1081

The perception of speech gestures

Aimée M. Surprenant[a]
Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907-1364

Louis Goldstein
Haskins Laboratories New Haven, Connecticut 06510 and Yale University, New Haven, Connecticut 06520

(Received 2 April 1996; revised 4 March 1998; accepted 31 March 1998)

Two experiments examined the effects of temporal overlap of speech gestures on the perception of stop consonant clusters. Sequences of stop consonant gestures that exhibit temporal overlap extreme enough to potentially eliminate the acoustic evidence of (at least) one of the consonants were obtained from x-ray microbeam data. Subjects were given a consonant monitoring task using stimuli containing stop sequences as well as those containing single stops. Results showed that (1) the initial consonant in the stop sequences was detected significantly less often than in the single stops; (2) bilabial gestures were considerably more effective at obscuring a preceding alveolar than the reverse; and (3) the detection rate correlated with an index of overlap between lip and tongue tip gestures. Experiment 2 employed stimuli that were truncated during the closure for the critical stop or stop sequence, so as to eliminate any information occurring in the acoustic signal at the stop release. This experiment showed that removing release information decreased detectability of the consonants generally. However, consistent with the observed gestural patterns, removing the release did not decrease detection of the alveolar stop when it was the first consonant of a sequence, indicating that there was no information about the alveolar stop present in acoustic realization of the second stop release. These experiments show that certain gestural patterns actually produced by English speakers may not be completely recoverable by listeners, and further, that it is possible to relate recoverability to particular metric properties of the gestural pattern. © *1998 Acoustical Society of America.* [S0001-4966(98)02707-6]

PACS numbers: 43.71.Es, 43.71.An [WS]

INTRODUCTION

The observation that the acoustic properties of phonological units vary considerably as a function of phonetic context has been the basis for a great deal of research in both speech perception and production. In speech, phonological units are produced in overlapping time frames (Liberman *et al.*, 1967; Fowler, 1986) and, as a result of this coproduction, the acoustic signal bears a complex relation to the perceived units. Because of the lack of context-invariant correspondence between the signal and the percept, some theories of speech perception have looked to places other than an acoustic description of the speech signal to find the invariants that allow us to achieve a stable percept in the face of such contextual dependence. In particular, the revised motor theory of speech perception (Liberman and Mattingly, 1985, p. 2) has identified the "intended phonetic gestures of the speaker, represented in the brain as invariant motor commands," as the units of speech perception.

Although the exact processes whereby the listener recovers the abstract gestures from the speech signal have not been specified in detail, the general idea is that the listener uses (innately specified) phonetic knowledge to find a gestural pattern that could have given rise to a particular acoustic input to the speech system. One question that can be posed from this perspective is, what acoustic conditions are necessary for a gesture to be (correctly) recovered? In the extreme, it is obvious that some evidence of a gesture in the acoustic signal is necessary in order for it to be recovered. As Mattingly (1981) has argued, it is possible to view the syllabic organization of speech gestures as a means by which gestures can be overlapped substantially (so as to maximize "parallel transmission") while still ensuring recoverability. However, it remains a question exactly how subtle the local acoustic consequences of a gesture can be and still lead to correct interpretation.

There do seem to be circumstances in which the listener may fail to recover the gestures intended by (and indeed produced by) the talker. Such circumstances may occur in informal or casual speech, which has traditionally been described as exhibiting deletion or assimilation of consonants in sequences (Brown, 1977). Browman and Goldstein (1990) have presented articulatory movement data showing that, in at least some of these cases, the gesture for a putatively deleted consonant is, in fact, produced by the talker, but is overlapped by surrounding consonant gestures to such an extent that it is perceptually "hidden." An example they present is the final [t] in "perfect," when produced in the fluent phrase "perfect memory." X-ray evidence in this case revealed motion of the tongue tip toward the alveolar ridge and then away from it, comparable in magnitude to the motion found in examples in which the [t] was completely audible (when there was a prosodic boundary between "perfect" and "memory"). In the fluent phrase, the closing and releasing motions of the tongue tip occur while the vocal tract is completely closed by the overlapping dorsal ([k]) and

[a]Electronic mail: aimee@psych.purdue.edu

labial ([m]) closure gestures. Browman and Goldstein (1991) have hypothesized that recovery failures of this kind (and others) can ultimately lead to "listener-based" historical sound changes (of the kind proposed by Ohala, 1981). If such perceptual "hiding" of overlapped gestures can be demonstrated systematically, it can provide evidence relevant to theories of speech perception that use the gesture as the unit of perception. In particular, it suggests that there are limiting cases where the nature of the acoustic output resulting from a pattern of overlapping gestures is such that recovery of the gestures by the listener is difficult, if not impossible.

The assimilations and deletions reported for fluent speech have primarily been observed by phoneticians through "careful listening." Taking the deletion of [t] in "perfect memory" as an example, what such listening actually reveals is that there is no local information in the signal that sounds like the formation or release of a tongue tip constriction. However, the word "perfect" might still be expected to sound completely normal to a naive listener, in context, and thus it is possible that the listeners are still recovering the "hidden" gesture, perhaps using knowledge of higher-order gestural dependencies. For example, the precise degree of temporal overlap between the dorsal and labial gestures (and their relations to other gestures in the utterance) might be different if the intended utterance had no tongue tip closure gesture at all (i.e., had the speaker intended to say "perfeck memory"). If that were the case, listeners *could* infer the existence of the tongue tip gesture from the pattern of overlap, even if there was no local acoustic indication of its formation or release. One way to see whether a potentially "hidden" gesture is actually recovered or not is to present listeners with utterances that differ minimally in whether that gesture is part of the "intended" phonetic structure of the speaker, and see if listeners can distinguish them.

Such a study was undertaken by Byrd (1992). She used the Haskins gestural model (Browman and Goldstein, 1990; Saltzman and Munhall, 1989; Rubin *et al.*, 1981) to create sets of stimuli that varied in the amount of overlap between two consonant gestures across a word boundary ("bad ban" and "bab dan"). The overlap between C_1 and C_2 was varied over 11 steps from 34% to 107% (the latter indicating that C_2 actually preceded C_1 slightly). These stimuli were presented to subjects in a forced-choice identification task in order to assess the perceptual effects of the overlap between gestures. The effect of overlap on the identification of a consonant was measured in two contexts, bilabial#alveolar ("ba$\underline{b}$ $\underline{d}$an") and alveolar#bilabial ("ba$\underline{d}$ $\underline{b}$an"). The size of the gestures was held constant in all cases. The results showed that as the amount of overlap increased, identification of C_1 was significantly reduced (C_1 was perceived as assimilated to C_2). This effect was stronger and appeared with a smaller amount of overlap in the [db] condition compared to the [bd] condition. It appears that the tongue tip gesture was more easily hidden by the subsequent labial gesture than vice versa. In a second condition, Byrd (1992) found similar results when the stimuli were truncated during the medial stop closure. However, in this truncated case, the effects were weaker and occurred only with a greater amount of overlap compared to the condition where the context word was present. The stronger effects obtained in the two-word condition are consistent with other results (Repp, 1978; Dorman *et al.*, 1979; Ohala, 1990) showing that, as stop closure duration in a VC_1C_2V decreases, only one consonant is perceived, and that CV information dominates over VC information in determining the identification of that consonant.

