NIH Public Access
Author Manuscript
J Mem Lang. Author manuscript; available in PMC 2010 July 9.

Published in final edited form as:
J Mem Lang. 2003 October ; 49(3): 396–413. doi:10.1016/S0749-596X(03)00072-X.

Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks☆

Carol A. Fowler[a,b,*], **Julie M. Brown**[a,b], **Laura Sabadini**[a,b], and **Jeffrey Weihing**[a]

[a] Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695, USA

[b] University of Connecticut, Storrs, CT 06269, USA

Abstract

Participants took part in two speech tests. In both tests, a model speaker produced vowel–consonant–vowels (VCVs) in which the initial vowel varied unpredictably in duration. In the simple response task, participants shadowed the initial vowel; when the model shifted to production of any of three CVs (/pa/, /ta/ or /ka/), participants produced a CV that they were assigned to say (one of /pa/, /ta/ or /ka/). In the choice task, participants shadowed the initial vowel; when the model shifted to a CV, participants shadowed that too. We found that, measured from the model's onset of closure for the consonant to the participant's closure onset, response times in the choice task exceeded those in the simple task by just 26 ms. This is much shorter than the canonical difference between simple and choice latencies [100–150 ms according to Luce (1986)] and is near the fastest simple times that Luce reports. The findings imply rapid access to articulatory speech information in the choice task. A second experiment found much longer choice times when the perception–production link for speech could not be exploited. A third experiment and an acoustic analysis verified that our measurement from closure in Experiment 1 provided a valid marker of speakers' onsets of consonant production. A final experiment showed that shadowing responses are imitations of the model's speech. We interpret the findings as evidence that listeners rapidly extract information about speakers' articulatory gestures.

Keywords

Direct realism; Motor theory of speech perception; Acoustic theories of speech perception; Choice reaction time; Simple reaction time

Scientists attempt to carve nature at its joints to select a subdomain on which to conduct research. For such an approach to yield meaningful outcomes, natural systems must be nearly decomposable (Simon, 1969). By some accounts (Van Orden, Holden, & Turvey, in press), however, systems under study by cognitive psychologists are not nearly decomposable, to the detriment of the interpretability of research findings.

One way in which cognitive systems have been partitioned for scientific investigation is into the distinct perceptual systems and action systems. Although ecological psychologists (e.g., Reed, 1996) have consistently resisted that partitioning, it has persisted in other domains. Recently, however, some theorists who take more conventional theoretical approaches to the study of cognition have also confronted the necessity of investigating perception as it guides action, and action as it informs perception.

☆The research reported here was supported by Grants HD-01994 and DC-03782 to Haskins Laboratories.

[*] Corresponding author. Fax: 1-203-865-8963. carol.fowler@haskins.yale.edu (C.A. Fowler).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Attention by researchers exclusively either to perceptual processing or to motor control has directed attention away from the question of how perception can guide action in the world (cf. Prinz, 1997). If, in a theory of perception, percepts are characterized in sensory terms and, in a theory of motor control, action planning is characterized in motor terms, it is not obvious how percepts can be mapped on to action plans. For crosstalk to be able to take place, Prinz (1997) suggests, there must be a "common coding" in the two domains. He proposes the "action effect" principle whereby action planning is coded in terms of the effects the actions will have in the world, that is, in terms of the distal events that perceivers perceive. His research, and that of others taking the same approach is designed to reveal the common coding (see Hommel, Musseler, Aschersleben, & Prinz, 2001; for a review).

The study of speech typifies the study of other perceptual and action systems. Investigators who study speech perception characteristically do not study production and vice versa. Nor, in general, are theories of perception and production developed with attention to the obvious fact that the two capabilities must interface. Their interfacing manifests itself in two ways at the phonetic level of description on which we focus here. First, in ordinary, between-person, speech communication, listeners perceive the phonetic forms that speakers produce. As Liberman and colleagues (e.g., Liberman & Whalen, 2000), put it, for speech communication to succeed, there must be a relation of parity between phonetic messages sent and received. Second, within speaker/listeners, it is well-known that perceiving speech has an impact on speech production.

This is manifest in a variety of ways. Some of them are special to experimental manipulations, as, to provide just one example among many, when transformations of the acoustic consequences of speaking lead to adjustments in the gestures that produce speech (e.g., Houde & Jordan, 1998). Some are manifest in more natural settings as when exposure to the speech of others leads to phonetic convergences with that speech (e.g., studies reviewed by Giles, Coupland, & Coupland, 1991; see also Sancier & Fowler, 1997).

It is likely that the direction of influence between speaking and listening at the phonetic level of description goes the other way as well; however, to date, this is little studied. Sams (personal communication, February 25, 2003) does report, however, that when speakers mouth one syllable synchronously with their perceiving an appropriately selected other syllable, something like a McGurk effect (McGurk & MacDonald, 1976) can occur in which the speaker/listener reports hearing a syllable that integrates phonetic properties of the acoustic and the mouthed syllable.

For the most part, theories of speech perception do not include accounts of how speech perception links to production. This has a variety of consequences both for understanding between-person communication and for understanding perceptual guidance of production within speakers. Prototypical theories of speech perception propose that listeners analyze acoustic signals to extract cues to the phonetic units (e.g., phonetic segments, syllables) that compose the speaker's phonetic message. Those cues are mapped to abstract phonetic categories in memory. We refer to these as "acoustic" theories of speech perception, to contrast them with the gestural theories on which we focus in our research. Examples are Massaro's Fuzzy Logical Model of Perception (e.g., 1998), Diehl and Kluender's auditory enhancement theory (e.g., 1989; Kluender, 1994), Lotto's model (2000) and others. Prototypical accounts of speech production (but see Guenther, Hampson, & Johnson, 1998) are that plans for speaking eventually control articulators or systems of them (e.g., Saltzman & Munhall, 1989). or muscles. If these prototypical accounts are put together to make a unified theory of listening and speaking, in the terms of Prinz (1997), there is no common coding at the level of description of the public consequences of speaking, that is, articulator movements. In the terms of Liberman and Whalen (2000), there is no account of the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

achievement of parity. This is not to say that one could not be devised. The obvious common currency would be the covert phonetic categories that serve as the end point of phonetic perception and might serve as the starting point of phonetic production planning. Here we are remarking only that there are important issues about between-person communication that remain unaddressed when the theories are not developed in relation to one another. The same omissions occur in developing an understanding of within-person speech. Specifically, there is no account in these theories for findings that speech perception guides speaking or that speaking can affect perception.

Our research focuses on within-person speech. It is designed to underscore the tight linkage between speech perception and production, and specifically the observation that perceiving speech has immediate effects on production. We describe the research from the perspective of gestural theories of speech perception, because, to date, these are the only theories that provide an understanding how and why listening to speech should have an immediate impact on production.

The motor theory of speech perception (e.g., Liberman & Mattingly, 1985) and direct realist theory (e.g., Fowler, 1986) are gestural theories that offer accounts of the link between speech perception and production. In the motor theory, the link has two related bases. One is in the brain mechanism (a phonetic module) that the theory claims supports both capabilities. In the theory, to recover phonetic primitives from coarticulated acoustic signals requires recruitment of the speech motor system and specifically a process of analysis by synthesis (e.g., Liberman & Mattingly, 1985). Recent findings by Fadiga, Craighero, Buccino, and Rizzolatti (2002) using transcranial magnetic stimulation provide supportive evidence. The research revealed enhanced tongue muscle activity when listeners heard lingual as contrasted with nonlingual consonants. This kind of finding suggests that motor recruitment does occur in speech perception, although, of course, it does not require the interpretation that analysis by synthesis occurs or that speech perception would be blocked if motor recruitment were blocked. Moreover the finding of mirror neurons in monkeys (e.g., Rizzolatti, Fadiga, Gallese, & Fogassi, 1996) and perhaps in humans as well (Fadiga, Fogassi, Povesi, & Rizzolatti, 1995) that respond both when a particular action (e.g., a particular grasping action) is perceived and when it is produced may suggest that links between action and perception are quite general rather than being special to speech as Liberman and colleagues proposed.

The second basis for a the link between speech perception and production in the motor theory is the common currency of the primitives that support both capabilities. The common currency in this case is the phonetic gestures that talkers produce and listeners are proposed to perceive.[1] This common currency not only underlies perceptual guidance of speaking within speaker/listeners, it also promotes between-person communication. In general, the phonological messages that talkers produce should be those that listeners perceive, and the common currency between talker and listener of gestural primitives fosters this achievement of parity.

Direct realist theory has remained agnostic regarding brain mechanisms supporting speech perception and production. In this theory, listeners use structure in acoustic speech signals as information for its causal source, as they do in perceiving generally. In the theory, the causal sources are the phonetic gestures, that is, the linguistically significant actions of the vocal

[1] In the motor theory, the gestures are abstracted from the actions of the vocal tract during speech even though they are the control systems that underlie speech production. Liberman and Mattingly supposed that coarticulation prevented intended gestures from becoming actual in the vocal tract. Accordingly, the common currency enabling crosstalk between production and perception, like the abstract phonetic categories of other theories of speech perception remain in the privacy of the mind.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

tract that talkers produce. In this account, the common currency of listening and speaking that allows both perceptually guided speaking and achievement of parity is provided by the phonetic gestures that occur publicly during speech (e.g., Goldstein & Fowler, in press).

