced by two speakers instead of one.

1. Method

Except for the introduction of a "different voice" condition, the procedures of this experiment were similar to those of experiments 2a and 2b, where, as the reader may recall, we were concerned with the effect of intersyllable silence on the perception of syllable-final stops in [bɛb dɛ] and [bɛg dɛ]. First, we recorded a male saying [bab], [bag], and [da]. Those utterances were digitized and stored in computer memory. We then modified the [bab] and [bag] syllables by removing all but 15 ms of the voicing that followed the final formant transitions. To create the set of stimuli for the "same-talker" condition, we appended the syllable [da] to [bab] and [bag], so as to create intersyllabic intervals of silence from 0 to 90 ms in steps of 10 ms. Three tokens of each stimulus were generated. The entire sequence was then randomized and recorded on audio tape with a 3-s interval between stimuli. To generate stimuli for the "different-talker" condition, we followed exactly the same procedure, but substituted a female voice saying [da]. Thus, we produced disyllables in which the first syllable ([bab] or [bag]) was in a male voice and the second syllable [da] in a female voice.

The subjects were ten volunteers, all undergraduates at Lehman College who had previously participated in experiment 1. For the "same-talker" condition, the subjects were told that they would hear a male voice saying [bab da], [bag da], or [ba da]. For the different-talker condition, the subjects were told that they would hear a male voice saying [bab], [bag], or [ba] followed by a female voice saying [da]. In both conditions the subjects were asked to respond by writing on their response sheets the identity of the sound ([bab da], [bag da], or [ba_da]) at the end of the first (male produced syllable). The stimuli of the same- and different-talker conditions were presented in blocks. To control for practice effects, the order of the blocks was counterbalanced across the listeners. To familiarize the listeners with the stimuli, we presented 20 stimuli before each trial block.

2. Results and discussion

The results for the same- and different-talker conditions are shown in Fig. 14. Looking first at the same-talker condition, we see a result very similar to the one obtained in the analogous condition of one of our earlier experiments (experiment 2b): At short intervals of silence listeners did not hear syllable-final stops; these were heard with 75% accuracy only when the silent interval was about 45 ms in duration.

The result of the different-talker condition is quite different. Eight of the ten subjects identified syllable-final stops with near-perfect accuracy at every inter-

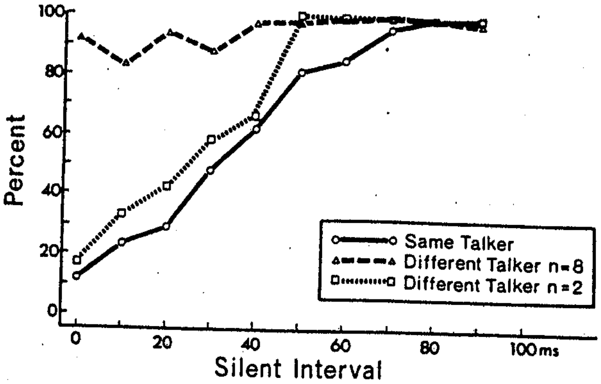

FIG. 14. Silence as a condition for stop manner when it reflects the behavior of one vocal tract or two: identification of syllable final stops in [bɛb dɛ]–[bɛg dɛ] in the same- and different-talker conditions.

val of silence, including even the very shortest. For these subjects, it is as if their perceptual machinery "knew" that, with two speakers, intersyllabic silence conveys no useful phonetic information. The remaining two subjects behaved in the different-talker condition almost exactly as they had when there was but a single talker. We cannot be sure why. We may note, however, that a single syllable by each talker provides very little information about the identity of the talker. Conceivably, therefore, the fact that the two syllables were produced by different talkers did not properly "register" with these two subjects. In that connection, it is relevant that one of these two subjects did remark at the end of the experiment, that she thought she had been listening to the same talker speaking on two different pitches. This suggests that the effect we obtained in the different-talker condition was not due solely to the acoustic differences between the voices as such, but rather to their role in informing the listeners that there were, indeed, two sources of speech.

B. Experiment 7

The purpose of this experiment was to determine if the effect of silence in converting "say shop" to "say chop" is different when the words on either side of the silence are produced by two talkers instead of one.

