Psychonomic Bulletin & Review
2008, 15 (2), 453-457
doi: 10.3758/PBR.15.2.453

Notes and Comment

The motor theory of speech perception revisited

Dominic W. Massaro and Trevor H. Chen
University of California, Santa Cruz, California

Galantucci, Fowler, and Turvey (2006) have claimed that perceiving speech is perceiving gestures and that the motor system is recruited for perceiving speech. We make the counter argument that perceiving speech is not perceiving gestures, that the motor system is not recruited for perceiving speech, and that speech perception can be adequately described by a prototypical pattern recognition model, the fuzzy logical model of perception (FLMP). Empirical evidence taken as support for gesture and motor theory is reconsidered in more detail and in the framework of the FLMP. Additional theoretical and logical arguments are made to challenge gesture and motor theory.

It is reasonable to believe that speech perception and speech production are intimately related. In most research, however, these behaviors have been studied separately. This research is based, at least implicitly, on the assumption that much of behavior can be carved at its joints and some behaviors can be studied independently of others. Galantucci, Fowler, and Turvey (2006; hereafter, GFT) reviewed productive research when these processes were studied together. Their review evaluated three central propositions of the motor theory: "speech processing is special" (p. 364), "perceiving speech is perceiving gestures" (p. 365), and "the motor system is recruited for perceiving speech" (p. 367). They favored the last two claims and opposed the thesis that speech processing is special. We heartily agree that speech is not special (one of us has advocated this for over 3 decades) but disagree with the latter two claims. We contrast an alternative theory with motor theory in explaining results that speak directly to these two claims, rather than results that simply show that perception and production are simultaneously active but not necessarily causally related.

The alternative we propose is a prototypical pattern recognition model, which explains perceptual processing independently of motor behavior. The fuzzy logical model of perception (FLMP) was developed to account for several important empirical phenomena. The FLMP's major assumptions are that (1) multiple sources of information influence speech perception, (2) perceivers have continuous information, not just categorical information, about each source, and (3) the multiple sources are used together in an optimal manner (Massaro, 1998). Figure 1 illustrates the FLMP's three major operations in pattern recognition: evaluation, integration, and decision. The three perceptual processes are shown to occur left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. In this hypothetical situation given face-to-face dialogue, the evaluation process transforms these sources of information into psychological values, which are then integrated to give an overall degree of support for each speech alternative. The implicit decision operation maps the outputs of integration into some interpretation, which in behavioral experiments can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

Differences between the perceptual and the learning processes are also schematized in Figure 1. Perception is a feedforward process, in the sense that processing outcomes at a later stage do not feedback and influence earlier stages. Similarly, top-down contextual effects do not modify bottom-up perceptual processes. Feedback after perception is assumed to tune the prototypical values of the features used by the evaluation process.

The second claim of GFT (2006), that perceiving speech is perceiving gestures, is related to Gibson's (1966) view of direct perception, in which we perceive the distal, rather than the proximal, world; that is, we perceive the distal events, rather than the proximal sensory input. We agree that our experience is often of distal events, but it is debatable whether the appropriate distal event for speech perception is the gesture. We might, instead, simply perceive speech in terms of a prototypical pattern recognition process, using sensory input along with various contextual constraints. In language processing, the goal is to understand, which is not necessarily linked to the actual gestures of the speaker/writer. Understanding is broadly conceived as including perception of nonlexical utterances, such as coughs, nods of agreement, syllables, meaningful words, and semantic, syntactic, and pragmatic context. The FLMP has successfully described the influence and integration of these top-down and bottom-up sources of information in speech perception (Massaro, 1987, 1996, 1998; Movellan & McClelland, 2001).