Research reviewed by Prinz (1997) that was designed to expose common coding in perception and action includes demonstrations of stimulus–response compatibility effects. In research by Hommel (1993), participants pressed buttons with the left and right hands to high and low pitched tones. The tones were presented randomly to the left and right sides of space. Responses were faster if the tone associated with the left hand was presented to the left rather than the right side of space. This compatibility effect between tonal stimulus and button press response might have either of two sources. Possibly it reflects an abstract compatibility ("left" hand corresponds to "left" side of space and "right" with "right"). Alternatively, it might reflect compatibility between the side of space of the tonal stimulus and the side of space of the responding hand. Research by Simon, Hinrichs, and Craft (1970) and Wallace (1971) ruled out the former compatibility by having participants cross their hands so that the right hand, on the left side of space, responded faster to stimuli presented on the left side of space. Hommel (1993) showed that what matters is the location of the responding action's effects. When participants' button presses caused lights to go on on the side of space opposite to the buttons, and participants were instructed, e.g., to make the right-hand light go on when they heard the high pitched tone and the left light when they heard the low pitched tone, compatibility effects were between the sides of space of the light and tone, not between the side of space of the light and the response button. In Prinz's terms, this is compatibility between an event code (a tone on one side of space or the other) and a planned action effect (turning on a light on one side of space or the other).

Some results of speech research have led to similar inferences that stimulus–response compatibility is present in speech when perception and production are represented in a common currency. In 1965, Kozhevnikov, Chistovich, and colleagues (Kozhevnikov & Chistovich, 1965) reported that shadowed vocal responses to speech were associated with shorter latencies than written responses to speech. They interpreted their findings as consistent with the idea that "the process of speech perception is associated with latent activity of the speech forming organs" (pp. 222–223), a proposal recently confirmed by Fadiga et al. (2002).

Porter and Lubker (1980), and, in a separate study, Porter and Castellanos (1980) replicated the study of Kozhevnikov et al., but with a new comparison. They compared latencies to make vocal responses in both a simple and a choice reaction time test. In both tests, a model speaker produced an extended vowel /a/ and then after an unpredictable period of time shifted to /o/, /æ/ or /u/ (Porter & Lubker) or to a consonant–vowel (CV) syllable (Porter & Castellanos). In the simple task, participants shadowed the extended /a/, but as soon as they detected the change to a new vowel or CV, they were to shift to /o/ (Porter & Lubker) or /ba/ (Porter & Castellanos). In this test, shifting from /a/ merely signaled that a change had been detected. In the choice task, in contrast, the subject's task was again to shadow /a/, but then to shift to whatever vowel or CV the model shifted to. Accordingly, listeners had to do more than detect a change from /a/; they had to determine what was being said and to say that themselves.

Luce's (1986) review of studies of simple response times shows that average latencies in the studies cover a considerable range; latencies depend on stimulus properties (such as the modality of the signal, its intensity, etc), on trial characteristics (is there a warning signal and if so how much later does the target stimulus occur?), and other properties. The graphical displays that Luce provides suggest that the fastest times can be as fast as approximately 120 ms, but times range upward to about 300 ms. Luce (1986, p. 208) reports

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

that, generally, simple responses times are faster than choice response times by 100–150 ms when the two tasks are made comparable. This latency difference is understandable in that the simple task only involves detection of a stimulus (or stimulus change) whereas the choice task requires that a choice among responses be made depending on the identity of the stimulus.

In contrast to this typical finding, Porter and Lubker found just a 15 ms difference between the choice and simple response times; Porter and Castellanos found a 50 ms difference. Moreover, response times were very fast, so the small difference between conditions was not due to simple responses having been slow. Why, under their experimental conditions, is the choice/simple response time difference so small?

Porter and Castellanos suggest that “subjects are able to directly and accurately realize the results of perception in articulatory terms” (p. 1354). This, of course, is highly compatible with interpretations by either the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000) or direct realist theory cited earlier (Best, 1995; Fowler, 1994). In these accounts, the results of the choice tasks reflect the substantial compatibility between perceived stimuli and responses to them. Underlying the compatibility is the common gestural currency of speaking and listening.

No doubt the compatibility can be accounted for in other ways, specifically in ways that are more consistent with current acoustic theories of speech perception. Theories tend to be elastic. However, to date, as noted, no accounts have been offered. Accordingly, we will refer to the gestural theories in interpreting our findings. Our research is designed to highlight the existence, within people, of a speech perception–production linkage by exploring further the marked stimulus–response compatibility of the speech shadowing task of Porter and colleagues. Perhaps the research will promote attention to the linkage among theorists who deny that speech gestures are perceived by underscoring that listeners are also speakers. This is a fact that has consequences for understanding the role of speech perception in the guidance of production. A second purpose of the research is to begin to explore the nature of the perception–production link.

In this first study, we provide a replication and extension of the study by Porter and Castellanos (1980). Our first experiment differs from theirs in several ways. First, in their simple response task, all participants used /aba/ as the simple response (or /ba/ in another condition). Therefore the latencies to make /ba/ responses in the simple task were being compared to the averaged latencies to produce any of five CVs in the choice task. We counterbalanced simple responses across our participants and then we only compared identical simple and choice responses (that is, say, /apa/ produced as a response to /apa/ in the simple as well as in the choice task). We used a different set of consonants than Porter and Castellanos. They had used /b/, /p/, /m/, /g/, and /k/. We used the three voiceless stops of English. Whereas Porter and Castellanos report that they used “practiced” participants without specifying the amount of practice, we tested our participants in eight sessions in each of which they responded to 144 simple and 144 choice response time trials. Finally, whereas most of the data that Porter and Castellanos report are measures of latency between closure in the stimulus and closure in the response, we measured and report latencies from closure to closure and from release to release. Measurement point turns out to have an effect on assessed performance.

Besides looking at the differences in response latencies in the choice and simple tasks, we tested for evidence of rapid access to gestural information by perceivers in a second way. In the motor theory, perceiving speech involves using the acoustic signal to determine the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

gestures that produced it by means of analysis by synthesis, Accordingly, a speech stimulus might be expected to “prime” or to serve as a goad for an imitative response. In direct realist theory, listeners perceive the speaker’s gestures, not necessarily because listeners involve their own speech motor systems in perceiving, but because the speaker’s gestures are what the structure in the acoustic signal provides direct information about. Listeners do imitate the speech they hear (e.g., Goldinger, 1998); perhaps they do so in part, because perceiving gestures serves as a prime or goad (Fowler, 2000). A third, imaginable, account of speech perception that predicts these findings is that listeners map acoustic cues onto phonetic categories in memory that link also to the motor control systems that underlie their production.

In these accounts, simple responses should be shorter when the model’s CV matches that of the speaker. That is, in the simple task, speakers are assigned a syllable. They produce that assigned syllable whenever they hear the model produce a CV. On most trials, the model’s CV will not match that of the speaker, but on some trials it will. We predict that latencies should be shortest when the CVs match, because of the common currency accessed while individuals speak or listen.

Experiment 1

Method

Participants—The six participants were native speakers of English who were paid for their participation in eight sessions each of approximately 45 min duration.

Materials—A female native speaker of English (J.M.B.) was recorded producing extended /a/s followed by /pa/, /ta/ or /ka/. She was cued to produce /a/ for each of eight durations before shifting to one of the CVs. This was achieved by presenting “a” on the computer screen for the designated duration and then replacing it with “pa,” “ta,” or “ka.” Accordingly, the extended /a/s were produced without the speaker’s knowledge of the identity of the CV, which, therefore, could have affected the vowel’s production only at the vowel’s end. Her utterances were edited by removing whole pitch pulses from the beginning of the vowel so that /a/s were 2, 2.5, 3, 3.5, 4, 4.5, 5, and 5.5 s long before each CV. These different durations of initial /a/ were used so that participants could not predict when the model would shift to producing the following CV. We used three tokens of each /aCa/ disyllable at each of the eight durations of initial /a/ giving us 72 unique stimulus tokens. In each simple and choice response session, the 72 stimuli were presented twice each to make a 144 item test.

Procedure—Participants took part in eight sessions over a two- to three-week period. Each session lasted approximately 45 min during which participants took both a 144 item simple response time test and a 144 item choice test. The order of the tests was counterbalanced over participants; however, each participant experienced the conditions in the same order in each of his or her sessions.

The tests were presented using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). The order of trials was randomized separately in each session and for each participant and task. For the simple task, participants were assigned a CV, /pa/, /ta/ or /ka/, as their response to produce when the model speaker shifted from /a/ to one of the three CVs. Two participants were assigned each CV. Participants were told to shadow the initial /a/ and then, as rapidly as possible, to shift to their assigned CV when they heard the model shift from /a/ to a CV. In the choice task, they were told to shadow /a/ and then, as rapidly as possible, shift to the CV to which the model had shifted. In both tasks they were told to respond as quickly as possible without making many mistakes. If they found themselves making many

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

mistakes, they were to slow down a little. However, we emphasized that speed was important. Trials were self-paced.

In the first and eighth sessions, participants' responses were recorded onto one channel of a cassette tape as the model's utterances were recorded on the other channel. (Sessions 2–7 were not recorded.) The two channels were simultaneously input to SoundScope (GW Instruments, Somerville, MA), in which latency measurements were made. We measured latencies using both the waveforms and spectrographic displays of the participants' and the model's speech: from the model's onset of closure for the consonant to the participant's and from the model's release burst to the participant's. Measurements were made by authors J.M.B., L.S., and J.W.

Results

Errors in which speakers produced the wrong consonant or made some other speech error were rare in the simple response task; in all, there were just 10 such errors (1% of the data). In contrast, many such errors occurred in the choice task. Error rates were 12.5% overall in the first session and 14.1% in the eighth session on the choice task. [Porter and Castellanos (1980) report a 23% error rate among their participants.] We will discuss the possible meaning of this difference in error pattern between choice and simple responses after reporting the primary findings on latency.

Table 1 provides mean latencies measured from closure and from release in the choice and simple response time tasks averaged across sessions 1 and 8. Latencies of identical simple and choice trials (e.g., in the simple and choice tasks, /apa/ produced as a response to /apa/ for the participants whose designated CV was /pa/) were entered into an ANOVA with four factors: Task (simple, choice), Session (first, eighth), Measurement Point (closure, release), and Duration of initial /a/ (2–5.5 s). Data of each participant were analyzed separately, and, in addition, an analysis was done on the averaged data of the six participants.