1. Method

The stimuli for this experiment were produced in the same manner as those of experiment 4, except for the addition of a "different-voice" condition. First we digitized and stored in computer memory a male speaker's recording of "please say shop." To produce stimuli for the same-talker condition, we imposed intervals of silence between "please say" and "shop" in 10 ms steps over the range 0–100 ms. Three tokens of each stimulus were recorded. The entire sequence was then randomized and recorded with a 3-s interval between stimuli. To produce stimuli for the different-talker condition, we first digitized a female's recording of "please say shop." The phrase "please say" was excised from the recording and stored in computer memory. We then appended the male-produced "shop" to the female-produced "please say," leaving intervals of silence between "say" and "shop." These intervals ranged from 0 to 100 ms in steps of 10 ms. Three tokens of each stimulus were generated. The resulting stimuli were randomized and recorded on audio tape with a 3-s interval between stimuli.

The subjects were ten volunteers, all undergraduates at Lehman College who had not previously participated in research on speech perception. For the "same-talker" condition, the subjects were told that they would hear a male voice saying either "please say shop" or "please say chop." For the different-talker condition, the subjects were told that they would hear a female voice saying "please say" and a male voice saying either "shop" or "chop." In both conditions the subjects were asked to write either "sh" (for "shop") or "ch" (for "chop") on their response sheets. The subjects were tested in two groups of five under the listening conditions described in experiment 1. The stimuli of the same- and different-talker conditions were presented in blocks. The order of the blocks was counterbalanced across the two groups of subjects. To familiarize the subjects with the stimuli, we presented 20 stimuli before each trial block.

2. Results and discussion

The results of experiment 7 are shown in Fig. 15. One sees in the same-talker condition a result similar to that we obtained in the analogus condition of experiment 4: the fricative in the word "shop" was heard as the affricate in the word "chop" when the silent interval between it, and the immediately preceding word exceeded about 45 ms. In contrast, silence had no effect in the different-talker condition: increases in the silent interval did not convert "shop" to "chop."

We should note that the utterance "please say shop" used in this experiment should have provided more information about the identity of the talker (or talkers) than did the two syllables of the previous experiment. This may account for the fact that, in this experiment, though not in the other, the effect of the same- versus different-talker conditions was found in every subject. Perhaps, however, the effect would not have been so large had we used other settings of the cues for the fricative-affricate distinction. Obviously, further

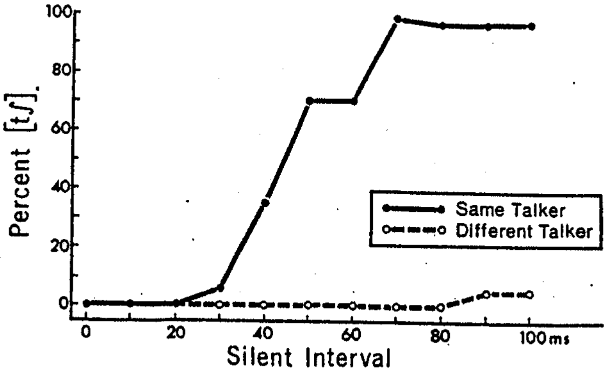

FIG. 15. Silence as a condition for affricate manner when it reflects the behavior of one vocal tract or two: identification of [tʃ] in patterns composed of "please say" and "shop" in the same- and different-talker conditions.

research is necessary to determine the limits over which the effect obtains. We should also wonder about the effect in connection with the trading relations among the fricative-affricate cues that we observed in our earlier experiments. Having found, for example in experiment 5, that duration of silence can be traded for friction duration, we might ask whether these cues also trade with the (perceived) magnitude of the difference between the voices.