GFT (2006) used four sets of evidence to support gesture as the object of speech perception. The first is the lack of signal–phoneme invariance in auditory speech, illustrated by the well-known /di/–/du/ schematic spectrograms. These show that the acoustic information for the phoneme /d/ is variable in different vowel contexts, whereas putatively the speech gesture is not. However, we believe that the speech gesture cannot be any more invariant than its acoustic consequences, because the latter event is directly formed by the former event. In fact, there is now some evidence that

D. W. Massaro, massaro@fuzzy.ucsc.edu

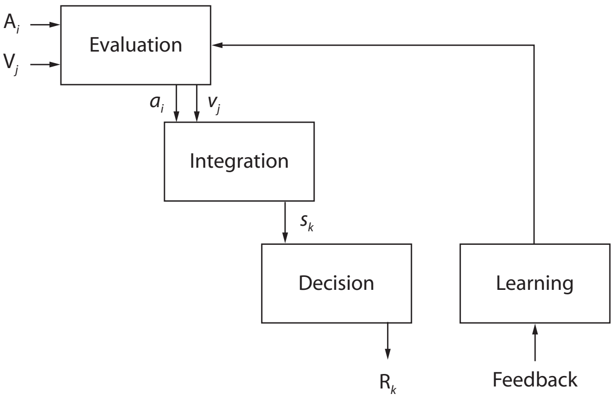

Figure 1. Schematic representation of the fuzzy logical model of perception for the processing of audible and visible speech in face-to-face dialogue. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k, which could be as small as a speech segment or as large as an utterance interpretation. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The learning process is also included in the figure. Feedback at the learning stage is assumed to tune the psychological values of the sources of information used by the evaluation process.

there may be less variability in the acoustic consequences of articulation than in the articulation itself. Vocal tract imaging and tracking techniques indicate that American speakers produce /r/ with many different tongue shapes, and yet all of these are perceived as /r/ (Nieto-Castanon, Guenther, Perkell, & Curtin, 2005). What appears to be critical for a /r/ to be perceived as such is that its acoustic stimulus must have a very low third formant (although we know that many other characteristics of the spectrum are influential, such as the direction of the formant transitions).

To solve the invariance problem between acoustic signal and phoneme, while simultaneously adhering to a preperceptual auditory memory constraint of roughly 250 msec, Massaro (1972) proposed the open syllable V, CV, or VC as the perceptual unit, where V is a vowel and C is a consonant or consonant cluster. This assumption was built into the foundation of the FLMP (Oden & Massaro, 1978). Assuming that this larger segment is the perceptual unit reinstates a significant amount of invariance between signal and percept. Massaro and Oden (1980, pp. 133–135) reviewed evidence that the major coarticulatory influences on perception occur within open syllables, rather than between syllables. Any remaining lack of invariance across open syllables could conceivably be disambiguated by additional sources of information in the speech stream. Thus, signal–phoneme variability does not support gesture theory and does not preclude prototypical pattern recognition processes.

The second finding cited as support for gesture theory is that vocal tract activity can be picked up directly from touching the speaker's mouth (Fowler & Dekle, 1991). College students placed their hands on a talker's mouth while she silently uttered one of the two syllables /ba/ and /ga/. Their task was to identify an auditory syllable from a /ba/-to-/ga/ continuum presented alone or presented simultaneously with the mouth movement. The mouth movement influenced the identification of the auditory syllable. GFT (2006) viewed this result as exactly analogous to the McGurk effect. In a reanalysis of their third experiment, however, Massaro (1998, pp. 352–355) used quantitative model tests to determine the size of the tactile effect and whether there was convincing evidence that the tactile and the auditory speech were actually integrated. The tactile effects were very small, and integration and nonintegration models gave indistinguishable descriptions of the results. Massaro's (1998) interpretation of the small influence of tactile speech and the failure to find integration was "that some learning is probably necessary to develop the ability to use a novel source of information" (p. 354).

Fowler and Dekle's (1991) first experiment showed bigger effects, which we will consider here. As can be seen in the right panel of Figure 2, the tactile input did appear to influence the identification of the auditory syllable. However, 10 of the 22 subjects identified the felt syllables at a chance level, and the authors plotted only the remaining 12 subjects. Thus, the size of the tactile effect shown in Figure 2 is probably overestimated by about 45%. In addition, contrary to the typical finding when two sources of speech information are factorially varied, the influence of the tactile input was not larger when the auditory information was ambiguous. This observation is borne out by the fit of the FLMP (RMSD = .0546), which was indistinguishable from the fit of a nonintegration model (RMSD = .0523).