In the overall analysis, the effect of duration was significant (*F* (7, 35) = 7.22, *p* < .0001); the means show an almost monotonic tendency to decrease as /a/ duration increases (from 206 ms with the 2 s /a/ to 173 with the 5.5 s /a/). Because this factor is not of great interest and because it did not participate in significant interactions with factors of interest (with one exception in the case of one participant), we will not consider it further.

Overall we obtained a highly significant effect of task (*F* (1, 5) = 120.11, *p* < .0001) with simple responses shorter than choice responses by 59 ms on average. Across participants the difference ranged from 40 to 68 ms. However, in the overall analysis and in each of the individual analyses, task interacted with measurement point (overall: *F* (1, 5) = 18.12, *p* = .008). Measured from closure, the choice task responses were slower than the simple responses by just 26 ms [this compares to the 50 ms difference reported by Porter and Castellanos (1980)]; measured from release, the difference was 91 ms. Measured from either point, simple response times averaged 157 ms. However, measured from closure, choice response times averaged 183 ms, but measured from release, they averaged 248 ms.

The significant task by measurement point interaction reflects the much larger difference in the size of the task effect when latencies were measured from release rather than closure. A test of the task effects at both measurement points revealed that, whereas the difference in response times to choice trials differed highly significantly from those to simple trials measured from release (*F* (1, 5) = 70.02, *p* < .001), measured from closure the 26 ms difference was only marginally significant (*F* (1, 5) = 5.51, *p* = .07). However, at this measurement point, five of the six participants showed numerically (and four of those significantly) slower responses to choice than to simple trials.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Overall, the effect of session was nonsignificant. Despite the nonsignificance of the session factor, five of the six participants showed numerical reductions in latency from session 1 to 8; of those, four were significant in individual analyses. In the overall analysis, session did not interact with any other factor. By session 8, the difference between the choice and simple response times measured from closure was 31 ms (range 2–55 ms); measured from release, it was 84 ms (range 32–156).

No other factor reached significance in the overall ANOVA. (Two factors were marginally significant: Task by Duration and Measurement Point by Duration. However, in both instances, the effect of task was present at every level of duration; it was larger at some durations than others in no intepretable pattern.)

We did another analysis meant to provide converging evidence that listeners access gestural information when they hear speech. This analysis focused on the simple response times. We asked whether response latencies were shorter when the participant's designated CV response matched that of the model than when it did not. This was the case numerically for five of the six participants. In individual analyses (with factors consonant, session, measuring point, and duration of /a/), the main effect of consonant was significant for four of the five participants showing the predicted effect To perform an overall analysis on the data as shown in Fig. 1, we created a factor Match with levels match and mismatch. Matches were response latencies on trials in which the model's consonant and the participant's response matched. Mismatches were response latencies to other trials. (So, for a participant whose simple response was /pa/, his or her latencies when the model produced /pa/ were averaged to provide a value for match, and his or her latencies when the model produced /ta/ or /ka/ were averaged to provide a value for mismatch.) We computed these values for each participant using the grand mean of simple response latencies averaged over both experimental sessions and over closure and release measures. We analyzed the means in an ANOVA with the between participant factor Consonant (that is, the consonant produced by the participant) and the within participant factor Match. The effect of Match was significant (F (1, 3) = 17.31, p = .025) with a 7 ms mean difference in latency between match and mismatch trials. The factor Consonant was not significant; nor was the interaction.

Discussion

Our findings provide a close replication of those of Porter and Castellanos (1980). Measured from closure, latencies of choice responses were slower than those of simple responses by just 26 ms. This is a remarkably small difference. Moreover, because the simple response times average 157 ms, a value at the fast end of simple times in Luce's review, we can know that the small difference between simple and choice reaction times occurs because choice times are fast, not because simple times are slow. If the difference is a reliable indicator of participants' quite short latency to begin producing the consonant in the choice task, it is a result highly consistent with a proposal that listeners perceive speech gestures, an interpretation like those offered by Kozhevnikov and colleagues and by Porter and colleagues.

We will elaborate on this interpretation, but not until we have addressed two other questions. One is whether the difference in error rates between the simple and choice response conditions means that a speed–accuracy trade-off may underlie the very similar response times measured from closure. The second is whether responses measured from closure are, in any case, valid indicators of participants' initiation of the response consonant. Latencies measured from release were considerably longer than those measured from closure. Moreover, whereas choice and simple response times differed only by 26 ms measured from closure, they differed by 91 ms measured from release. This difference approaches the canonical difference between choice and simple response times (between 100 and 150 ms)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

according to Luce. We will address the first question in Experiment 2 and the second in Experiment 3 and a subsequent acoustic analysis.

As to the first question, the different error rates neither demonstrate a speed–accuracy trade-off nor rule out that there is one. This is because the kinds of errors that occur in simple and choice tasks are quite different. This may explain why, in his explicit comparison of simple and choice tasks, Luce provides no comparison of their relative error rates. Essentially the only kind of error that participants make in a simple response task is a false alarm. That is, they respond with the response they know they will be making before they get the signal to respond. This kind of error is rare in a choice task, because participants do not know in advance what their response will be. Participants rarely produce the wrong response in a simple response task, because they make the same response on every trial. However, errors in which the wrong consonant occurs or two consonants are produced are essentially the only kind of errors that occur in the choice task.

We have no nonarbitrary way to identify a false alarm, and so no responses were excluded on that basis in our report of error rates in Results. If we arbitrarily identify response latencies 110 ms or shorter as false alarms, across sessions 1 and 8, we obtain a 15.3% false alarm rate in the simple task but less than a 1% rate (.3%) in the choice task. In contrast, errors in which the wrong consonant or two consonants were produced constituted less than 1% of simple responses but 13.3% of choice responses.

Accordingly, if false alarms reflect a speed–accuracy trade-off, responses in the simple task focus on speed and those in the choice task on accuracy. If reports of incorrect responses reflect a speed–accuracy trade-off then the tradeoff goes the other way.

Because comparisons of errors can neither demonstrate a trade-off nor rule one out, rather than comparing choice to simple response times in Experiment 2, we will compare choice responses in the highly compatible stimulus-responses condition of Experiment 1 with a lower compatibility condition.

Experiment 2

Aside from a comparison of simple and response latencies as provided in Luce's review (1986) and in the research by Porter and colleagues, there is a second way to show that there is a special compatibility between stimuli and responses in the choice task of Experiment 1. It is to show that, when the possibility of stimuli providing instructions for a choice response is eliminated, choice times are much slower than when that possibility is present. The outcome of the present experiment is not in doubt; however, the latencies it provides will allow us to index the extent to which the stimulus–response compatibility of Experiment 1 reduced latencies.

Method

Stimuli—We used a precursor tone followed by three tones of high (1500 Hz), medium (1000 Hz), and low (500 Hz) frequency. The precursor tone was the same frequency as the medium tone (1000 Hz) and varied in duration from 2.0 to 5.5 s in .5 s increments. The tone that followed the precursor tone was 400 ms in duration. There was 80 ms of silence between the precursor and following tone. These durations were modeled after the durations of the model's speech in Experiment 1. There was a 1500 ms ISI between trials. There were 3 tones × 8 durations for 24 stimuli and 3 repetitions of each stimulus for a total of 72 trials in each condition. There were 144 trials in the experiment. The same stimuli were used for the simple and choice conditions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Participants—Participants were six graduate students and one recent psychology graduate. They were paid for participating in the experiment. Participants had normal hearing and were native speakers of North American English (six from the US; one from Canada).

Procedure—Participants were asked to shadow the tones by saying /a/ when the precursor tone started and then switch to either /pa/, /ta/, or /ka/ when the tone switched. Speed was emphasized, but participants were instructed to slow down if they found they were making many errors. In the simple reaction time condition, they were asked to switch to one consonant. In the choice task they were asked to switch to one of three consonants based on which tone they heard. The designated consonant in the simple task was counterbalanced across participants. Likewise, the tone that was paired with each consonant was counterbalanced in the choice task. Due to a counterbalancing error, four of the participants' designated consonants in the simple condition were the same consonants shadowed with the low tone in the choice condition. One of the participants' simple consonants was the same consonant shadowed with the high tone, and one with the medium tone. The order of the simple and choice tasks was counterbalanced across participants. Participants' responses were recorded on one channel of a cassette tape. The tone stimuli were recorded on the other channel. The experiment took approximately 30 min. After the first test, simple or choice, subjects received instructions for the second test. They were given practice trials for both the simple and choice conditions.

Results and discussion

The data of one participant were discarded due to an error in the recording of the simple condition. We counted as errors trials on which participants responded with the wrong consonant or a double consonant.

There were no errors in the simple condition. The average error rate for the choice condition was 9% with a range from 3 to 18%. A between subjects ANOVA revealed no significant difference between the choice reaction error rates in this experiment and those in Experiment 1.

Latencies were measured from release. A repeated measures ANOVA with factors Test (simple or choice), Tone (low, medium, or high), and Duration of the precursor tone (2.0–5.5 s) was conducted, as in Experiment 1, on just the trials on which the simple and choice responses matched. In this ANOVA there was only a significant effect of Test (F (1, 5) = 57.57, p < .001). Simple responses were faster (mean 237 ms) than choice responses (mean 637 ms).

We conducted an ANOVA to compare the release latencies in Experiment 1 and in the present experiment, with factors Experiment (Experiments 1 and 2), Test (simple or choice), and Duration. The analysis revealed a significant effect of Experiment (F (1, 10) = 10.15, p < .01). Release latencies were faster in Experiment 1 (mean 224 ms) than in the present experiment (mean 437 ms). There was also a significant difference between simple and choice latencies (206 ms vs. 455 ms; F (1, 10) = 85.71, p < .001), and a significant interaction between Test and Experiment (F (1, 10) = 30.92, p < .01). Reaction times were 63 ms longer in the simple condition of Experiment 2 than Experiment 1, and 363 ms longer in the choice condition.