We should emphasize that in both experiments the two talkers were male and female. Thus, the acoustic difference between the voices was relatively large. We are now conducting experiments contrived specifically for the purpose of helping us to determine whether the phenomenon we have here described depends critically on such an acoustic difference, or, alternatively, on an inference by the listener that he did or did not hear different sources of speech. At this point, we believe it is the latter.[5]

IV. GENERAL DISCUSSION

We should now assemble the results of our experiments in terms of their bearing on the three questions we raised at the very beginning. As for the first question—Is silence a cue to stop manner?—the answer is quite straightforward, and wholly in accord with the results of previous research. Silence is a cue, necessary in some cases, sufficient in others. Thus, given spectral cues appropriate for a stop in absolute initial position (e.g., [gɛ]), silence preceding those cues was found to be necessary if a stop was to be perceived as the second element of a fricative-stop-vowel syllable (e.g., [ʃkɛ]). Similarly, in the case of stops in syllable-final position (e.g., [bɛb]) silence following the spectral cues was necessary if they were to give rise to the perception of a stop when a second syllable was added (e.g., [bɛb dɛ]). More interesting, perhaps, is the finding that even in the absence of sufficient spectral cues, silence did, in some circumstances, produce the perception of a stop or affricate. Thus, prefixing the noise of [s] to the syllable "lit" produced "split" when the correct amount of silence was interposed; inserting silence between the words "say" and "shop" converted them to "say chop."

Our second question asked whether the effect of silence was exclusively auditory, or also phonetic. If auditory, we should expect to find explanations in terms of masking or any one of a variety of interactions. If phonetic, we should assume that silence informs the listener that the speaker did or did not make the closure that is the distinguishing characteristic of the stops, and further that the listener is sensitive to that information, just as he would be if his perception of speech were constrained by knowledge of what a vocal tract must (or must not) do when it makes a linguistically significant gesture. This question is, by its nature, more problematic than the first one, and the answer is correspondingly harder to find. We believe, however, that the pattern of results obtained in the experiments reported here lend support to the assumption that the effect of silence is, to a significant extent, phonetic. Having presented those data at various places in this paper, we should collect them here.

First, we should consider again the basic fact that silence was an important cue, and then note how difficult it is, given our results, to account for that solely in auditory terms. Thus, we found that the transition cues for the stops were neither appreciably masked nor altered by interaction when, having been isolated from the speech patterns, they were heard as nonspeech chirps. It is also relevant, of course, that, under some conditions, silence was a sufficient cue. There were, in those cases, no other sufficient cues to be masked. It is also telling that silence was effective as a cue only over a limited range, just as we should expect given the assumption that it provides information about a stop closure that lasts for a limited amount of time. Further evidence for a link between perception and production is provided by those of our experiments that showed an equivalence in phonetic perception between duration of silence and duration of friction (or between duration of silence and the rise time of the friction). That result—similar, as we have pointed out, to the results of other investigators—seems easiest to interpret on the assumption that the acoustically different cues give rise to the same phonetic percept because they are normally the correlated (but distributed) acoustic consequences of the same gesture.

Having said that the data of our experiments (and those of others) imply that perception of the silence cue is constrained as if by knowledge of what vocal tracts can do, we should offer a few parenthetical comments about what the data do not imply. First, they most certainly do not imply that a listener can hear only what a vocal tract can do. Indeed, it is for that reason that we have so often added the qualification "when the vocal tract makes linguistically significant gestures." For we know that synthetic speech can be readily perceived (as speech), though it departs, sometimes appreciably, from those acoustic patterns that real vocal tracts can produce. Thus, synthetic patterns sometimes contain only two formants, and the transitions are sometimes made to change direction instantaneously. But such departures, we should note, are not linguistically relevant. Languages cannot enforce a distinction between phones made with two formants and those made with the greater number of formants that real vocal tracts produce, nor can they contrast instantaneous changes in formant slope with those more gradual changes that must characterize the behavior of such real masses as the tongue. In cases like these, an experimenter can take all manner of liberties with the stimulus patterns without destroying or even distorting phonetic perception, provided he manages to include the acoustic information that enables the listener to hear the stimuli as speech. All this is to say that if the speech perceiving mechanism is "tuned" to a vocal tract, as we have implied it might be, then such "tuning" must hold only for those maneuvers that have linguistic significance.