We now will evaluate whether the tactile information influences speech perception in the same manner as the visible influence from the talker's face, as in the McGurk effect. Two potentially important aspects of their study are that the subjects were given instructions to report what they heard and that the speech sounds were presented over speakers, rather than over headphones. These conditions were present in an early study in which subjects were given auditory and auditory–visual syllables and had to identify which syllable they *heard* (Massaro, 1987, p. 68). The results of this study are shown in the left panel of Figure 2. Like many other results in a variety of conditions (Massaro, 1987, 1998), these results document a large visual influence, especially when the auditory speech is ambiguous. This type of significant interaction is a signature prediction of the FLMP. The fit of the FLMP (RMSD = .0420) was twice as good as the fit of a nonintegration model (RMSD = .0883). Given the differences in the fine-grained results in the tactile and visible speech experiments, it is premature to conclude that tactile–auditory speech perception is analogous to auditory–visual speech perception.

As was noted earlier, one possible hypothesis is that the subjects in the tactile condition had no experience in bimodal speech perception in which information was given from the hand and the ear. Clearly, further research is warranted and, perhaps, would provide a potential test between the gesture theory and the FLMP. Learning the proximal cues for speech categories is essential for the FLMP, whereas it seems less so for gesture theory. In car-

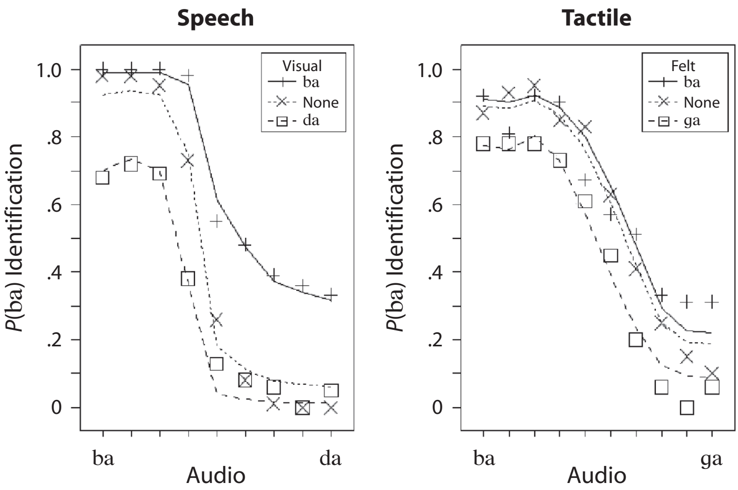

Figure 2. Left: Observed (points) and predicted (lines) probabilities of /ba/ judgments as a function of the auditory and visual information—Predictions of the fuzzy logical model of perception (estimated observed results from Massaro,1987, Figure 9, p. 68). Right: Observed (points) and predicted (lines) probabilities of /ba/ judgments as a function of the auditory and tactile information—Predictions of the fuzzy logical model of perception (estimated observed results from Fowler & Dekle, 1991, Figure 2). Note that the auditory continua in the two panels, although different, are reasonably comparable, because both vary monotonically in small steps between two speech alternatives.

rying out new research, however, it would be more valid to use a physical analogue to Baldi, our virtual animated agent, in order to control the tactile input exactly and even vary its ambiguity. A symmetrical expanded factorial design would provide the strongest test, and model tests should be carried out on the results from individual subjects, rather than simply group results (Massaro, 1998).