A comparison of the results of the first two experiments underscores the effect of stimulus–response compatibility when listeners' responses shadow those of a model speaker. This comparison buttresses the simple-choice comparison in Experiment 1. When choice responses are shadowing responses to the speech of a model speaker, latencies approach simple response times and are considerably faster than choice responses to unrelated stimuli.

NIH-PA Author Manuscript

A theory of speech perception needs to address why. It is not sufficient to note that stimuli and responses are the same. As Prinz (1997) has pointed out, theories need to address how stimuli can be mapped onto responses. In most accounts, stimuli are coded in one format and responses in a different format and how stimuli then can guide responses remains unaddressed and unexplained. We propose that the very fast responses in the choice condition of Experiment 1 as compared to Experiment 2 reflect the common currency of perceived and produced speech. Listeners gain rapid access to speech gestures.

However, a conclusion that responses were especially fast and close to simple responses depends on the validity of our measures of choice response time from closure. Experiment 3, and the acoustic analysis that follows it, addresses that issue.

Experiment 3

Porter and Castellanos largely reported response times measured from closure, but they also measured response latencies from release. Although they did not report all of their findings with the release measure, they did provide two sets of histograms of closure and release latencies, and they provided the mean values for the two measures. In both sets of histograms, closures are about 50 ms longer in the choice than in the simple responses. This would make their choice–simple difference about 100 ms measured from release (that is, the 50 ms that is measured from closure plus the 50 ms longer closure interval). Our difference was 91 ms. Accordingly, our finding likely is not a failure to replicate, but rather a consistent finding between the two studies. What does it mean?

It might be that measurements from closure are not valid indicators of the participants' latencies to produce CVs. Measurement from the model's closure point would be invalid if participants are insensitive to coarticulatory information for the model's consonant in the closing transitions of the vowel. Measurements to the participant's closure are not valid if what we identified as closure was not really consonant closure. We instructed our participants to respond as quickly as they could following the model. Perhaps these instructions led them to cease production of the vowel as soon as they heard the model do so. However, at that point in time, they may not have known what consonant to produce. The measurements from closure are only valid if there is detectable information about the model's CV at consonant closure, if our participants did initiate their own CVs at vowel's end, and if they did so having used information for the CV in the model's closing transitions.

We attempt to test the validity of the measurements from closure in two ways, one designed to evaluate the left edge of our closure to closure measurements and the other to evaluate the right edge. As for the left edge, if measurements from there are valid, we should find that listeners to our model's speech are sensitive to the consonantal information in the model's closing transitions so that they can initiate the production of the consonant before they hear the model's release. As for the right edge, the participant's closure onset, we should find, in acoustic measures of the formants of our participants' own speech at the point we identified with closure, that the formants differ depending on whether the following syllable was /pa/, /ta/, or /ka/. Finally, participants' closures should not occur sufficiently later than the model's release that it is plausible that post-release information, not closure information, guided their responses at closure. In short, measurements from closure are valid if we find that listeners are sensitive to consonantal information in the closing transitions of the model's speech and that their own closing transitions are different for /pa/, /ta/ and /ka/ even if their response times measured from the model's release are not plausible response latencies. In that case, participants must have used the consonantal transitions in our model's speech to guide initiation of their own consonant productions.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Our third experiment assesses listeners' sensitivity to the model's consonantal information in the initial vowel of her disyllables. A subsequent acoustic analysis provides the converging test of the participants' own speech and of the temporal alignment of their speech with that of the model's consonantal release.

In Experiment 3, we tested the validity of the left end of our measure from closure by asking whether our participants detected the model's production of a consonant at closure. We made this test in two ways, both involving cross-splicing.

We ran two experiments in which we cross-spliced our /a/-C-/a/ disyllables at closure onset so that closure transitions provided misleading information about the forthcoming consonant.

In one experiment (cf. Martin & Bunnell, 1981, 1982; Whalen, 1984) that we report very briefly here, we asked listeners to make speeded button press responses identifying the consonants in cross-spliced and "spliced" disyllables. We created spliced disyllables by splicing CVs beginning at closure onset from one token of a VCV to another, having the same duration initial /a/ vowel and the same consonant. In our instructions to participants we did not tell them that half of the utterances were cross-spliced or might sound as if they contained a sequence of two consonants. We found very accurate responses identifying the consonant on both spliced and cross-spliced disyllables, averaging 98% correct if "correct" meant that listeners identified the consonant as that of the CV, rather than that specified by the model's closing transitions. Despite this high accuracy and the absence of indications during debriefing that listeners were puzzled about the task, because they, in fact, heard two consonants, their response latencies were a significant 44 ms slower to cross-spliced as contrasted with spliced disyllables. This experiment shows that listeners are sensitive to consonantal information at consonant closure; however, when they are asked to identify the consonant, they identify it as the second consonant of the pair specified by information at closure and at release.

The experiment we report in detail here, addressed the question of the detectability of consonantal information in the closing transitions of the model's initial vowel in a way more directly related to the method of Experiment 1. Listeners performed the choice task with the cross-spliced and spliced disyllables. The major measure of interest is the number of responses in which they produced two consonants in their shadowing responses and produced the two consonants specified by the cross-spliced stimuli. We expect such dual consonant responses to exceed those produced in Experiment 1 and those produced to the spliced stimuli of Experiment 3.

Method

Participants—Participants were six undergraduate students from the University of Connecticut who received course credit for their participation. They were native speakers of English who reported normal hearing.

Materials—Stimuli were provided by our model's utterances to which participants in Experiment 1 had responded. We made cross-spliced and spliced versions of each utterance. In cross-spliced versions, for example, we took /pa/ from one /apa/ disyllable, and spliced it after /a/ from /ata/ or /aka/ (having the same duration initial /a/ as /apa/). Likewise, we made cross-spliced versions of /ata/ and /aka/ disyllables. In spliced tokens, we took, say, /pa/ from one token of /apa/ and spliced it after /a/ from a different token of /apa/ having he same duration initial /a/. In each spliced and cross-spliced disyllable, the splice point was the onset of closure. That is, for example, from a token of /ata/, the /ta/ syllable was copied beginning at the offset of the extended /a/ vowel. It was pasted immediately after the

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

extended /a/ of a token of /apa/, /aka/ or /ata/. In this way, in the cross-spliced tokens, the extended /a/ vowel provided misleading information about the consonant; the closure interval provided information, if any, for the pasted consonant. This allowed us to determine whether listeners are sensitive to misleading information in the closing transitions of the extended vowel.

We had three tokens of each disyllable at each duration of /a/. The tokens provided the initial /a/ for one spliced trial and two cross-spliced trials. There were eight /a/ durations, three consonants, three tokens of each CV, and three types of trials (for a /pa/, one spliced trial and two cross-spliced, one with /ta/ and one with /ka/). This made 216 trials in the test.

Procedure—The stimuli were presented using PsyScope. Participants shadowed each disyllable. Their instructions were those of Experiment 1. Their shadowing responses were recorded on one channel of a cassette tape; the model’s spliced and cross-spliced disyllables were recorded on the other channel.

Results

The most direct test of whether listeners detected the consonant specified by the closing transitions of the model’s speech is provided by response errors rather than response latencies. We made three predictions. One was that, when participants made errors in which they produced two consonants instead of one on cross-spliced trials (for example, producing /apta/), the two consonants would tend to be the two specified by the model’s closing transitions and release. (In effect, these are accurate shadowings of the model’s speech, but are classified as errors, because participants were instructed to shadow the initial vowel and the CV.) A second prediction was that participants would make more double consonant errors on the cross-spliced trials of this experiment than they made in session 1 of Experiment 1. Third, we predicted that they would make more double errors on the cross-spliced than on the spliced trials of the present experiment.

The proportions of double responses in which the two consonants produced were those specified by the closing transitions and the releases of the model’s speech ranged from .56 to .94 of double responses, averaging .84 of trials. This clearly exceeds the chance value of . 17 (that is 1/6 of all possible responses in which two different consonants are produced), and, with every participant showing the effect, the result is highly significant. Participants produced double consonants on .177 of cross-spliced trials, compared to .068 of trials in session 1 of Experiment 1, a significant difference (t (10) = 2.10, p = .03; one-tailed). Participants produced double consonants on .097 of spliced trials, a difference that is only marginally different from .177 (t (5) = 1.53, p < .10), with just four of the six participants showing the predicted direction of difference.

Latencies were slower on cross-spliced than on spliced trials (F (1, 5) = 14.23, p < .05), but only on measures from release; the interaction of measurement point and trial type was marginal (F (1, 5) = 5.55, p < .10). Latencies from closure averaged 212 ms; those from release averaged 276 ms.

Discussion

We have several indications that listeners were sensitive to information in our model’s closing transitions. In our briefly reported button press experiment, participants were slower to identify the consonant in the CV syllable on cross-spliced than spliced trials. In our shadowing experiment, when double responses were produced on cross-spliced trials, they were highly likely to be the very two that the model’s speech had specified. Finally, participants produced more double consonant responses on the cross-spliced trials of

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Experiment 3 than participants had done in the first session of Experiment 1. That they did not produce significantly more double responses on cross-spliced than on spliced trials of the present experiment may reflect spillover from cross-spliced to spliced responses.

It need not follow from these results, of course, that participants in Experiment 1 used that information to commit themselves to producing one of the three CV syllables. In the next analysis, we attempt to determine whether they did by examining measurements of their vowel-final second formant frequencies (that is, second formant frequencies (F2s) at closure) and then locating that point in time relative to the model's closure and release. This will validate the right edge of our latency measures from closure in Experiment 1. If talkers do differentiate the consonants at closure onset, and closure onset is less than a plausible response time later than the model's release, then we can infer that our participants not only committed themselves to a consonant at the point we identified with consonant closure, but they did so based on information in the model's initial vowel.

Acoustic analysis

We measured F2 at the end (the last visible pitch pulse) of the first /a/ vowel in all trials of each participant's choice responses excepting those in which the participant made an error on the consonant. Measurements were made from spectrographic displays of the disyllables using Macquirer (Scicon, Los Angeles, CA).