Byrd's experiment demonstrates that gestures may, in fact, fail to be recovered under conditions of extreme overlap. However, her study employed synthetic speech generated by a gestural model (as is true of similar experiments in German by Geumann and Kröger, 1995), and it is not clear that her stimuli preserve subtleties of gestural organization, from which listeners might recover an overlapped gesture, as discussed above. Nolan (1992) used naturally produced materials for a perceptual study of assimilation. He examined electropalatographic (EPG) production data for utterances with sequences of alveolar-velar consonants across word boundaries, such as "road collapsed." In addition to examples produced with complete alveolar closure, he found examples in which the EPG indicated that the tongue tip gesture is reduced in magnitude so as to no longer produce a complete closure ("residual-alveolar") and examples in which the tongue tip gesture is so reduced that there is no apparent tongue tip contact at all ("zero-alveolar"). Listeners were asked to identify these utterances and a set of control utterances in which the tongue tip gesture is not part of the phonetic intent (e.g., "rogue collapsed"). Listeners correctly identified the "residual-alveolar" utterances with a greater than chance frequency, although the error rate was higher than for the examples with complete alveolar closure. "Zero-alveolar" utterances were not distinguished from the controls. Thus it appears that a tongue tip gesture is not recovered by the listener when it is reduced in magnitude by a certain degree. However, although Nolan suggested that there is a continuous change from full to nonexistent gestures, due to limitations of the contact measures employed, he was unable to systematically demonstrate continuous changes in perception with changes in degree of closure of the tongue tip. In addition, Nolan did not measure temporal overlap of the gestures involved, and thus, the study has nothing to say about the possibility of gestural hiding.

Taken together, the experiments by Nolan (1992) and Byrd (1992) indicate that the perceived assimilation can result from the failure to recover an intended gesture and that recoverability may be affected by at least two factors, the size of the gesture and the amount of overlap between two adjacent gestures. The two experiments reported here combine both factors and investigate listeners' responses to stimuli with naturally produced consonant sequences that show substantial gestural overlap. The stimuli, taken from an x-ray microbeam study, were chosen in order to maximize variation in overlap as well as in the size of the relevant gesture. X-ray data provide a means for quantifying the production data more precisely than would be possible in electropalatographic studies. The experiments were also designed to test the hypothesis (Browman and Goldstein, 1990, 1992; Byrd, 1992) that tongue tip gestures are more sensitive to such hiding than lip gestures, a hypothesis that could account for the fact (Byrd, 1992) that in the phonology of many languages, coronal consonants assimilate more commonly to noncoronals than vice versa.

I. EXPERIMENT 1

The purpose of experiment 1 was to explore how the detection of a stop consonant gesture is affected by its size (as actually produced by a speaker) and the degree to which it is overlapped by a subsequent stop gesture. We compared the detection of stop consonant gestures when they are followed by a second (partially overlapping) stop gesture (''stop sequence'' condition) to the detection of stop gestures not immediately followed by another stop (''single stop'' condition) and then performed statistical analyses to investigate the potential contributions of gesture magnitude and overlap to the detectability scores. Natural speech tokens that were tracked by x-ray microbeam technology were used to measure the actual movements of the articulators.

A. Methods

1. Subjects

Subjects were 22 Purdue University undergraduates who volunteered to participate in exchange for credit in introductory psychology courses. Data from two subjects were discarded because they were not native speakers of American English.

2. Stimuli

The stimuli for the perceptual experiments were utterances for which articulatory data had been collected using the NIH x-ray microbeam system at the University of Wisconsin (Nadler *et al.*, 1987). Gold pellets were placed in a number of locations on the subject's tongue, jaw, and lips. The subject (a male student at the University of Wisconsin, who grew up in California) then read a set of three-word sentences (printed in normal English orthography), each starting with the word ''my'' and followed by a CVC word such as ''top.'' The CVC words had either voiceless labial (''p'') or voiceless alveolar (''t'') stops as the final consonant. The third word was a two-syllable verb (or nonsense verb) ''_uddles'' with the blank being filled by ''p,'' ''t,'' ''c,'' ''h.'' The subject was prompted to place contrastive accent on one of the three words of the phrase, by means of a contextualizing sentence that was displayed to the subject, but not read aloud, and by capitalizing the to be accented word. In the course of the experiment, all three stress patterns were recorded. Each phrase was produced five times in succession by the same male speaker and the movements of the pellets in relation to the fixed parts of the vocal tract were tracked by the system and synchronized with the acoustic signal. Analysis of gestural magnitude in these utterances, and in particular, the difference in magnitude between the initial and final C's in the CVC word can be found in Browman and Goldstein (1995).

TABLE I. Stimuli used in the experiments. The stimuli for experiment 2 were identical to those of experiment 1 except that the stimuli were truncated during the acoustic closure following the initial CVC word.

Target		Present	Absent
''t''	stop sequence	tot puddles	top cuddles
	single stop	tot huddles	top huddles
''p''	stop sequence	top tuddles	tot cuddles
	single stop	top huddles	tot huddles

The stimuli for the present study were chosen to compare perception of stimuli in which the oral gesture for a final stop is overlapped in time by the oral gesture for the stop consonant that begins the following word (''stop sequence context'') with stimuli in which there is no oral stop gesture at the beginning of the following word, and therefore the final stop gesture is not overlapped by another stop (''single stop context''). In addition, they were designed to compare the effect of overlap of labials on alveolars to the effect of alveolars on labials (e.g., ''tot puddles'' versus ''top tuttles''). All three stress patterns were used. The stimuli chosen for each monitoring condition are shown in Table I above. An equal number of stimuli that lacked the target (*Absent* column in Table I) were chosen as filler items. They also occurred in two forms: those involving a stop consonant sequence (''cuddles''), and those involving a single stop (''huddles'').

To choose particular stimulus tokens from the recorded x-ray material we examined the movements of pellets on tongue tip and lips, in particular, the vertical position of the tongue tip pellet (tongue tip height) and the vertical distance between the upper and lower lip pellets (lip aperture). Movements of these parameters have been used (Browman and Goldstein, 1988, 1995) to infer the timing (and size) of tongue tip closure gestures (e.g., in ''tot''), and bilabial closure gestures (e.g., in ''puddles''). An example of an output from the x-ray microbeam system is given in Fig. 1. The top of the figure shows the time functions of tongue tip height and lip aperture, while the bottom shows the two-dimensional positions of the relevant pellets) at a particular point in time (shown by the arrow just below the time axis),

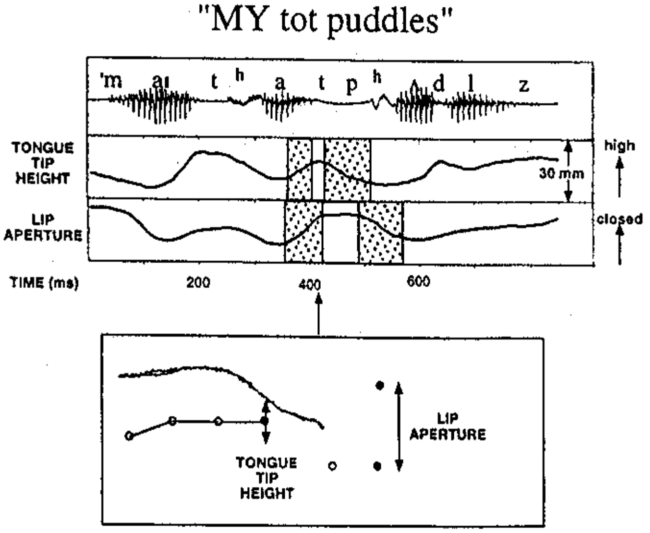

FIG. 1. An example of the output from the x-ray microbeam system. See text for further explanation.

with respect to a tracing of the subject's hard palate. The relevant pellets (one on the tongue tip and one on each lip) are represented by filled circles. The unfilled ones represent other pellets on the tongue and lower teeth. The subject was saying "MY tot puddles." For both the tongue tip and lip closure gestures, three kinematic events associated with each gesture are shown. The first event, closing movement, is marked by the first shaded region; the second, held constriction peak or plateau, is the clear region between the two shaded regions; and the third, opening movement, is marked by the second shaded region. Note that, in this example, the closing movements of the two gestures overlap almost completely. For the perceptual experiment, we chose those three tokens (of the five produced in succession for each stress pattern) that exhibited a range of variation in both the magnitude of the tongue tip gesture (amount of movement shown by the closing and opening events) and degree of overlap between tongue tip and lip closing movements.