We are using the term *McGurk effect* because of its familiarity, but we employ it to mean the interaction between auditory and visual speech, which the FLMP has quantitatively predicted in many different individuals, languages, and tasks (Massaro, 1998, chaps. 5 and 6). Although the McGurk effect has been used as support for both gesture and motor theory, we might ask how these theories actually explain and predict the outcome. The auditory input might be /ba/ and the visual input /ga/, and perceivers often experience /da/, /va/, or /ða/. Given that there are contradictory inputs to the motor system, which putatively would elicit contradictory motor actions, how does the perception emerge? The explanation of the FLMP, on the other hand, gives an exact algorithm for prediction that can be quantitatively tested against the results (see Figure 2).

The third result involves how quickly we can shadow speech. The relevant hypothesis is that the additional time required for choice responses, relative to simple responses, is much less for speech than in other domains. Fowler, Brown, Sabadini, and Weihing (2003) asked subjects to always respond /pa/ when /pa/, /ta/, or /ka/ were presented (simple response time [RT]) or to repeat back the exact syllable that was said (choice RT). RT was measured from the onset of closure of the stimulus to the onset of closure of the response or from the release burst of the stimulus and response. They found a 26-msec increase in choice over simple RT when RT was measured from the closure period but a 92-msec increase when measured from the release. They did not convincingly account for this discrepancy and favored the smaller difference, even though the larger difference was compatible with the differences that are usually found in other domains. Given that the choice response was always one of three voiceless stop consonant syllables, the subject might initiate a voiceless stop closure without yet knowing the exact stop to utter. The error rate in the choice task was about 13%, relative to just 1% in the simple task. The extra $92 - 26 = 66$ msec was evidently required to determine the exact consonant to utter once the closure was initiated. On the basis of the existing literature, we do not find that speech imitation is exceptionally fast, relative to other domains that do not involve imitation but have similar stimulus–response compatibility.

Simple and choice tasks are a promising avenue for testing between gesture theory and the FLMP. Gesture theory predicts only a small increase in choice RT, relative to simple RT, when the stimuli and responses are spoken. This follows because speech stimuli supposedly provide nonarbitrary information for the spoken response: "What is perceived provides instructions for the required response—indeed, reducing the element of choice" (GFT, 2006, p. 366). We interpret this to mean that speech provides instructions for the required response because the gesture is the object of speech perception. A written letter, on the other hand, is not produced by a speech gesture and, thus, should not be so easily spoken. It follows that one test of this hypothesis would be to compare spoken and written stimuli with spoken responses. For example, one could use CV syllables versus letters as stimuli: /bi/, /di/, and /pi/ versus *B*, *D*, and *P*. Gesture theory would predict that the difference in RTs from simple and choice tasks would be significantly smaller with spoken than with written stimuli. The FLMP predicts that choice RT to letters will also be fast, because a spoken letter name is also a well-learned compatible response to a letter. Contrary to gesture theory, the FLMP predicts no significant advantage in the choice task with spoken stimuli, relative to written stimuli.

The fourth result used to support the gesture theory is the finding that perceivers appear to be sensitive to coarticulatory information in CV syllables. We agree with these empirical outcomes, although our interpretation differs from that of GFT (2006), who must describe how the perceiver parses the signal to recover the separate gestures of the consonant and vowel in a CV syllable. This operation is not necessary in the FLMP, in which the CV syllable is a perceptual unit. The FLMP is able to quantitatively predict the results of Mann and Repp (1980) and Whalen (1984) in the same manner that it has been able to predict so-called trading relations (Massaro, 1987, pp. 108–110). Coarticulatory constraints in production influence the acoustic and visual consequences, and perceivers learn these properties and use them in their prototype definitions (Massaro, 1998).

For the third claim that the motor system is recruited for perceiving speech, GFT (2006) reviewed findings from a variety of research areas to support a general and common link between perception and production/action. We endorse this proposal but question the motor theory. GFT used Kerzel and Bekkering's (2000) results to provide evidence for the motor theory. Subjects viewed the symbols && or ## (or the letters *ba* or *da*) and had to say /ba/ or /da/ as quickly as possible. Sometime before or simultaneously with the letters, the subjects viewed the lower half of a face articulating these same syllables. The talking face had a significant influence on the RT for producing the syllable indicated by the symbols or letters. In our interpretation, visible speech influences speech perception directly, which then can influence speech production, which is analogous to a Stroop effect.