Participants showed a remarkably uniform pattern. F2s were lowest for forthcoming /pa/ syllables and (with one exception) highest for forthcoming /ta/ syllables. The average ending F2 was 1443 Hz for /pa/, 1614 Hz for /ta/, and 1547 Hz for /ka/. In analyses of individual talkers, the difference between /pa/ and /ta/ was significant in every instance; those between /pa/ and /ka/ and between /ta/ and /ka/ were each significant in five of the six comparisons. In the three comparisons between pairs of syllables, every talker had at least two significant differences. We can be confident that our participants in Experiment 1 both detected the early consonantal information provided by the model (Experiment 2) and began producing the detected consonant during their initial /a/.

The next issue concerns where in time the participants' closures occurred relative to the model's closures and releases. Possibly, participants waited to terminate the initial /a/ vowel and begin to produce the CV until they heard the model's release. Were our listeners using information in the model's closure to initiate their own closure responses?

Fig. 2 provides histograms of the latencies of participants' closure onsets measured from the model's closure onset (Fig. 2a) and from the model's release (Fig. 2b). Participants' closure onsets lagged those of the model by 183 ms on average in the choice task; their closures lagged the model's releases by 116 ms on average (range 86–138 ms on average for the six participants). These latter response times are short enough, particularly for two of the participants (averaging 86 and 88 ms), that we can rule out their having waited to initiate consonant production until they heard the model's release. It is likely, rather, that participants used information in the model's closing transitions to guide their choice responses.

We infer, therefore, that the longer latencies when measures are taken from release rather than closure do not reflect a delay in participants' initiation of the consonant. It is more likely to reflect a lack of confidence in the consonant they have begun producing. Participants may remain ready to shift their place of articulation if their initial place perception proves to be in error?[2]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We conclude that our findings are consistent with those of Porter and Castellanos (1980) and Porter and Lubker (1980) that participants' choice response times in a speech shadowing task are remarkably fast and close to their simple response times.

Interim discussion

Having found two converging indications that measurements from closure are valid indications of participants having initiated the post vocalic consonant, we return to our claim that findings such as those in Experiment 1 show that listeners to speech have rapid access to gestural information, which then guides their choice responses, and it slows their simple responses when they differ from the model's.

We doubt that the effect is special to speech, an idea that we intend to pursue (see the General Discussion). Rather it is likely special to tasks with high stimulus response compatibility.

As Prinz (1997) and Hommel et al. (2001) remark, theorists need to address the question of how perception can guide action. How, as in our shadowing task, can percepts, coded in terms of their perceptual properties, "talk to" motor codes reflecting preparations for action? They conclude, and their research suggests, that there is a common coding (we prefer a common currency[3]) between perceptual objects and response preparations, that allows perceived properties to have an impact on planned responsive actions. When there is common currency and stimulus–response compatibility, percepts and planned actions share many properties, and this facilitates response preparation, as research reviewed by Hommel et al. has shown.

Kerzel and Bekkering (2000) draw a similar inference of common currency for the domains of speech production and perception. In their research participants saw a videotaped face mouth either /bʌ/ or /dʌ/, an irrelevant stimulus for the participant's task. Simultaneously with it, or a variable interval later, they saw a pair of symbols (either ## or &&) one of which cued a vocal /ba/ response whereas the other cued a /da/ response. Response times were faster when the observed vocal gestures of the irrelevant prime were compatible with the gestures that the speakers went on to produce. How can perceived gestures prime produced gestures? They can, if in both domains the relevant objects are gestures. We would apply the same account when speech is perceived acoustically as when it is perceived optically.

A similar idea, finally, can be found in Meltzoff and Moore's (1997) account of how even newborn infants can imitate the facial gestures of an adult, for example a tongue protrusion gesture. Infants can see the adult's tongue protruding between the lips, but not their own. They can feel their tongue proprioceptively and can feel their tongue protruding between their lips, but they cannot feel the tongue of the adult proprioceptively. How do infants (how does anyone) know how to imitate when information is cross-modal in this way? Meltzoff and Moore propose the idea of a supramodal representation, a common code. From stimulation, the infant develops representations of the adults' vocal organs and of their own that transcend the perceptual modes. The representations accomplish that by being about

[2] A reviewer suggested as an alternative account that two processes underlie responding. One is a fast automatic mapping from percept to response, which is responsible for response production at closure. The other is a slower, explicit choice of response with a time course closer to that of choice responses in the absence of stimulus–response compatibility. We note, however, that, even measured from release, response times are considerably faster in Experiment 1 (248 ms) than in Experiment 2 (637 ms).

[3] By *common code*, Hommel et al. are referring to featural representations of percepts or action plans. By *common currency* we are not referring to anything in the head of actors/perceivers. We refer to properties in the world. Speakers produce gestures and listeners perceive them (or, when the information they extract does not specify the gesture, they may misperceive them as other gestures). Gestures are the common currency.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

distal objects and events (tongues and protrusion gestures) not about proximal visual and proprioceptive stimulation. We focus here on the similar idea of the need for a common currency or code for crosstalk, now between perceptions across the sensory modalities.

We invoke the idea of common currency in conditions of high stimulus–response compatibility to interpret our findings that replicate those of Kozhevnikov et al. and Porter and colleagues. Choice response times can be very fast, because perceived events are the same kinds of things as produced events. There is a common currency. What must that currency be? We propose that it is gestural. Phonetic gestures, like the facial gestures of Meltzoff and Moore (1997), are abstractly equivalent across speakers. If so, perceivers must have rapid access to gestural information to perform the choice task as they do. The motor theory of speech perception and direct realist theories are the only theories of speech perception so far proposed in which listeners access gestural information when they perceive.

Even so, there is more than one way in which perceivers can have such access. As the motor theory and direct realist theory propose in somewhat different ways, they can use acoustic information about gestures as such. In these accounts, listeners perceive the gestures (or intended gestures in the motor theory) of the speaker. A second way, however, is compatible with acoustic theories of speech perception, augmented a little to account for the data of Experiment 1. In acoustic accounts (such as those of Diehl & Kluender, 1989; Massaro, 1998; Sawusch & Gagnon, 1995; and others) listeners extract acoustic features or cues from a speech signal and map the cues onto the abstract phonological category with which they are most compatible. No articulatory information is said to be accessed, and so the theories do not predict the results of Experiment 1. However, the theories could be modified just a little to allow articulatory access from acoustic perception. After all, listeners are talkers too, and it might be useful if the phonological categories that serve listening also serve talking. Accordingly, the augmentation of acoustic theories would be to allow articulatory properties as well as acoustic ones to be associated with phonological categories.

From this perspective, in the choice task, listeners perceive the disyllables acoustically, but the consequence of mapping the cues onto a phonological category is that articulatory properties are made available. Perhaps it is these properties, not the gestures of the speaker as perceived that listeners use to guide their choice responses.

We have concluded that Experiments 1 and 2 are most consistent with an idea that listeners to speech access gestural speech information in perception. Experiment 4 was not designed to provide additional evidence on that point. Rather, we designed Experiment 4 to distinguish the two ways outlined above of explaining how gestural information is accessed. That is, under the assumption that the prior experiments show that gestural information is accessed, we now ask which gestures guide responding in the choice task. In the motor theory and direct realism, they are the gestures of the model speaker as the listener perceives them (accurately or not). In the augmented acoustic account, they do not include the speaker's perceived gestures. They are gestural prescriptions for the talker's own productions.

Experiment 4

To discriminate the accounts, we asked whether choice responses were modulated by the manner of speaking of the model. They should be according to the motor and direct realist theories; they should not be according to the acoustic account that we have generated. We approximately doubled the voice onset times of half of the stop consonants that participants heard and shadowed. Increased VOTs imply a change at least in the phasing of the stops' oral constriction and devoicing gestures. If rapid shadowing occurs because listeners

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

perceive the (ostensible) gestural patterning of the speaker and use that perceived patterning to guide their shadowing responses, then shadowing responses on extended-VOT trials should exhibit longer VOTs than those on unchanged-VOT trials. In contrast, if listeners hear the acoustic signals produced by talkers, classify the tokens as tokens of a familiar phonological category, and use articulatory information stored with the category to guide responding, no VOT increases are expected on extended-VOT trials.

We ran this experiment twice, because error rates were high the first time relative to error rates in Experiment 1. We will refer to the two versions as Experiments 4a and 4b. We considered two reasons why error rates were high in Experiment 4a. One was that our extended-VOT items were unnatural, and this threw off performance on all trials. The second was that our participants were highly compensated graduate students in Experiment 1, but were undergraduates compensated by course credit in Experiment 4a. The latter group may have been less motivated than the former. We will briefly report the results from both experiments, because except for having reduced error rates, comparable to those in Experiment 1, Experiment 4b provides a close replication of 4a.

Methods

Participants—The 12 participants in Experiment 4a were University of Connecticut undergraduates who were native speakers of English and who reported no speech or hearing problems. All were compensated with course credit. The 12 participants in Experiments 4b were paid.

Materials—We selected one of the three tokens of each of the 24 utterance types of Experiment 1 and edited them using SoundScope to extend the VOTs. The average original and extended VOTs, respectively, were 73 and 130 ms, a 79% extension. To make the extended tokens, original VOTs were measured, and medial portions of the aspiration was copied and pasted back into the waveform. The selected extensions were pasted immediately after the original copied portion. The investigator editing the tokens (LS) selected portions of the waveforms that created a natural sounding extension of the VOT segment of the wave. The average duration of each selection was 26 ms; more than one selection was pasted into each VOT to extend the VOTs by an average of 57 ms. The exact duration and location of the portion selected for extension varied across edited utterances. It was not possible to generalize a location or duration of a section to be selected due to natural variations in the model's different productions of the CV syllables. In particular, it was necessary to ensure that each selection be free of highly identifiable acoustic information such as bursts or transitions. The selections were then made from the "cleanest" portion of the VOT segment, the relative placement of which varied across productions.