Stimuli were digitized from videotape (recorded during the x-ray session) onto an Apple Macintosh computer, using eight-bit digitization and a sampling rate of 22 kHz. The initial word ("my") was excised (up to the silence for the initial stop in the CVC word) resulting in stimuli that ranged from approximately 600 to 800 ms in length.

The two monitoring conditions (monitor for "t" or "p") were blocked and presented in a counterbalanced order to each subject. Three different tokens of each of the four utterances for each of three different stress patterns were selected and presented four times for a total of 144 trials in each condition, half of which contained the target and half of which did not.

3. Apparatus

Stimuli were presented by an Apple Macintosh IIci computer, and played through Sony MDR-V5 headphones. Subjects were permitted to adjust the level of the sounds, during the practice phase, to a comfortable setting. Subjects responded by pressing a button on an external hardware box.

4. Procedure

The task used in this experiment was a phoneme monitoring (vigilance) task in which subjects were asked to press a button as quickly as possible when they heard a specific consonant. This task has been used extensively to explore the time course of word recognition (Foss and Blank, 1980; Marslen-Wilson and Tyler, 1980). The dependent variable in this type of experiment is usually response time. However, because we were mainly interested in determining whether the listeners heard the target, accuracy, rather than speed, was stressed to the subjects.

The targets ("t" and "p") were presented in blocks and subjects were instructed to monitor for a different target after the first block. The following instructions were read to the subjects:

"You will be hearing nonsense utterances that consist of two words. What we would like you to do is to press a button as quickly as you can when you hear the sound ("tee as in tot"/"pee as in top"). You will be hearing only four possible sounds (t condition—tot puddles, top cuddles, tot huddles, top huddles; p condition—top tuddles, tot cuddles, top huddles, tot huddles). If you do not hear the sound ("tee"/"pee") in the sequence, just don't press the button and wait for the next trial. If you do hear it, press the button as quickly as you can. Half of the sounds will have a ("tee"/"pee") in them and half will not."

In the "t" condition, subjects were given the following additional instructions:

"There will be a "tee" sound beginning each utterance but you are to ignore that and decide if there was a "tee" anywhere else in the word."

There were ten practice trials before each block. On any given trial a message appeared in the center of the screen informing the subject that a trial was about to begin. After 100 ms, a letter appeared in the middle of the screen as a reminder of the target, and the stimulus was presented over the headphones. Subjects either pressed a button indicating that they had heard the target or waited for the next trial. Two seconds later the next trial began. The stimuli were presented in a new random order for each subject and the two target conditions were blocked and counterbalanced across subjects.

5. Stimulus measurements

For each stimulus token, the magnitude of the gestures and the amount that they overlapped were measured, so that these variables could be used as predictors of subjects' perceptual responses. The measurements were based on the three kinematic events associated with each gesture, as shown in Fig. 1. These events were detected automatically by finding curve extrema (peaks and valleys) with a noise band of 1 mm, that is, extrema plateaux within which the pellets move less than 1 mm. In Fig. 1, the held constriction peaks for tongue tip height and lip aperture were plateaux detected in this way. The closing movement for a stop gesture begins at the right edge of the preceding vowel's plateau and continues to the left edge of the constriction plateau for the stop gesture. The opening movement begins at the right edge of the stop's constriction plateau and continues to the left edge of the plateau for the following vowel. Once these events had been detected, the eleven magnitude and overlap measures listed in Table II were calculated. The first nine characterize properties of the individual gestures, and the last two characterize the temporal overlap between the gestures.

B. Results

The data were analyzed in two ways. First, we divided responses into categories of "correct" and "incorrect" based on what the speaker was asked to produce (and presumably intended to produce). Thus, if the speaker in the x-ray microbeam study was reading "tot puddles," the "correct" response in the "t" condition would be a button press, i.e., the consonant was detected. Then, a simple analysis of variance was performed on the number of correct detections. This analysis is presented in the next section. However, given that we have measurements on continuous movements of the gestures, we can also use correlation analyses to relate

TABLE II. Gestural variables correlated with perceptual responses.

Lip Aperture (LA)	Vertical distance between upper and lower lips during constriction plateau (mm)
Δ Lip Aperture Closing (Δ LA Closing)	Change in Lip Aperture from beginning of closing movement to constriction plateau (mm)
Δ Lip Aperture Opening (Δ LA Opening)	Change in Lip Aperture from constriction plateau to the end of opening movement (mm)
Duration of LA Constriction Plateau (LA Duration)	Time from end of LA closing movement to the beginning of LA opening movement (ms)
Tongue Tip Height (TT)	Vertical position of tongue tip during constriction plateau (mm)
Δ Tongue Tip Closing (Δ TT Closing)	Change in tongue tip height from beginning of closing movement to constriction plateau (mm)
Δ Tongue Tip Opening (Δ TT Opening)	Change in tongue tip height from constriction plateau to end of opening movement (mm)
Duration of TT Constriction Plateau (TT Duration)	Time from end of TT closing movement to the beginning of TT opening movement (ms)
Duration of Cl Closing Movement	Time from beginning of closing movement (TT or LA, as appropriate) to constriction plateau (ms)
Closing Lag	Time between end of TT closing movement and end of LA closing movement (ms)
Opening Lag	Time between beginning of TT opening movement and beginning of LA opening movement (ms)

the relative sizes of the gestures and the amount of their overlap to the consonant monitoring responses. These analyses are reported in the subsequent section.

1. Analysis of variance

Overall, subjects were performing the basic task accurately. Figure 2 shows the mean proportion of consonants detected as a function of utterance type for the "t" and "p" conditions. In addition, the figure shows the proportion of consonants incorrectly detected when the stimulus was not present (filler). As is evident in the top panel of the figure, the proportion of "t" responses to stop sequences was substantially lower than to the single stop stimuli. This difference is much smaller in the "p" condition (bottom panel). In both conditions there were few false alarms to the filler items.

With these data, we can perform a signal detection analysis. However, because in some conditions we have a large number of perfect discriminations which produce a false alarm rate of 0 or a hit rate of 1.0, or both, we used the nonparametric analog of d', A' (Pollack, 1970). This measure ranges from 0 to 1 where chance responding is 0.5. The resulting A' data were analyzed by means of analysis of variance, with factors corresponding to condition ("t" vs "p"), context ("stop sequence" versus "single stop"), accent, and order of production.