Second, our assumption of a link between perception and production is not meant to imply anything about the nature of the mechanism that mediates the link, or about the relative contributions of nature and nurture to its formation. In regard to the nature of the mechanism, there are aspects of our results (and those of others) that speak against at least one very simple possibility: feature detectors that have evolved in such a way as to be "tuned" to respond to fixed acoustic consequences of articulatory gestures, and to be "sprung" when those consequences are present in the signal. In that connection, we note, first, that the relations among cues that we have found suggest that the setting of one detector (e.g., the silence detector) must be, in effect, variable and conditioned by the "value" of the other cues (e.g., duration of the noise). We should then note that, according to the results of the experiment on identification of syllable-final stops, a detector for the syllable-final transition cues could not respond directly upon sensing these cues, but would, instead, have to wait until it had information about the next syllable. At the least, it would have to know about that next syllable how far removed in time it was from the syllable containing the target phone and what kinds of phones it comprised. The consequence for a detector model is that it loses much of the appeal that it would otherwise have by virtue of its simplicity.

As for questions about the contributions of nature and nurture to the assumed link between perception and production, we should emphasize that such questions stand apart from those that pertain to the existance of such a link. Our experiments bear only on the latter.

We turn finally to the third question: Whose vocal tract is perception linked to? Given the results of our experiments with same and different talkers, we should suppose that the answer is quite clear: the relevant vocal tract is not that of the listener nor is it that of the speaker; it is rather some very abstract conception of vocal tracts in general. But those same results add support to the view that a link to some vocal tract, however abstract, does figure in the perception of speech.

ACKNOWLEDGMENTS

This research was supported by a grant (HD-01994) from the National Institute of Child Health and Human Development to Haskins Laboratories. We wish to thank Bruno Repp for helpful comments on an earlier version of this paper. We also wish to thank Anthony Levas and Suzi Pollack for their assistance in collecting and tabulating portions of the data. Michael F. Dorman is also at Arizona State University; Lawrence J. Raphael at Herbert H. Lehman College and the Graduate School of the City University of New York; and Alvin M. Liberman at the University of Connecticut and Yale University. Address reprint requests to Michael F. Dorman, 270 Crown Street, New Haven, Connecticut 06511.

[1] When native speakers of English produce [ʃpɛ] and [ʃkɛ], [p] and [k] are realized as voiceless inaspirates. It is for this reason that, when the fricative noise is removed from [ʃpɛ] and [ʃkɛ], listeners hear the stops that remain as voiced. In our experiment, it was necessary, therefore, to record [bɛ] and [gɛ] (rather than [pɛ] and [kɛ]), so that, when the fricative noise and vocalic segment were combined, the listeners would hear a normal sounding [ʃpɛ] and [ʃkɛ].

[2] The term "geminate" is ordinarily used to refer to the doubling of a consonant within a word. Such doubling as we find in English occurs only across word boundaries. We nevertheless here use the term, though our subjects were native speakers of English and were accustomed to consonant doubling only at word boundaries.

[3] Since writing this paper, a somewhat similar result by Rudnicky and Cole (1977) has come to our attention. Having recorded [ba ga] they found: (a) that after removing the [ga] their listeners heard [bag], (2) that after replacing the [ga] with [da] placed close in time to the first syllable, listeners heard [bai da], and (3) that when the second syllable was separated from the first syllable by a sufficient interval of silence, listeners heard [bag da]. This result is of particular interest from our point of view because, in the condition when the second syllable [da] was close to [ba] and the subjects heard [bai da], it is clear that the transition cues at the end of the first syllable were not being (backward) masked by the second syllable; they were being perceived, but as a glide to [i] rather than as a stop. That result is similar to the findings of Liberman and Pisoni (1977), referred to earlier in this paper, that ʃ noise placed close to [gɛ] causes listeners to perceive [ʃjɛ].

[4] We have not commented on the difference between the identification function for [b] and [g] because we have found that difference to change, even to be reversed, depending on the surrounding vocalic environment. We emphasize the geminate versus nongeminate contrast because it remains more nearly stable across vowel environments.

[5] Using stimulus patterns and procedures very different from ours, Darwin and Bethell-Fox (1977) have, nevertheless, obtained results that are quite compatible. After synthesizing a pattern that was heard as an uninterrupted sequence of semivowels and vowels, they found that introducing changes in fundamental frequency at appropriate places in the pattern (without changing formant frequencies) caused the semivowels to be heard as stops. Their interpretation was that the rapid shifts in fundamental caused the sequence to "stream," thus permitting the listener to hear two voices; that, in turn, provided the silence necessary to convert semivowel to stop.