The original and extended VOT productions were digitized in SoundScope and saved as files for use in PsyScope.

Procedure—Each participant took part in an individual session that lasted approximately 25 min. Participants took a 144 item choice test following six practice trials. The 144 items included the 48 tokens presented three times each in random order using PsyScope.

Participants were told that they would hear a model saying a prolonged vowel /a/, and then the model would switch to either /pa/, /ta/ or /ka/. They were asked to shadow the vowel, then when the model switched to one of the three CV syllables, to switch as quickly as possible to the same CV syllable. In this experiment, participants were asked to be as accurate as possible and were advised not to sacrifice accuracy for speed. In this respect, speed was emphasized less than in Experiment 1. This was to help ensure that participants

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

heard the model's VOTs in time to allow them to have an effect on their own responses. Trials were self-paced.

Participants' responses were recorded on one channel of a cassette tape and the stimuli were recorded on the other. The two channels were digitized simultaneously using SoundScope, and latency and participants' VOT measurements were made.

Measurements in Experiment 4a were made by LS and JMB. For purposes of assessing reliability, 25% of the VOT data (data from three participants) were measured independently by both investigators. The inter-rater reliabilities were high for two of the participants (R^2 = .98 and .99), but modest for the third (R^2 = .72). That participant's VOTs were difficult to measure, because it was difficult to pinpoint when voicing began. However, the measurements were not biased in favor of finding an effect of VOT length. The mean differences in VOT between JMB and LS in the original and extended conditions for this subject were 6 and 7 ms, respectively, with JMB providing generally longer VOT measurements than LS. Accordingly, JMB did not provide longer VOTs especially in the extended VOT condition. In analyses below, for those participants on whom we had two sets of measurements, we took the measurements that were made first. Measurements in Experiment 4b were made by L.S. and J.W.

Results

Errors were common in Experiment 4a. Across participants, the average error rate was 21%. Error rates were 21 and 22% on extended and original VOT syllables, respectively. Error rates were 11 and 12% on the extended and original VOT trials in Experiment 4b. This difference between the experiments pinpoints the participant group rather than any unnaturalness of the extended VOT stimuli in the experiment. Error rates in Experiment 4b are quite similar to those of Experiment 1.

Latencies averaged 198 ms from closure and 385 ms from release in Experiment 4a. (We were unable to measure latencies for one participant, because, on the recording of the model's speech, the participants' speech had bled through making marking of model closure and release impossible.) The measurement from closure is about 15 ms slower than that in Experiment 1; the measurement from release is about 140 ms slower. This difference between the experiments may be consequent on our instructions not having emphasized speed as much as in Experiment 1. Corresponding latencies from Experiment 4b were 212 and 343 ms.

Fig. 3 presents participants' VOTs produced in response to the extended and original VOT syllables from Experiment 4a. Measures are presented separately for the three consonants, /p/, /t/, and /k/. An analysis of variance on VOT durations was performed with factors Model's utterance (original, extended) and Consonant (/p/, /k/, /t/). We collapsed over duration of initial /a/, because the high error rate produced some empty cells.

In the ANOVA, the effect of Model's utterance was highly significant (*F* (1, 11) = 18.75, *p* = .001); VOTs for shadowed extended utterances averaged 69 ms; those for original utterances averaged 61 ms. Although the effect of lengthening VOTs was small, it occurred numerically for each of the 12 participants. The effect of consonant was also significant (*F* (2, 22) = 19.67, *p* < .001). VOTs increased from /pa/ (57 ms) to /ta/ (67 ms) to /ka/ (71 ms) as expected (e.g., Zue, 1980). The interaction was not significant.

In the ANOVA performed on the VOTs of Experiment 4b, the effect of Model's utterance was highly significant (*F* (1, 11) = 22.54, *p* < .001); VOTs for shadowed extended utterances averaged 57 ms; those for original utterances averaged 53 ms. The 4 ms effect is

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

smaller than that of Experiment 4a; that direction of effect occurred for 11 of the 12 participants. The effect of consonant was also significant (F (2, 22) = 33.16, p < .0001). VOTs increased from /pa/ (44 ms) to /ta/ (57 ms) to /ka/ (63 ms). The interaction did not approach significance.

We asked whether the difference between extended and original VOTs was related to participants' response times. Possibly fast responders responded too quickly sometimes to hear the model's VOTs. The correlations between the difference in VOTs of original vs extended model VOTs (extended minus original) and participants' average latency on extended VOT trials measured from closure or release were nonsignificant. The correlation between VOT difference and latency from release did approach significance, but in Experiment 4a only (r = .56, p = .07). In that analysis, larger VOT differences were associated with slower responses. We performed the same analysis on the data of individual participants to ask whether, on extended VOT trials on which response latencies were slow, participants showed more lengthening than on trials where latencies were faster. In short, in both Experiments 4a and 4b, very few analyses reached significance, and those that did were equally likely to be in the predicted and the nonpredicted direction. There is no evidence that more lengthening of VOTs occurred when latencies were long.

Discussion

Error rates were much higher in Experiment 4a than in Experiment 1. The findings of Experiment 4b suggest, however, that the reason for the difference is not because of any perceived unnaturalness of our extended VOT stimuli. The paid participants of Experiment 4b had an error rate comparable to those of the paid participants of Experiment 1. Therefore, we ascribe accuracy difference between Experiment 4a and Experiments 1 and 4b to a difference in participant group and, perhaps level of motivation.

In Experiments 4a and 4b, participants' VOTs were significantly longer when they shadowed lengthened VOTs, as predicted by a hypothesis that perceived gestures of the model guided participant responses. If acoustic information is used to access phonological categories, and attached to each category is motor information that is used to produce the category member, then influences from the model should not have occurred under the conditions of Experiment 4. That participants lengthened their own VOTs when they shadowed lengthened-VOT syllables shows that the information guiding their shadowing response included information about the model speaker's gestures. This is the account of the motor theory and of direct realism. Participants can imitate unusually long VOTs, because their perception of phonemes includes motor information.

To different extents, both the augmented acoustic perceptual theory that we offered and gestural theories can explain the results of Experiment 1. However, in our view, gesture theories provide the best account of the data of Experiment 4.

A criticism of the gestural interpretation can be that participants' responses to extended VOT items did not have VOTs anywhere near as long as the model's productions. If participants were perceiving gestures that guided their responses, why did not they lengthen their VOTs to a much greater extent? This question can be addressed in several ways.

First, we should not be surprised that talkers' own habitual ways of producing /pa/, /ta/, and / ka/ have a major, even predominant effect on VOTs in the experiment. However, a significant difference of *any* magnitude between participants' VOTs on original and extended VOT trials provides information that distinguishes the acoustic and gestural accounts of the findings of Experiment 1 that we offered. All but one of the 24 participants showed the expected direction of VOT difference; this suggests that they perceived the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

model's gestures and that the perceived gestures modulated their own characteristic speech behavior.

Second, this magnitude of effect is not unprecedented. Sancier and Fowler (1997), found a similar magnitude of VOT lengthening in an individual who traveled between English and Brazilian-Portuguese speaking populations. After some *months* of experience in the United States, the participant's VOTs in both languages were approximately 6 ms longer than VOTs following experience in Brazil. Our participants showed a 6 ms shift (averaging across Experiments 4a and 4b) with about 20 min exposure.

A final possibility relates to our finding that our instructions de-emphasizing speed in this experiment had only a small impact on latencies measured from closure. That is, response times measured from closure were just 15 ms slower in Experiment 4a (29 ms in Experiment 4b) than in Experiment 1. If our earlier calculations are correct, much of the time participants were initiating consonant production before they could have taken the information in the model's release into account. Of course, they did hear the model's release before they produced their own. The small effects we see may reflect any late adjustments in laryngeal timing they may have made. Our correlational analyses rule out that participants only produced lengthened VOTs on trials where their response latencies were slow.

General discussion

Our research provides a close replication of Porter and Castellanos' (1980) earlier study. Measured from the earliest point in the acoustic signal at which we can detect that our participants had begun to produce a consonant, in Experiment 1, we found just a 26 ms latency difference between their simple and choice response times. This is well below the canonical 100–150 ms difference between responses in these tasks as reported by Luce (1986). Following Porter and Castellanos, we tentatively inferred from this small difference that the choice task almost eliminates the element of choice, because listeners perceive gestures, and their task is to reproduce them. We obtained a compatible finding in our simple response task. There we found that response latencies were significantly shorter on those trials on which the model's CV matched that of the participant. Although the difference was small (just 7 ms), it was present in the responses of all but one of the participants.

In case the comparison between simple and choice response times is rendered questionable, because of the possibility that there are different speed–accuracy trade-offs in performance of the two tasks, we made a different comparison in Experiment 2. Here we used a task in which the element of choice is very much present and found much longer response times than in Experiment 1.

An unexpected finding of Experiment 1 was that latencies measured from release rather than from closure were much longer in the choice task, leading to a difference between choice and simple latencies that was in line with canonical estimates of the difference. Experiment 3 and an acoustic analysis of the participants' responses in Experiment 1 were designed to determine whether the latency measures from closure were spurious (because participants may have ceased producing the vowel but not because they had achieved consonant closure) and to determine a reason why latencies from release were not in line with those from closure. Experiment 3 verified that listeners were sensitive to the earliest information about the consonant in the model's speech—that is, information in the closing transitions of the first vowel. Acoustic analysis verified that participants in Experiment 1 also distinguished among the consonants at the offset of their first vowel. Accordingly, their vowel offsets were, in fact, consonant closure onsets and, therefore, appropriate points at which to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

measure the participants' latencies. Moreover, latencies measured from the model's release to the participants closure were improbably short, particularly for two of the six participants to conclude that they had waited until the model's release to initiate their consonantal responses. Accordingly, we infer that measurements from model closure to participant closure provide appropriate measures of latency, and that measures from release were longer only because participants were unconfident of their consonantal response and were giving themselves time to make a rapid change in constriction location should information at release warrant it.