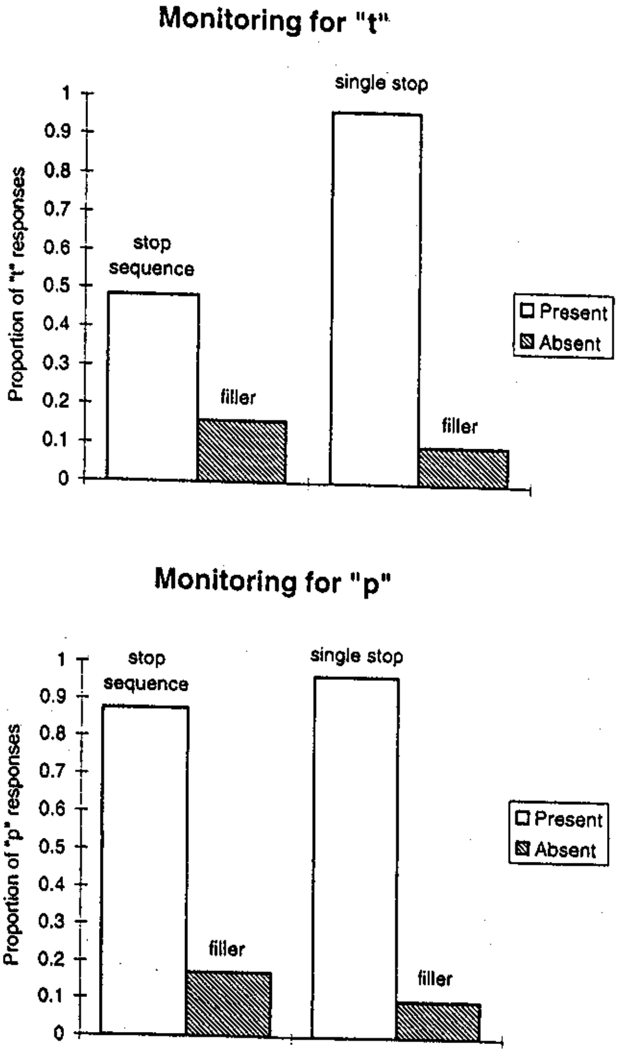

FIG. 2. Proportion of "t" (top panel) and "p" (bottom panel) responses as a function of presence or absence of the target for experiment 1.

When the to-be-detected consonant was actually present, subjects responded more than when it was not present [$F(1,19)=95.31$, MSe=0.05, $p<0.01$]. In addition, the overall detectability (A') in the "p" condition (0.93) was greater than in the "t" condition (0.83), $F(1,19)=42.46$, MSe=0.04, $p<0.01$. That main effect is largely due the interaction of context and condition [$F(1,19)=41.09$, MSe =0.04, $p<0.01$]: In the "p" condition, there was a small effect of context with an A' of 0.89 for the stop sequence items compared to 0.70 for the stop sequence items in the "t" condition. In both conditions A' was 0.96 for the single stop targets.

For the present purposes, the main variable of interest is the difference in responses to the stop sequences compared to the single stops. Planned contrasts show that the effect of context was statistically significant in both the "p" [$F(1,19)=16.71$, MSe=0.02, $p<0.01$] and the "t" [$F(1,19)=91.22$, MSe=0.06, $p<0.01$] conditions. In addition, item analyses (see Clark, 1973, for a rationale and formulas for using MinF') show both effects to be reliable ["p"—$F(1,38)=6.33$, $p<0.05$; "t"—$F(1,38)=61.46$, $p<0.01$] although the MinF' was reliable only in the "t"

TABLE III. Experiment 1. Spearman's correlation coefficients for ranked data (r_s) between gestural variables and detections of the target (**$p<0.05$; *$p<0.10$). First row—"tot puddles," second row—"top tuddles."

	LA	Δ LA closing	Δ LA opening	LA duration	TT	Δ TT closing	Closing duration	Δ TT opening	TT duration	Closing lag	Opening lag
t	−0.19	−0.76**	0.71**	0.07	0.52	−0.10	−0.23	0.63*	0.46	0.64*	0.74**
p	−0.01	0.26	−0.05	0.36	0.20	0.17	0.15	0.21	0.09	0.45	0.19

condition ["p"—Min$F'(1,54)=3.76$, $p=0.06$; "t"—Min$F'(1,54)=35.37$, $p<0.01$].

Each token was spoken by the subject in the x-ray microbeam study in three possible stress patterns; either the initial, the middle, or the final word of the phrase was stressed ("MY," "CVC," or "_UDDLES"). Although we had made no *a priori* predictions about the effect of stress pattern on detection of a particular consonant, the interactions of condition, stress and context were interesting. There was a main effect of stress, with detectability increasing as the stress moved from the first to the last word [$F(2,19)=14.41$, MSe$=0.02$, $p<0.05$]. All of the two-way interactions with stress were reliable and can be interpreted by examining the three-way interaction of condition, stress, and context [$F(2,38)=10.16$, MSe$=0.02$, $p<0.05$]. This interaction can be attributed to the effect of stress being different for the "t" and the "p" conditions. In both conditions, there was a ceiling effect for the single stop stimuli with all stress conditions having an A' greater than 0.95. However, in the stop sequence conditions, when the target was "t," there was an increase in detectability from the first to the final word stressed (0.64, 0.65, 0.81, respectively) but when the target was "p" that increase occurred when the second word was stressed (0.83, 0.94, 0.91). The reduced effect of context, resulting in increased detectability, for "t" when the final word is stressed could be due to the fact that, in the production subject's data, final alveolar stops are sharply reduced in magnitude when accent is on the first or second words, but they are not as reduced when the accent is on the final word (Browman and Goldstein, 1995).

The final variable of interest was the order in which the tokens were initially produced. Each phrase was produced five times in rapid succession and, because deletions and assimilations generally increase in rapid speech (Browman and Goldstein, 1995), it is likely that the amount of assimilation or consonant deletion increased from the first utterance to the last. The main effect of order was statistically significant with detections decreasing as the utterance was pronounced in the beginning to the end of the trial [$F(2,38)=3.65$, MSe$=0.02$, $p<0.05$]. The three-way interaction among condition, context and order was also reliable [$F(2,38)=3.67$, MSe$=0.07$, $p<0.05$]. As with the stress variable, there was a ceiling effect in the single stop condition, with delectability greater than 0.95 for all orders of pronunciation. For the stop sequence conditions, detectability decreased with order of pronunciation for the "t" (0.78, 0.66, 0.66) condition but not for the "p" condition (0.90, 0.88, 0.89).

One other effect was statistically significant: There was a reliable three-way interaction involving context, stress, and order [$F(2,38)=4.23$, MSe$=0.02$, $p<0.01$] due primarily to a lack of effect of stress and order in the single stop conditions.

2. Correlational analyses

The analyses reported above indicate that alveolar stops were detected significantly less often in the stop sequence condition than in the single stop condition. Given the availability of articulatory data on different tokens used in the experiment, we wanted to see to what extent variation in the detectability of these alveolar consonants in the stop sequence condition could be predicted from variation in the gestural pattern: the spatial extent of the alveolar gesture, the spatial extent of the overlapping labial gesture, and the degree of overlap between the two.

Each token of "tot puddles" was analyzed separately and gestural extent and overlap was characterized in terms of the factors in Table II. These factors were correlated, using Spearman's correlation coefficient for ranked data (r_S) with the mean number of target detections for each utterance. The correlations between the detection of the target and articulatory factors for the utterances containing "t" are shown in the top line of Table III. It is evident that factors characterizing the size of both gestures and the degree of overlap help determine detectability, although the correlations reach significance only for the three largest correlations: Opening Lag (the time between the beginning of lowering of the tongue tip and the beginning of opening of the lips—$r_S=0.74$, $p<0.05$), Δ LA closing (the change in lip aperture from the beginning of the closing movement to the constriction plateau—$r_S=-0.76$, $p<0.05$), and Δ LA opening (the change in lip aperture from the constriction plateau to the end of the opening movement—$r_S=0.71$, $p<0.05$).

The significant correlation of Opening Lag and detection supports the role of gestural overlap in detectability. The longer this interval (indicating greater separation, and thus less overlap between the gestures), the greater the detectability of the target. The correlations involving lip aperture are also consistent with the hypothesis that the tongue tip gesture may be "hidden" by the lip gesture. The change in lip aperture (LA) during lip closing correlated negatively with detectability: larger lip movements resulted in reduced detectability. The change in LA during opening also correlated with detectability, but with the opposite sign. The reason for the reversal of sign seems to be that the value of LA attained during the "p" closure is fairly stable across all tokens, and that differences in LA change are associated with different values of LA during the preceding and following vowels. The LA during the vowels of the words "tot" and "puddles" are negatively correlated with one another because of stress: when stress is on "tot" LA is much wider for the vowel in "tot" than for the vowel in "puddles." The reverse is true when "puddles" is stressed. Thus if larger changes of LA during the closing movement help to reduce detectability of "t," then changes of LA during opening will, necessarily, be correlated positively with detectability.