Abbs, M. (1971). "A study of cues for the identification of voiced stop consonants in intervocalic contexts," Doctoral dissertation, University of Wisconsin (unpublished).

Bailey, P., Summerfield, Q., and Dorman, M. "Friction duration and friction offset as cues to stop manner in fricative-stop-vowel sequences," (unpublished).

Bastian, J., Eimas, P., and Liberman, A. (1961). "Identification and discrimination of phonemic contrast induced by silent interval," J. Acoust. Soc. Am. 33, 842 (A).

Bastian, J. (1962). "Silent intervals as closure cues in the perception of stops," Haskins Laboratories, Speech Res. Instrum. 9, Appendix F.

Darwin, C. J. (1971). "Ear differences in the recall of fricatives and vowels," Q. J. Exp. Psychol. 23, 46–62.

Darwin, C. J., and Bethell-Fox, C. (1977). "Pitch continuity and speech source attribution," J. Exp. Psychol.; Hum. Perform. and Percept. 3, 665–672.

Erickson, D., Fitch, H., Halwes, T., and Liberman, A. (1977). "Trading relation in perception between silence and spectrum," J. Acoust. Soc. Am. 61, S46 (A).

Fujisaki, H., Nakamuro, K., and Imoto, T. (1975). "Auditory perception of duration of speech and non-speech stimuli," *Auditory Analysis and the Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London).

Ganong, W. (1975). "An experiment on 'phonetic adaptation,'" Research Laboratory of Electronics, MIT, Progress Report 116, 206–210.

Gerstman, L. J. (1957). "Perceptual dimensions for the friction portions of certain speech sounds," Doctoral dissertation, New York University (unpublished).

Harris, K. S. (1958). "Cues for the discrimination of American English fricatives in spoken syllables," Lang. Speech 1, 1–7.

Kuypers, A. (1955). "Affricates in intervocalic position," Haskins Laboratories, Q. Prog. Rep. 15, Appendix 6.

Liberman, A. M., and Pisoni, D. B. (1977). "Evidence in a special speech-processing subsystem in the human," *Recognition of Complex Acoustic Signals*, edited by T. H. Bullock (Dahlem Konfrerenzen, Berlin) Life Sciences Research Rep. 5.

Liberman, A. M., and Studdert-Kennedy, M. (1979) "Phonetic perception," in *Handbook of Sensory Physiology*, edited by R. Held, H. Leibowitz, and H. L. Teuber (Springer-Verlag, Heidelberg), Vol. VIII, "Perception" (in press).

Lisker, L. (1957a). "Closure duration and the voiced-voiceless distinction in English," Language 33, 42–49.

Lisker, L. (1957b). "Closure duration, first-formant transitions and the voiced-voiceless contrast of intervocalic stops," Haskins Laboratories, Q. Prog. Rep. 23, Appendix 1.

Lisker, L. (1977). "Closure hiatus: cue to voicing, manner and place of consonant occlusion," J. Acoust. Soc. Am. 61, S48 (A).

Pickett, J. M., and Decker, L. (1960). "Time factors in perception of a double consonant," Lang. Speech 3, 11–17.

Port, R. (1976). "The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words," Doctoral dissertation, University of Connecticut (unpublished).

Raphael, L. J., and Dorman, M. F. (1977). "Perceptual equivalence of cues for the fricative-affricate contrast," J. Acoust. Soc. Am. 61, S45 (A).

Repp, B. (1976). "Perception of implosive transitions in VCV utterances," Haskins Laboratories, Status Rep. Speech Res. SR-48, 209–234.

Repp, B. (1977). "Perceptual integration and selective attention in speech perception: further experiments on intervocalic stop consonants," Haskins Laboratories, Status Rep. Speech Res. SR-49, 37–70.

Rudnicky, A., and Cole, R. (1977). "Vowel identification and subsequent context," J. Acoust. Soc. Am. 61, S39 (A).

Summerfield, A. Q., and Bailey, P. (1977). "On the dissociation of spectral and temporal cues for stop consonant manner," J. Acoust. Soc. Am. 61, S46 (A).

Truby, H., (1955). "Affricates," Haskins Laboratories, Q. Prog. Rep. 11, 7–8.