Experiment 4 provided support for the interpretation that participants' shadowing responses are guided by their perception of the model's gestures, not merely by perception of acoustic signals, followed by classification into phonological categories to which articulatory information is linked. In that experiment, when the model's VOTs were long on a given trial, so were those of the participant, albeit by a much smaller amount. In a theory that perceptual objects are acoustic, the model's own VOTs should have no effect on the participants'. However, they had a highly significant effect.

These findings are most compatible with the motor theory of speech perception and direct realist theory. The motor theory provides the simplest account of the findings. Perceiving the model's CV engages the speaker's own speech motor system to recover the gestures that produced the signal. This may prime the production system specifically to produce the gestures overtly that it just generated internally.

Direct realism provides a different account that is also compatible with the findings. In that theory, listeners perceive the speaker's gestures, because the gestures have causally and distinctively structured the acoustic speech signal, which, therefore, provides information about them. In the choice task, these perceived gestures constitute instructions for the required response. In the simple task, they constitute instructions for the response on one-third of the trials, and responses on these trials, accordingly, are faster than responses on the others.

The motor theory of speech perception and direct realism are the only theories in the literature of which we are aware that predict the set of findings across the four experiments. Undoubtedly, acoustic theories can be adjusted to post-dict them. However, prediction prevails.

It is an interesting question whether our findings are special to speech. According to the motor theory of speech perception, the results should not obtain for other acoustic stimuli, because percepts of other acoustic stimuli are of the proximal, not the distal stimulus. The theory is almost silent on expectations given stimuli from other modalities. It might predict a replication of the findings of Experiment 1 when stimuli are perceived by specialized brain modules (e.g., for speech, a phonetic module). The expectation from direct realism is that the findings are not special to speech. A small difference should obtain between choice and simple response times whenever there is common currency between perception and action planning (so always, according to direct realist theory) and sufficient stimulus–response compatibility.

Consider first an intuitively clear example in which stimuli are visual. If participants in an experiment respond by button pressing to observed button presses, there is common currency (assuming that visual perceivers see distal events (buttons being pressed) rather than proximal ones (changing patterns of reflected light)). There is stimulus–response compatibility because, in response to observed button presses, participants press buttons (and, in the choice task, press the buttons corresponding to those pressed by the model). A simple response test would provide each participant with a designated button. The model

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and participant would sit pressing a home key for an unpredictable interval. When the model shifted to press one of three response buttons, participants would press their designated button. In the choice task, participants shift to press the response button corresponding to that pressed by the model. Direct realist theory predicts the same outcome as in Experiment 1. Simple and choice response times should be close, and simple responses should be faster when the model presses the response button corresponding to the participant's designated button.

The theory predicts a similar outcome when stimuli are acoustic, but nonspeech. For example, modeled stimuli might be a kissing sound, a lip smack, and a raspberry. Or they might be an audible finger tap, a hand clap and a finger snap. The motor theory does not predict a replication of the findings of Experiment 1 in this case, because, in that theory, percepts of acoustic stimuli, speech excepted, are proximal not distal. Accordingly, there is no common currency between stimulus and response. The direct realist theory predicts a replication of Experiment 1 whether stimuli are heard or heard and observed. The critical elements of the experiments should be common currency between (distal) percepts and planned actions and stimulus–response compatibility.

References

Best, C. A direct realist perspective on cross-language speech perception. In: Strange, W., editor. Cross-language speech perception. Timonium, MD: York Press; 1995. p. 171-204.

Cohen J, MacWhinney B, Flatt M, Provost J. PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behavior Research Methods and Instrumentation 1993;25:257–271.

Diehl R, Kluender K. On the objects of speech perception. Ecological Psychology 1989;1:121–144.

Fadiga L, Craighero L, Buccino G, Rizzolatti G. Speech listening specifically modulates the excitability of tongue muscles: A TMS study. European Journal of Neuroscience 2002;15:399–402. [PubMed: 11849307]

Fadiga L, Fogassi L, Povesi G, Rizzolatti G. Motor facilitation during action observation: A magnetic stimulation study. Journal of Neurophysiology 1995;73:2608–2611. [PubMed: 7666169]

Fowler CA. An event approach to the study of speech perception from a direct realist perspective. Journal of Phonetics 1986;14:3–28.

Fowler, CA. Encyclopedia of language and linguistics. Vol. 8. Oxford: Pergamon Press; 1994. Speech perception: Direct realist theory; p. 4199-4203.

Fowler, CA. Imitation as a basis for phonetic learning after the critical period. Paper presented at the Twenty-fifth Annual Meeting of the Berkeley Linguistics Society; Berkeley, California. 2000.

Giles, H.; Coupland, N.; Coupland, J. Accommodation theory: Communication, context, and consequence. In: Giles, H.; Coupland, J.; Coupland, N., editors. Contexts of accommodation: Developments in applied sociolinguistics. Cambridge: Cambridge University Press; 1991. p. 1-68.

Goldinger S. Echoes of echoes? An episodic theory of lexical access. Psychological Review 1998;105:251–279. [PubMed: 9577239]

Goldstein, L.; Fowler, CA. Articulatory phonology: A phonology for public language use. In: Schiller, NO.; Meyer, A., editors. Phonetics and phonology in language comprehension and production: Differences and similarities. Berlin: Mouton de Gruyter; (in press)

Guenther F, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech. Psychological Review 1998;105:611–633. [PubMed: 9830375]

Hommel B. Inverting the Simon effect by intention. Psychological Research 1993;55:270–279.

Hommel B, Musseler J, Aschersleben G, Prinz W. The theory of event coding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences 2001;24:849–878. [PubMed: 12239891]

Houde J, Jordan M. Sensorimotor adaptation in speech production. Science 1998;227:1213–1216. [PubMed: 9469813]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Kerzel D, Bekkering H. Motor activation from visible speech: Evidence from stimulus–response compatibility. Journal of Experimental Psychology: Human Perception and Performance 2000;267:634–647. [PubMed: 10811167]

Kluender, K. Speech perception as a tractable problem in cognitive science. In: Gernsbacher, M., editor. Handbook of psycholinguistics. San Diego: Academic Press; 1994. p. 21-173.

Kozhevnikov, V.; Chistovich, L. Speech: Articulation and perception. 543. Vol. 30. Washington, DC: Joint Publications Research Service; 1965.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review 1967;74:431–461. [PubMed: 4170865]

Liberman A, Mattingly I. The motor theory revised. Cognition 1985;21:1–36. [PubMed: 4075760]

Liberman AM, Whalen DH. On the relation of speech to language. Trends in Cognitive Sciences 2000;4:187–196. [PubMed: 10782105]

Lotto AJ. Language acquisition as complex category formation. Phonetica 2000;57:189–196. [PubMed: 10992139]

Luce, RD. Response times. New York: Oxford University; 1986.

Martin J, Bunnell HT. Perception of anticipatory coarticulation effects. Journal of the Acoustical Society of America 1981;69:559–567. [PubMed: 7462478]

Martin J, Bunnell HT. Perception of anticipatory coarticulation effects in vowel–stop consonant–vowel syllables. Journal of Experimental Psychology: Human Perception and Performance 1982;8:473–488. [PubMed: 6212634]

Massaro, D. Perceiving talking faces. Cambridge, MA: MIT Press; 1998.

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976;264:746–748. [PubMed: 1012311]

Meltzo A, Moore MK. Explaining facial imitation: A theoretical model. Early Development and Parenting 1997;6:179–192.

Porter R, Castellanos F. Speech production measures of speech perception: Rapid shadowing of VCV syllables. Journal of the Acoustical Society of America 1980;67:1349–1356. [PubMed: 7372922]

Porter R, Lubker J. Rapid reproduction of vowel–vowel sequences: Evidence for a fast and direct acoustic–motoric linkage. Journal of Speech and Hearing Research 1980;23:593–602. [PubMed: 7421161]

Prinz W. Perception and action planning. European Journal of Cognitive Psychology 1997;9:129–154.

Reed, E. Encountering the world. New York: Oxford University Press; 1996.

Rizzolatti G, Fadiga L, Gallese V, Fogassi L. Premotor cortex and the recognition of motor actions. Cognitive Brain Research 1996;3:131–141. [PubMed: 8713554]

Saltzman E, Munhall K. A dynamical approach to gestural patterning in speech production. Ecological Psychology 1989;1:333–382.

Sancier M, Fowler CA. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. Journal of Phonetics 1997;25:421–436.

Sawusch J, Gagnon D. Auditory coding, cues and coherence in phonetic perception. Journal of Experimental Psychology: Human Perception and Performance 1995;21:635–652. [PubMed: 7790838]

Simon, H. The sciences of the artificial. Cambridge, MA: MIT Press; 1969.

Simon JR, Hinrichs JV, Craft JL. Auditory S–R compatibility: Reaction time as a function of ear–hand correspondence and ear–response–location correspondence. Journal of Experimental Psychology 1970;86:97–102. [PubMed: 5482039]

Van Orden G, Holden J, Turvey MT. Self-organization of cognitive performance. Journal of Experimental Psychology: General. (in press).