Motor theory might predict that we perceive our own actions more accurately than those made by others. Subjects were much more accurate in recognizing audio recordings of their own hand-clapping than they were in recognizing the clapping of other familiar persons (Repp, 1987). This result does not demand a motor interpretation, because we probably attend more to our own hand-clapping than to that of others. If so, there would be better memory representations for our own clapping than for that of others, which would give a recognition advantage. However, this personal advantage does not seem to hold for speechreading. Schwartz and Savariaux (2001) used video-recorded speech and found that the 6 talkers could not recognize (speechread) their own videotaped utterances any better than those from the other talkers. We do not often watch ourselves speaking, as in a mirror, and, therefore, we will not have a personal advantage. If motor processes are functional in speechreading, however, there is reason to expect a personal advantage.

One of our biggest concerns about the motor theory is that seldom do its theorists describe how gesture and motor processing actually solve perceptual outcomes. What are the computational problems that must be solved, and how do gesture perception and motor processing actually provide algorithms for a solution? Given the interest in gesture and motor theories, we expect to learn how they make perception possible to solve. One potential reason to postulate motor processing is to somehow make perceptual processing faster and/or easier. As has been pointed out so persuasively by Rosenbaum (2002), however, motor programs and actions must solve the degrees-of-freedom problem to initiate a specific action and require substantial processing to be formed. Thus, any involvement of motor processes in perception may actually impose an additional burden on speech perception. We believe that gesture and motor theories have yet to address important issues in accounting for speech perception: What are the sources of information supporting speech perception; are the sources bottom up and/or top down; are they continuous or categorical; are the sources integrated; if integrated, how are they integrated; and what is the time course of perception? We are unsure how gesture and motor theory account for top-down context effects in speech perception. For example, how do perceiving gestures and motor recruitment explain the biasing influence of lexical constraints on phoneme identification (Massaro & Oden, 1995)? In contrast, the FLMP has adequately described the influence of phonological, lexical, syntactic, and semantic context (Massaro, 1987, 1998). By addressing these issues, we would obtain a more thorough understanding of gesture and motor theory, perhaps weak and strong versions, and the boundary conditions under which the theory holds.

The discovery of mirror neurons has apparently rejuvenated motor theories. A mirror neuron fires both when an animal performs an action and when the animal observes the same action performed by another animal (Rizzolatti & Craighero, 2004). Mirror neurons could serve as a basis for imitation and, therefore, learning. Our understanding, however, is that mirror neurons cannot account for perception, because they would overgeneralize. The macaque certainly experiences the difference between seeing a conspecific action and performing its own action, but the same mirror neurons are activated by these very different events and experiences. Therefore, mirror neurons alone cannot account for perception.

There is a paucity of specific algorithms of how motor processes contribute to perception. Vihman (2002), however, describes one possible alternative of how motor processes might work in speech acquisition. The infant practices canonical babbling and produces CV sequences at 6–8 months of age. This practice in production sensitizes the infant to similar input patterns, which are now easily recognized because they pop out of the acoustic stream. However, although these patterns would become familiar with practice and this increase in familiarity might facilitate perception, it does not mean that the motor processes involved in babbling were functional during perception. In fact, infant speech perception is much more sophisticated than what could be predicted from canonical babbling. We know that receptive language is acquired before productive language, so it is difficult to understand how motor behavior would contribute to speech perception. Toddlers (13- to 15-month-olds) have few words in their productive vocabulary but can compute the relations in a sentence (Hirsh-Pasek & Golinkoff, 1996). Furthermore, infants are highly sensitive to the statistical properties of segmental speech input at 6–8 months (Saffran, 2003), which could not be anticipated by canonical babbling. There are also significant correlations between individual infants' speech perception skills at 6 months and their language abilities at 13, 16, and 24 months (Tsao, Liu, & Kuhl, 2004). Thus, research on language acquisition does not support motor theory.