The size of the tongue tip gesture, as measured by the height of the tongue tip at gesture peak (TT—$r_s = 0.52$, $p = 0.14$), or by the change in tongue-tip height during opening (Δ TT Opening—$r_s = 0.63$, $p = 0.06$), also showed substantial positive correlations with target detection, although these did not reach significance. The larger the tongue tip gesture, the greater the likelihood of it being perceived.

Correlational analyses of "p" responses with properties of the gestures in "top tuddles" utterances (shown in the second line of Table III) were not very revealing. This is because the range of responses was quite restricted (0.74–0.98). The comparable range for the proportion of "t" responses is 0.18–0.74. The actual Spearman rank-order correlations of each factor and number of responses are included in Table III; none were even marginally reliable.

C. Discussion

Experiment 1 showed, using natural speech tokens, that a stop gesture was less likely to be detected when it was followed by an overlapping stop gesture than when there was no following stop. This result, by itself, is consistent with previous studies which show that the release a of stop gesture is more perceptually informative than its closing movement (Malecot, 1958; Wang, 1959; Repp, 1978; Fujimura *et al.*, 1978; Dorman *et al.*, 1979; Ohala, 1990). In addition, however, for the t#p condition of the present experiment, we were able to use correlations between articulatory measures and detection rates to show that token-to-token variation in detectability was related to the degree of overlap of the two closure gestures and to the size of the potentially interfering (or "hiding") closing movement for the second stop. The greatest predictor of detection of a final "t" was the time between the end of the tongue tip closure and the end of the lip closure. In other words, the larger the temporal distance between the gestures in production (as indexed by this measure), the more likely the first was to be detected. The size of the lip gesture (Δ LA Closing, Δ LA Opening) also correlated significantly with detectability of "t."

It is worth considering why it is that Opening Lag is the overlap index that best predicts detectability (better than Closing Lag). This correlation would seem to point to the perceptual importance of the tongue tip opening movement, yet this is not likely the case for these stimuli, since in every case the tongue tip gesture is always acoustically "unreleased." That is, there is no release burst associated with the lowering of the tongue tip. This is due to the fact that for every stimulus the beginning of tongue tip lowering occurs during the held LA constriction, that is, when the lips are closed. In fact, virtually all of the TT opening is completed by the time the lips begin to open. Thus the perceptually relevant information for the tongue tip gesture in this context must be the closing movement of the tongue tip. (This will be tested directly in experiment 2, in which the stimuli are truncated during the acoustic closure, so that only the closing movements can be perceptually relevant.) Detectability may be hindered to the extent to which lip closure is also taking place during this tip closing interval, and so the ideal predictor of detectability ought to be how much acoustically relevant lip closure is taking place during the tip closure interval. Since the onsets of tongue tip and lip movements are generally synchronous, the length of time between end of closing movements (Closing Lag) might have been expected to be a good approximation to this ideal predictor. However, the movement data show that the end of closing movement is ambiguous for LA. There is typically a long shallow slope atop the lip plateau, and the automatically determined point for the end of closing movement can be located at a variety of points along that plateau. On the other hand, the beginning of lip opening is clear in these signals, and well captured by the automatically determined point. Thus if there is a systematic relation between the time of opening of the lips and the "true" (and acoustically relevant) end of the lip closing movement (not an unreasonable assumption), Opening Lag may be providing a better measure of acoustically relevant overlap during the tip closing interval than is Closing Lag.

Another important finding of the present experiment is the asymmetry between the t#p and p#t utterances. The effect of a following stop on detectability was much greater for the t#p utterances. Byrd (1992) reports a similar finding with her synthesized stimuli, although the extent of the asymmetry does not appear to be as strong as in the present experiment. A number of explanations for such an asymmetry can be considered. One possible explanation that Byrd gives is that the formant frequencies that result (just before complete closure) from exactly simultaneous lip and tongue tip constrictions are closer to those that result from a lip closure alone than those that result from a tongue tip closure alone. This explanation cannot be tested directly here, as we have no stimuli in which the two gestures are perfectly synchronous.

Browman and Goldstein (1990) suggest an explanation for this asymmetry based on the kinematic properties of tongue tip and lip gestures. In particular, they note that tongue tip gestures have been found (e.g., by Kuehn and Moll, 1976) to involve higher velocities of movement than lip or tongue dorsum gestures and they hypothesize that a faster gesture might be easier to hide perceptually. However, Kuehn and Moll's observations of nonsense VCVC utterances do not appear to extend to the contexts tested here, as the mean of the average closing velocities for the tongue tip in t#p utterances was 141 mm/s, and that for Lip Aperture in p#t utterances was 198 mm/s.

Another hypothesis for this asymmetry implicates the role of acoustic closure duration. As is well known (see, e.g., Repp, 1978; Dorman *et al.*, 1979; Ohala, 1990), when closure durations of VCCV utterances are (artificially) shortened beyond the values typically observed in natural speech, listeners report hearing only a single consonant, usually the second of the two. Thus it would be important to determine if the t#p and p#t utterances in the present experiment differed in closure duration. However, the average closure du-

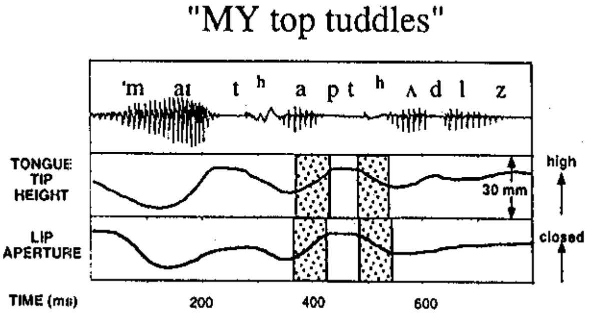

FIG. 3. Example of a token of "MY top tuttles." See text for further explanation.

ration was, in fact, about the same in the p#t (87 ms) and the t#p (85 ms) utterances.

Another possible explanation implicates the role of overlap during the release phase of the two closure gestures. As noted above, in the t#p stimuli the tongue tip release movement is already completed by the time of release of the lips. This can be seen in the example token shown in Fig. 1. However, the parallel state of affairs does not hold for the p#t stimuli. Figure 3 shows the time functions of tongue tip height and lip aperture in an example token of "MY top tuddles." Here, the lips are just starting to open as the tongue tip is lowered. Thus there is potential lip information available at the release of p#t stimuli, but no tongue tip information available at the release of t#p stimuli. To the extent that this is a contributing factor to the obtained asymmetry, we would expect that if the stimuli were truncated during the closure, the asymmetry between the conditions would be reduced, as the p#t condition responses would become less detectable (because information at release is removed). Experiment 2 tests this prediction.

II. EXPERIMENT 2

Experiment 2 was an exact replication of experiment 1 with one change: The utterances were truncated at the closure at the end of the CVC word. As argued above, the release information explanations for the order asymmetry obtained in experiment 1 would predict that the asymmetry will be reduced when the stimuli are truncated.

A. Methods

1. Subjects

Twenty different Purdue University undergraduates volunteered to participate in exchange for credit in introductory psychology courses. All were native speakers of American English.

2. Stimuli

The stimuli for experiment 2 were identical to those of experiment 1 except that they were truncated during the closure interval at the end of the CVC word. This was done by digitally cutting the stimulus in this silent interval, resulting in stimuli that ranged from approximately 200 to 300 ms in duration.