Wallace RA. S–R compatibility and the idea of a response code. Journal of Experimental Psychology 1971;88:354–360. [PubMed: 5090926]

Whalen DH. Subcategorical mismatches slow phonetic judgments. Perception & Psychophysics 1984;35:49–64. [PubMed: 6709474]

Zue, V. Acoustic characteristics of stop consonants: A controlled study. Bloomington, IN: Indiana University Linguistics Club; 1980.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

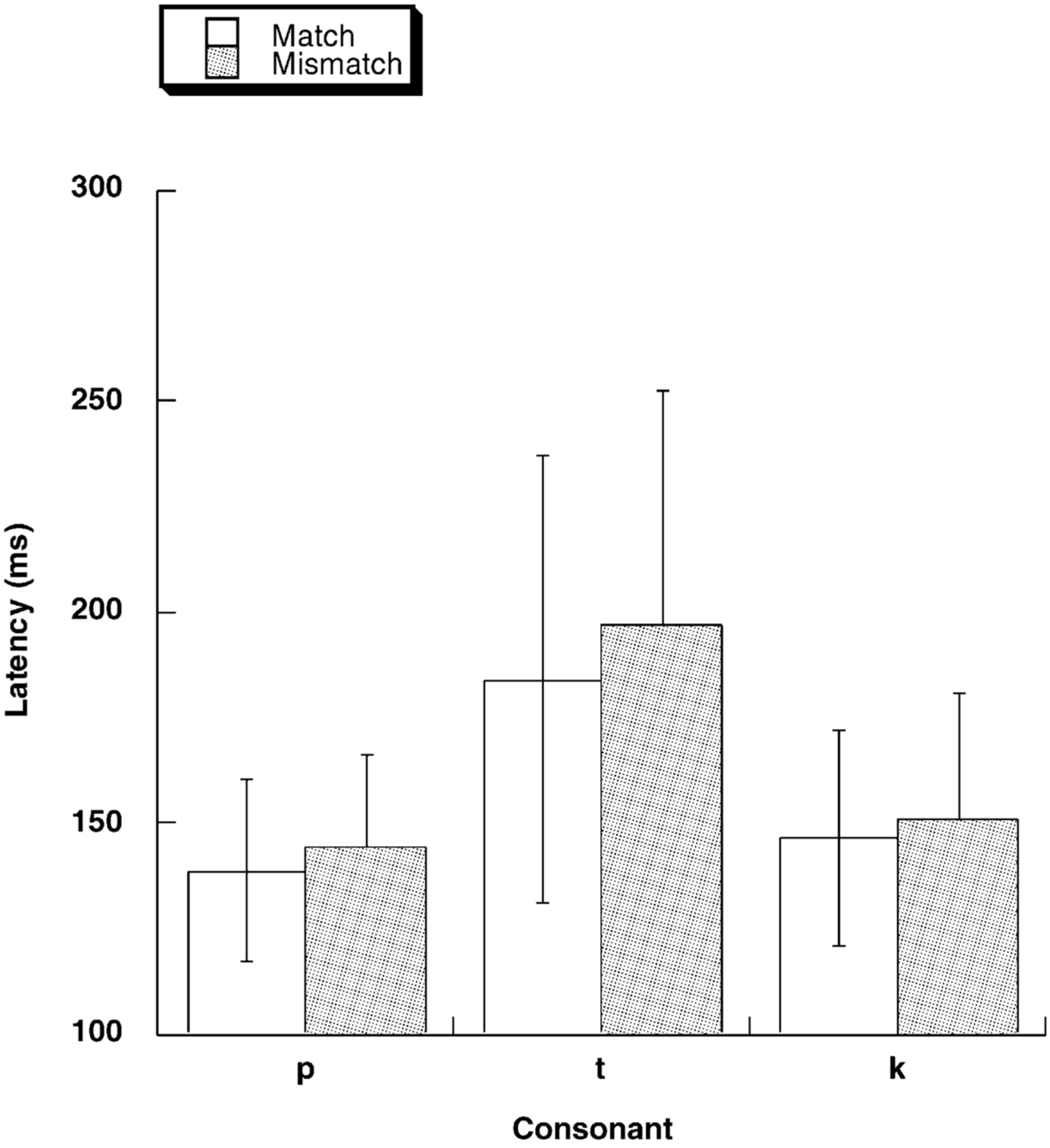

Fig. 1.
Simple response times in Experiment 1 as a function of match or mismatch between the model's and the participant's response consonant.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

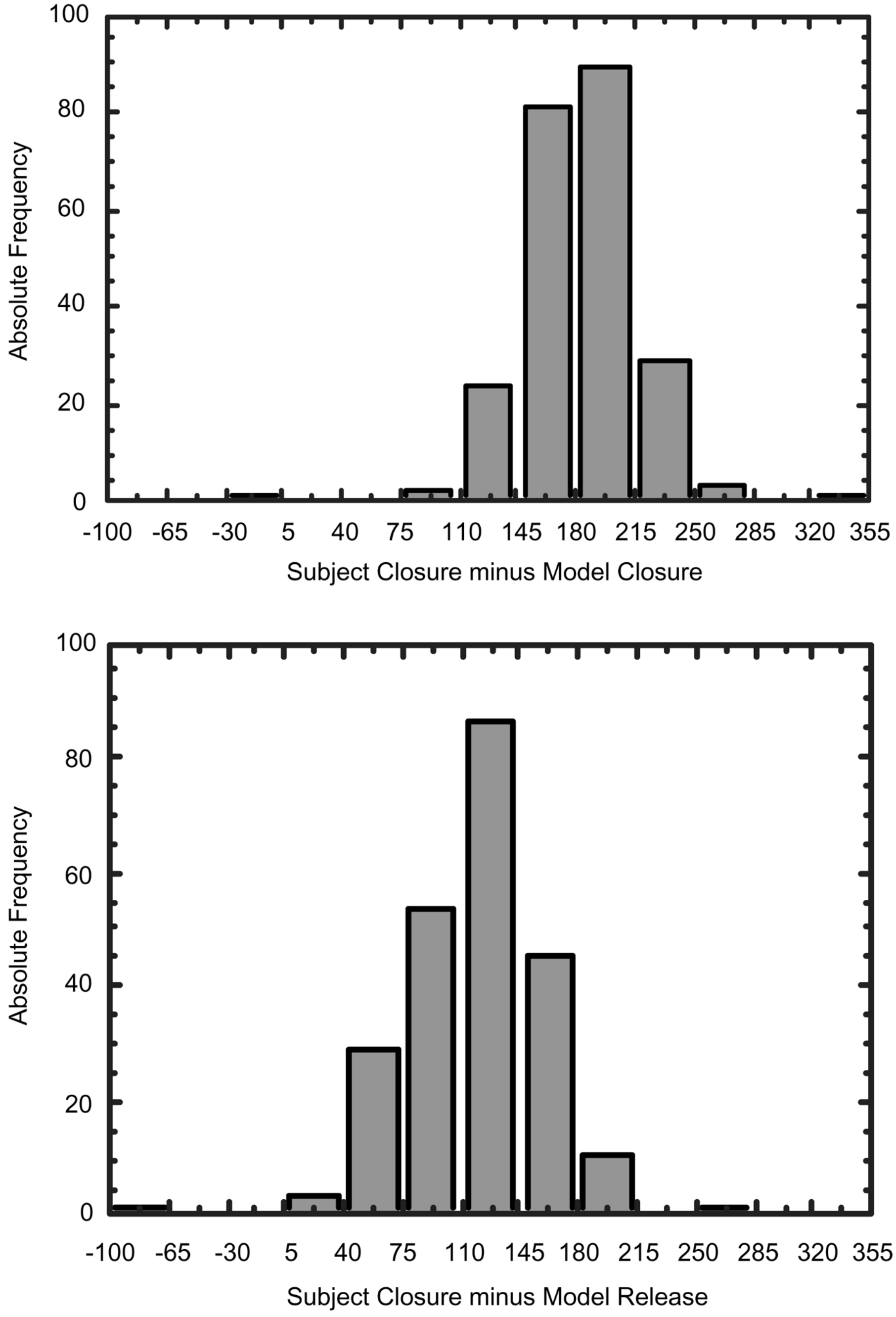

Fig. 2.

Histograms of latency differences between participant closure and model closure and between participant closure and model release.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

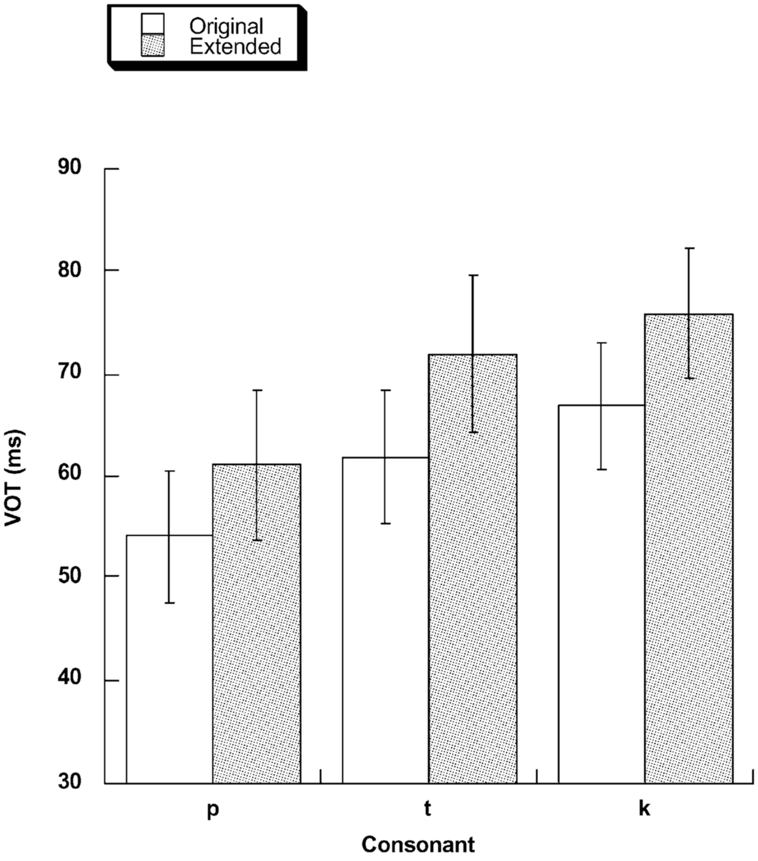

Fig. 3.
Mean VOTs (and standard errors) of participants' voiceless stops, presented separately for the original and extended conditions and for the consonants, /p/, /t/, and /k/.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Average latencies (in ms) measured from closure and from release in the choice and simple response tasks of Experiment 1

Task/measurement point	Experiment 1	
	Closure	Release
Choice response	183	248
Simple response	157	157
Difference	26	91