How does the FLMP account for language acquisition? It is assumed to be gradual and continuous and without innate categories. Infants gradually learn the meaningful distinctions in their language by weighting the appropriate sensory cues. If the caregiver asks a child to get a "ball" but the child does not yet know the auditory and visual cues to distinguish a /ba/ from a /da/, his or her perception will not completely resolve the /ba/. However, noticing that there is only a ball and not a doll, the child can now learn that the cues at word onset that he or she has just heard and seen were for /ba/ and not /da/. Similar learning opportunities for /da/ will eventually lead to the distinction between /ba/ and /da/. Learning these cues to accurately perceive these segments will, in turn, allow the child to have accurate auditory and visual targets in his or her speech production. According to the DIVA (directions into velocities of articulators) model (Guenther, 1995), auditory input (and we would add, visible input) help establish auditory targets for speech output. When the hearing infant speaks, auditory feedback from his or her own speech helps the infant attain accurate speech production.

Our conclusion is that motor theory, in its current formulation, is inconsistent with empirical evidence and logical analyses. The evidence reviewed by GFT (2006) reveals an association between perception and production/action, but they do not show what role, if any, motor processes play in perception. We conclude that there is still insufficient evidence that the gesture is the object of speech perception and that speech perception recruits motor-related processes. We look forward to future research on perception, production, and their interaction to advance this debate.

AUTHOR NOTE

This work was supported in part by NSF Grant BCS-9905176, Public Health Service Grant PHS R01 DC00236, cooperative grants from the Intel Corporation and the University of California Digital Media Program, the Perlino award, and other grants from the University of California, Santa Cruz. The authors are grateful for the technical help of Michael M. Cohen, the insightful comments of Bruce Bridgeman, Alan Kawamoto, Qiang Liu, Bill Rowe, David Rosenbaum, and Saul Sternberg, and the helpful reactions of Art Glenberg and the other reviewers. We are especially grateful for the many readings and productive discussions with Marcus Perlman. Correspondence concerning this article should be addressed to D. W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

REFERENCES

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory & Language*, **49**, 396-413.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 816-828.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, **13**, 361-377.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, **102**, 594-621.

Hirsh-Pasek, K., & Golinkoff, R. M. (Eds.) (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 634-647.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception & Psychophysics*, **28**, 213-228.

Massaro, D. W. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, **79**, 124-145.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1996). Integration of multiple sources of information in language processing. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 397-432). Cambridge, MA: MIT Press.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3, pp. 129-165). New York: Academic Press.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1053-1064.

Movellan, J., & McClelland, J. L. (2001). The Morton–Massaro law of information integration: Implications for models of perception. *Psychological Review*, **108**, 113-148.

Nieto-Castanon, A., Guenther, F. H., Perkell, J. S., & Curtin, H. D. (2005). A modeling investigation of articulatory variability and acoustic stability during American English /r/ production. *Journal of the Acoustical Society of America*, **117**, 3196-3212.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Repp, B. H. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America*, **81**, 1100-1109.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, **27**, 169-192.

Rosenbaum, D. A. (2002). Motor control. In H. Pashler (Series Ed.) & S. Yantis (Vol. Ed.), *Stevens' Handbook of experimental psychology: Vol. 1. Sensation and perception* (3rd ed., pp. 315-339). New York: Wiley.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, **12**, 110-114.

Schwartz, J.-L., & Savariaux, C. (2001). Is it easier to lipread one's own speech gestures than those of somebody else? It seems not! In D. W. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of International Conference on Auditory–Visual Speech Processing* (pp. 18-23). Santa Cruz, CA: University of California, Perceptual Science Laboratory.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, **75**, 1067-1084.

Vihman, M. M. (2002). The role of mirror neurons in the ontogeny of speech. In M. Stamenov & V. Gallese (Eds.), *Mirror neurons and the evolution of brain and language* (pp. 305-314). Amsterdam: Benjamins.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, **35**, 49-64.

(Manuscript received January 25, 2007;
revision accepted for publication July 30, 2007.)