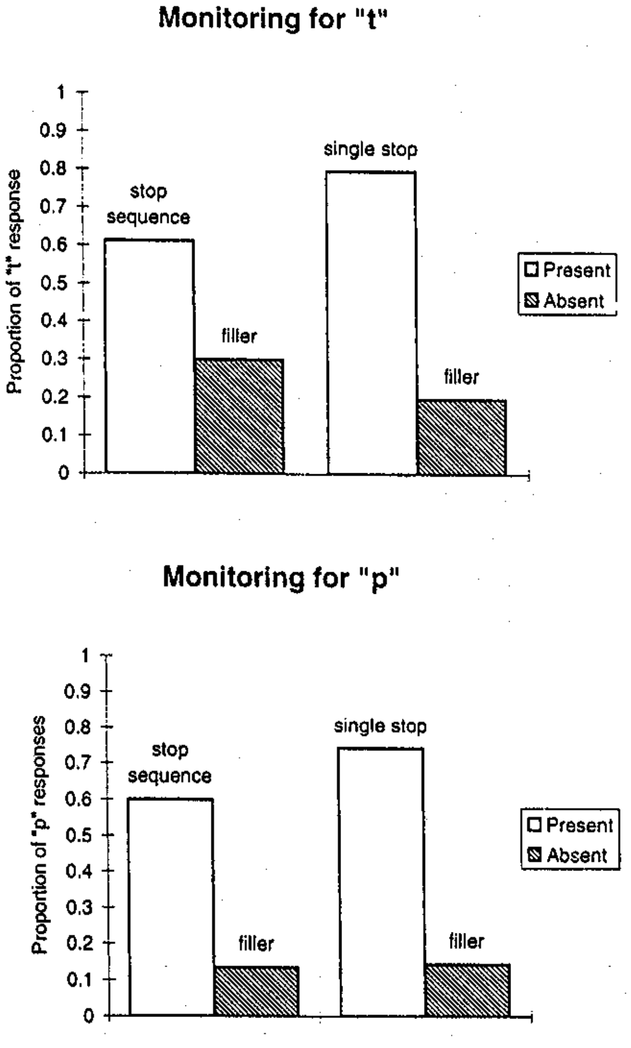

FIG. 4. Proportion of "t" (top panel) and "p" (bottom panel) responses as a function of presence or absence of the target for experiment 2.

B. Results

1. ANOVAs

As in experiment 1, we first divided responses into categories of "correct" and "incorrect" based on what the speaker was asked to produce (and presumably intended to produce). Thus if the speaker in the x-ray microbeam study was reading "tot puddles," the "correct" response in the "t" condition would be a button press, i.e., the consonant was detected. Although subjects were performing the basic task accurately (see Fig. 4), there were more false alarms to the filler utterances (dark bars), especially in the "t" condition (top panel) than in experiment 1.

As in the previous experiment, we transformed the hits and false alarms into A' for the statistical analyses. Unlike in experiment 1, there was no reliable difference overall between "t" and "p" responses [$F(1,19)=4.00$, MSe=0.07, $p=0.06$]. The main effect of context was significant [$F(1,19)=38.2$, MSe=0.05, $p<0.01$] as was the interaction of context and condition [$F(1,19)=5.9$, MSe=0.05, $p<0.05$]. Looking at the means, it seems that the size of the context effect was still larger in the "t" condition (0.70, 0.84) than in the "p" condition (0.78, 0.84). The difference

TABLE IV. Experiment 2. Spearman's correlation coefficients for ranked data (r_s) between gestural variables and detections of the target (**$p<0.05$; *$p<0.10$). First row—"tot [puddles]," second row—"top [tuddles]."

	LA	Δ LA closing	Δ LA opening	LA duration	TT	Δ TT closing	Δ TT opening	Closing duration	TT duration	Closing lag	Opening lag
t	−0.23	−0.62*	0.54	0.17	0.51	−0.03	0.60*	−0.19	0.26	0.41	0.82**
p	−0.07	0.70**	−0.63*	0.46	−0.05	0.49	−0.37	−0.20	−0.50	−0.71**	−0.46

in the size of the context effect is, however, much smaller than for experiment 1 (see below). In addition, item analyses show context to be significant for both conditions ["p"—$F(1,38)=5.28$, $p<0.05$; "t"—$F(1,38)=18.37$, $p<0.01$] although the MinF' was reliable only in the "t" condition ["p"—Min$F'(1,54)=3.64$, $p=0.06$; "t"—Min$F''(1,54)=10.98$, $p<0.01$].

There was a main effect of stress [$F(2,38)=7.51$, MSe $=0.05$, $p<0.05$] with detections increasing from the first to the second word stressed but no appreciable difference between the second and third word stressed (0.75, 0.82, 0.81). As in experiment 1, there was little effect of stress on detection of the single stop stimuli (0.81, 0.84, 0.89 for the three stress levels) but there was an increase in detection as a function of stress for the stop sequence stimuli (0.68, 0.80, 0.73). This is evident in the interaction of context and stress [$F(4,76)=2.80$, MSe$=0.02$, $p<0.05$]. There was also an interaction of stress and production order with the largest effect of order being on the initial stress pattern with decreasing effects of order on the second and third position stress patterns. Finally, there was a three-way interaction of condition, overlap, and stress [$F(2,38)=9.34$, MSe$=0.05$, $p<0.05$]. Inspection of the means reveals that this is due to an actual decrease in detectability of the "p" segment in the single stop condition as compared to the stop sequence context when the stress is on the word "top" in the stimulus.

Finally, there was a main effect of order with detections decreasing from the first to the third repetition produced [$F(2,38)=8.08$, MSe$=0.04$, $p<0.05$]. In addition to the effects described above, pronunciation order also interacted with the monitoring condition [$F(2,38)=17.80$, MSe $=0.03$, $p<0.05$] with a larger effect of order emerging in the "t" condition than in the "p" condition. There was also significant interaction of order and context [$F(2,38)=6.53$, MSe$=0.03$, $p<0.05$] with a larger effect of order on the stop sequence as compared to the single stop stimuli.

In order to test the hypotheses about the basis for the t#p—p#t asymmetry observed in experiment 1, it is useful to compare the results of the two experiments quantitatively. Because the two experiments are identical in design as well as the number of subjects, it is possible to consider them to be two conditions of a between-subjects design and compare the experiments directly in a statistical analysis.

The purpose of the following analysis is to compare the two experiments in terms of the two different types of context and the two monitoring conditions so we collapsed across stress pattern and order. Thus we performed a three-way mixed ANOVA with experiment (between-subjects), monitoring condition ("t," "p"), and context (single stop, stop sequence) as the factors. All of the main effects and interactions were statistically significant with no F value below 5.0 and no p value of less than 0.05. The significant main effect of experiment indicates that subjects performed generally worse in experiment 2, as would be expected based on the removal of release information. However, the effect of experiment was not uniform, as indicated by the significant interactions. It is evident comparing Figs. 2 and 4 that the t#p—p#t asymmetry is weaker in experiment 2 (as predicted by the released information hypothesis), and the significant three-way interaction can be interpreted as confirming this reduction statistically.

To further evaluate the release information hypothesis, we examined the two-way experiment by context interaction separately for "t" and "p" (which is justified in light of the significant three-way interaction). The two-way interaction was significant in the "t" condition [$F(1,38)=8.09$, MSe $=0.007$, $p<0.01$] but did not approach significance in the "p" condition [$F(1,38)<1$]. Performance in both contexts of the "p" condition fell by the same amount. In the "t" condition, however, detection performance decreased only in the single stop context and stayed almost exactly the same in the stop sequence (t#p) context.

This pattern of results can be explained by the fact that release information is removed in experiment 2. Obviously in the single stop condition, releases contain information about the critical consonant, and that information is removed. In addition, according to the release information account of the asymmetry in experiment 1, there is significant information about "p" in the p#t releases. This account is strongly supported by the fact that the decrement in performance in experiment 2 is about the same for the p#t condition as it is in the p#h condition: "p" release information is important in both cases. Likewise it explains why there is no decrement in the t#p condition, since it hypothesized that there no information about "t" in the t#p releases (see Fig. 3).

Finally, the fact that there is still a significant context by condition interaction in experiment 2, even though it is weaker, suggests that other factors, such as the acoustic one hypothesized by Byrd (1992) may also be contributing to an asymmetry between t#p and p#t.

2. Correlational analyses

Table IV shows Spearman's rank order correlations between gestural measurements and the detection of the target consonant. For "t" (first row), the pattern of correlations are similar to those in experiment 1, although almost universally smaller in value. The factor showing the largest correlation with the detection of the target is the Opening Lag, as it was in experiment 1. That this factor also showed the largest correlation with detectability in this experiment ($r_s=0.82$,

$p<0.05$), supports the analysis suggested above, that the strong correlation of Opening Lag was not due to any perceptual importance of the relative timing of the gestures' releases, but rather occurred because Opening Lag might capture the amount of acoustically relevant overlap during the gestures' closing phases. In experiment 2, the actual constriction openings are necessarily perceptually irrelevant: The stimulus is cut before either gesture is released. Although listeners only hear the gestures' closing phases, Opening Lag is the largest predictor of perceptual responses.

A closer look at the data reveals that responses to the truncated "tot p" in this experiment were highly correlated with the intact "tot puddles" in experiment 1 ($r_s=0.92$; $p<0.05$). In contrast, the correlation between responses to "tot h" and "tot huddles" was -0.42 ($p=0.19$). This is consistent with the view that all of the information relevant to perception of final "t" in the context of a following "p" is, in fact, in the closing movements, even in the intact forms of experiment 1.

In contrast to experiment 1, there were some reliable correlations between the number of "p" responses and the articulatory measurements. In particular, Δ LA Closing was highly correlated with the detectability of the target ($r_s=0.70$, $p<0.05$) indicating that when the movement into the lip gesture is large, detectability of "p" increases. The Closing Lag was a reliable predictor of detectability ($r_s=-0.71$, $p<0.05$) showing that the larger the separation between the two gestures, the larger the detectability of the target. The sign is reversed in the Closing Lag factor (compared to the Opening Lag) because the measurement is from the beginning of the TT movement to the beginning of the lip movement. The fact that such correlations only emerge in experiment 2 again supports the hypothesis that lip information is being perceived during the release of p#t stimuli. When such information is removed, and the only information available occurs during the closing movements, the kinematic characterization of the closing gestures and their overlap affect detectability.

C. Discussion

The results of this experiment were consistent with those of experiment 1, especially as regards the perception of a final "t" when it is produced in the context of a following overlapping stop. One difference between the experiments is that there were generally fewer correct detections of the target in experiment 2, when the stimuli were truncated during the medial closure(s). This effect was especially evident for the single stop condition, when the final gestures were not followed by an overlapping oral stop beginning the next word, but instead by /h/. The reason for this difference seems straightforward. In experiment 1, there are two sources of information that can lead to the detection of the final stop in these single stop stimuli: the closing movement of the stop gesture, and the opening (or release) movement. In experiment 2, this opening movement has been cut off, and the closing movement is all that is available. Without the opening movement, the amount of information specifying the gesture is reduced, and uncertainty increases. Moreover, the ability of listeners to perceive place information correctly based only on VC movements is relatively poor, as found in the studies listed previously. This would account for the fact that the decrement in performance is primarily observed for the single stops in the "t" condition, since it is only these overt finals that have audible opening movements in experiment 1. For stimuli like "tot puddles," the opening movement for the final alveolar stop is almost completely achieved by the time the labial closure begins to open (within 7 ms) so this tongue tip opening information is no more available in experiment 1 than it is in experiment 2 for these stimuli. For the "p" condition, decrement occurs in both single stop and stop sequence contexts. This also seems to be explicable on the basis of the timing of the release movements. For overlapped stimuli like "top tuttles," the opening movements for the final labial and the following tongue tip gestures are almost synchronous (the labial leads by a very small amount, about 10 ms), but this means that some of the labial opening information is available to listeners in this condition in experiment 1, but not in experiment 2. Thus for "p" stimuli, performance would be expected to be reduced in both single stop and stop sequence contexts.

The differences in results between experiments 1 and 2 cannot be well explained by the hypothesis that listeners are using closure duration as a "cue" for whether they hear one or two consonants. That hypothesis might predict an improvement in detectability for the stop sequence condition in experiment 2 over that in experiment 1, as the potentially short closure durations of experiment 1 could bias listeners to hear a single consonant, whose identity would be dominated by the release information. Such an improvement was, in fact, exhibited in Byrd's (1992) study. While it is not clear how to explain this difference between the studies, it should be noted that even in Byrd's study, the asymmetry between alveolar-labial and labial-alveolar sequences is less pronounced when truncated stimuli are employed.

The absence of the opening movements in experiment 2 could also account for the patterns observed in the correlation results. If the t#p stimuli have essentially the same gestural information available to listeners in experiments 1 and 2, as suggested above, then one would expect the same pattern of correlations with the articulatory measures in the two experiments, and that is, indeed, what was found. On the other hand, the gestural information available to the listeners in the p#t condition is different in the two experiments, because lip opening information *is* available to listeners in experiment 1, but not in experiment 2. In the absence of the opening information (and the attendant increase in uncertainty), aspects of the magnitude of the labial closure and the degree of overlap become important determinants of detectability, and this accounts for the significant correlations observed in experiment 2, but not in experiment 1.

III. GENERAL DISCUSSION

These two experiments show a significant reduction in the number of detections of a word-final alveolar consonant if it is partially or completely overlapped by a labial consonant gesture that begins the following word. A final labial does not suffer this decrement in detection to such an extent when overlapped by a following alveolar. This asymmetrical pattern replicates the perceptual results obtained by Byrd (1992) with synthetic speech and can be used to explain the direction of the assimilation tendencies observed in English and other languages (Browman and Goldstein, 1990; Byrd, 1992). The results of the two experiments combined suggest that one major factor contributing to this asymmetry is that a word-final tongue tip gesture is short in duration and that its release does not, therefore, overlap the release of the following labial. This reduced duration could be viewed as part of a general reduction of the final tongue tip gestures, which involves spatial reduction (Browman and Goldstein, 1995; Byrd, 1996; Giles and Moll, 1975) as well as this temporal reduction.

More generally, the results obtained show that variation in degree of overlap among gestural units has measurable consequences in listeners' ability to detect the presence of those gestures, even when that variation is within the range of natural speech. Detection of final "t" is poorer when the tongue tip closing gesture is more overlapped by a following labial closure gesture. Such results provide some realistic boundary conditions for theories that propose that listeners recover the gestures of the talker (e.g., Liberman and Mattingly, 1985; Fowler and Rosenblum, 1991). For his/her part, the talker may respect these limitations on recoverability. Recall that, in general, tokens produced early in the list by the talker exhibited better recognition than those produced late in the list. The talker may "intend" to allow more gestural reduction and overlap in circumstances in which the utterance can be assumed (by virtue of repetition) to be redundant (for related views of the role of informational context on production, see Fowler and Housum, 1987; Lindblom, 1990; Sotillo *et al.*, 1995).

Interestingly, the results support the view that gestural overlap can also have beneficial perceptual consequences. The detection of final "p" in "top tuddles" appears to benefit from the lip release movement that overlaps the tongue tip release for the initial "t." The listener appears to be able to use the acoustic signal at the release of the closure in "top tuddles" to provide information about the presence of both a tongue tip gesture and a lip gesture. Such a result is consistent with the "vector-analysis" view (e.g., Fowler and Smith, 1986) of how overlapped gestures are decomposed perceptually as well as with motor theory's general claim that the speech system is specifically designed to recover coproduced gestures. However, previous observations have focused on the perception of overlapping consonant and vowel gestures, while the present results argue that similar evidence can be found in the domain of consonant–consonant gesture overlap. It remains for future work to determine in what circumstances overlap can be correctly parsed by listeners (and therefore used beneficially in recovery) and in what circumstances overlap impedes recoverability. The stimuli used in these experiments include examples of both types.

Knowledge of the conditions under which gestural recovery can fail may contribute to an account of the kinds of gestural structures that languages tend to employ in their phonologies (e.g., Mattingly, 1981; Byrd, 1992; Silverman, 1995), as well as to an account of how gestural structures may change over historical time (e.g., "listener-based" sound changes, of the kind that have been discussed by Ohala, 1981). Surprisingly, few studies have systematically examined correlations between articulatory variability and perception of speech. These experiments add to the small literature correlating speech production measures with perceptual data. Although it seems obvious that the variability in movements of the vocal tract would be correlated with how gestures are perceived, there are few direct experimental demonstrations of this.

ACKNOWLEDGMENTS

This research was supported by an individual NIH National Research Service Award to the first author and by NICHD Grant No. HD 01994 to Haskins Laboratories. Portions of these experiments were presented at the 125th Meeting of the Acoustical Society of America, Ottawa, Ontario, Canada. Correspondence may be addressed to Aimée M. Surprenant, 1364 Psychological Sciences Bldg., Purdue University, West Lafayette, IN 47907-1364, email: aimee@psych.purdue.edu. The authors wish to thank Dr. Ian Neath for help conducting the experiments and for comments on an earlier version of the manuscript, and Dr. Michael Studdert-Kennedy and Dr. Ignatius Mattingly for their comments.

Browman, C. P., and Goldstein, L. (1988). "Some notes on syllable structure in articulatory phonology," Phonetica **45**, 140–155.

Browman, C. P., and Goldstein, L. (1990). "Tiers in articulatory phonology, with some implications for casual speech," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge U.P., Cambridge), pp. 341–376.

Browman, C. P., and Goldstein, L. (1991). "Gestural Structures: Distinctiveness, phonological processes, and historical change," in *Modularity and the Motor Theory of Speech Perception*, edited by I. G. Mattingly and M. Studdert-Kennedy (Erlbaum, Hillsdale, NJ), pp. 313–338.

Browman, C. P., and Goldstein, L. (1992). "Articulatory phonology: An overview," Phonetica **49**, 155–180.

Browman, C. P., and Goldstein, L. (1995). "Gestural syllable position effects in American English," in *Producing Speech: Contemporary Issues. For Katherine Safford Harris*, edited by F. Bell-Berti and L. J. Raphael (American Institute of Physics, Woodbury, NY), pp. 19–34.

Brown, G. (1977). *Listening to Spoken English* (Longman, London).

Byrd, D. (1992). "Perception of assimilation in consonant clusters: A gestural model," Phonetica **49**, 1–24.

Byrd, D. (1996). "Influences on articulatory timing in consonant sequences," J. Phonet. **24**, 209–244.

Clark, H. H. (1973). "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research," J. Verb. Learn. Verb. Beh. **12**, 335–359.

Dorman, M. F., Raphael, L. J., and Liberman, A. M. (1979). "Some experiments on the sound of silence in phonetic perception," J. Acoust. Soc. Am. **65**, 1518–1532.

Foss, D. J., and Blank, M. A. (1980). "Identifying the speech codes," Cogn. Psychol. **12**, 1–31.

Fowler, C. (1986). "An event approach to the study of speech perception from a direct-realist perspective," J. Phonet. **14**, 3–28.

Fowler, C. A., and Housum, J. (1987). "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction," J. Mem. Lang. **26**, 489–504.

Fowler, C. A., and Rosenblum, L. D. (1991). "The perception of phonetic gestures," in *Modularity and the Motor Theory of Speech Perception*, edited by I. G. Mattingly and M. Studdert-Kennedy (Erlbaum, Hillsdale, NJ), pp. 33–59.

Fowler, C. A., and Smith, M. (1986). "Speech perception as "vector analysis": An approach to the problems of segmentation and invariance," in *Invariance and Variability in Speech Processes*, edited by J. Perkell and D. Klatt (Erlbaum, Hillsdale, NJ), pp. 123–136.

Fujimura, O., Macchi, M. J., and Streeter, L. A. (1978). "Perception of stop consonants with conflicting transitional cues: A cross-linguistic study," Lang. Speech. **21**, 337–346.

Geumann, A., and Kröger, B. (1995). "Some implications for gestural underspecification as a result of the analysis of German /t/ assimilation," in Proceedings of the XIIIth International Congress of Phonetic Sciences, edited by K. Elenius and P. Branderud (ICPhS 95) (KTH and Stockholm University, Stockholm).

Giles, S., and Moll, K. (1975). "Cinefluorographic study of selected allophones of English /l/," Phonetica **31**, 206–227.

Kuehn, D. P., and Moll, K. (1976). "A cineradiographic study of VC and CV articulatory velocities," J. Phonet. **4**, 303–320.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. **74**, 431–461.

Liberman, A. M., and Mattingly, I. G. (1985). "The motor theory of speech perception, revised," Cognition **21**, 1–36.

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H and H theory," in *Speech Production and Speech Modeling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, Dordrecht, The Netherlands), pp. 403–439.

Malecot, A. (1956). "Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique," Language **32**, 274–284.

Marslen-Wilson, W., and Tyler, L. K. (1980). "The temporal structure of spoken language understanding," Cognition **8**, 1–71.

Mattingly, I. G. (1981). "Phonetic representation and speech synthesis by rule," in *The Cognitive Representation of Speech*, edited by T. Myers, J. Laver, and J. Anderson (North-Holland, Amsterdam), pp. 415–420.

Nadler, R. D., Abbs, J. H., and Fujimura, O. (1987). "Speech movement research using the new x-ray microbeam system," University of Wisconsin Speech Motor Control Laboratories Preprints, pp. 181–184.

Nolan, F. (1992). "The descriptive role of segments: Evidence from Assimilation," in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by G. Docherty and D. R. Ladd (Cambridge U.P., Cambridge).

Ohala, J. (1981). "The listener as a source of sound change," in *Papers from a Parasession on Language and Behavior*, edited by C. S. Masek, R. A. Hendrick, and M. F. Miller (Chicago Linguistics Society, Chicago), pp. 178–203.

Ohala, J. (1990). "The phonetics and phonology of aspects of assimilation," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge U.P., Cambridge), pp. 258–275.

Pollack, I. (1970). "A nonparamentric procedure for evaluation of true and false positives," Behav. Res. Methods Instrum. **2**, 155–156.

Repp, B. (1978). "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants," Percept. Psychophys. **24**, 471–485.

Rubin, P., Baer, T., and Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research," J. Acoust. Soc. Am. **70**, 321–328.

Saltzman, E., and Munhall, K. G. (1989). "A dynamical approach to gestural patterning in speech production," Ecol. Psych. **1**, 333–382.

Silverman, D. (1995). "Phasing and recoverability," Doctoral dissertation, Department of Linguistics. UCLA.

Sotillo, C., McAllister, J., Bard, E. G., Doherty-Sneddon, G., and Newlands, A. (1995). "Word intelligibility and place assimilation in spontaneous speech," in *Proceedings of the XIIIth International Congress of Phonetic Sciences*, **2**, edited by K. Elenius and P. Branderud (KTH and Stockholm University, Stockholm, Sweden), pp. 550–553.

Wang, W. S. (1959). "Transition and release as perceptual cues for final plosives," J. Speech Hearing Res. **2**, 66–73.