A REALIST PERSPECTIVE ON SOME RELATIONS AMONG SPEAKING, LISTENING AND SPEECH LEARNING

Carol A. Fowler
Haskins Laboratories, New Haven CT USA

ABSTRACT

In a realist theory of speech perception, listeners perceive gestures of the vocal tract. These gestures can be shown to *be* the phonological components of utterances. Accordingly, by perceiving gestures, listeners perceive the talker's phonological message. I suggest that this tight coupling between what talkers do and what listeners perceive fosters listeners' imitation of talker's gestures and that this, in turn, fosters phonetic learning.

A REALIST PROGRAM OF RESEARCH ON SPEECH

In a realist theory of speech, speaking is a true expression of the phonological message that a talker intends to convey to a listener. That is, the phonological structures that a listener must perceive in order to recognize the speakers' words are the linguistically significant actions of the vocal tract that constitute speaking.

For their part, realist listeners hear the phonological message. They use structure in the acoustic signal, and sometimes in the optic array, not as structures to be perceived in themselves, but as information for their causal source in the world. In speech, the causal source of structure in the acoustic speech signal, and sometimes in the optic array, is, at bottom, the articulating vocal tract. As noted, however, appropriately described, the articulations of the vocal tract achieve the phonological constituents of spoken utterances. Accordingly, when listeners perceive what speakers do, they hear the talker's phonological message.

Studying the relation of speaking to listening

In any theory of speech, the relation of speaking to listening is an intimate one. Speaking and listening to speech jointly constitute the primary means by which linguistic communication can take place. However, the nature of the intimate relation of speaking to listening is different in a realist theory than in other theories, and that difference fosters a difference in the research programs that the different theoretical approaches are disposed to develop.

In most alternatives to a realist theory, phonological elements of spoken messages have their primary reality as covert, mental categories. Although these mental categories may be referred to as "phonological representations," they are not, in fact, considered to represent anything themselves. To the contrary, speakers represent *phonological categories* to listeners by moving their articulators so as to structure acoustic speech signals appropriately. In these views, articulation is a flawed vehicle for representing phonological segments, because coarticulation prevents iconic representation of their discrete, context-free character.

From this theoretical perspective, there is little to learn about listening by studying speaking aside from studying the acoustic signal that speaking creates, and speaking has, in fact, not been a central topic of investigation among most perception researchers. However, in the same way that the realist theorist, James Gibson devoted the first major section of his final book, *The ecological approach to visual perception* [1], to a description of the to-be-perceived ecological niche of a visual perceiver, so the realist investigator of speech must study what speakers do in order to understand what speaking makes available to be perceived. Therefore, studying speech production constitutes one important part of a realist research program the ultimate goal of which is to understand speech perception.

Research on speech production over the last approximately 15 years has provided an important new perspective on speaking that, however, has not changed the way that many perception researchers write about production. Investigators still cite with approval Hockett's striking metaphor in which coarticulated phonemes are likened to

smashed Easter eggs having passed through a wringer [2], or they refer to coarticulation as distortion [3]. However, in my opinion, findings on speech production show clearly that these characterizations are mistaken. Indeed, the recent findings make a central claim of a realist theory of speech perception plausible. It is that phonological primitives of spoken utterances are linguistically-significant actions of the vocal tract.

The major findings are these. Just as other components of the body do for every action that we perform, articulators of the vocal tract form transient coalitions during speech. These "synergies" are physiologically implemented couplings of articulators that are organized to achieve a task [4] or goal. Synergies are best detected in experiments in which an articulator is unexpectedly perturbed while it is moving in some direction. If the articulator is the jaw, and it is tugged down unexpectedly while it is raising for a bilabial closure [5], extra activity in a muscle of the upper lip can be detected within 20-30 ms of the perturbation, with the result that the upper lip lowers more on perturbed than on unperturbed trials, and the extra lowering compensates for the unexpectedly low position of the jaw. Biliabial closure is achieved despite the perturbation. Responses to perturbations are functional--that is, they are specifically responses that compensate for the perturbation [6-9]. That the latency of the responses is so short indicates that responses cannot arise far from the site of the perturbation. Synergies are physiological systems.

Of course, the function of synergies cannot be to counteract unexpected, externally applied tugs on the articulators. A plausible function is to compensate for internally applied tugs on the articulators arising in coarticulated speech. A low vowel coarticulated with a /b/ may tug the jaw down as it is raising for bilabial closure. The closure synergy ensures on-line compensation for that perturbation.

Synergies have two corresponding functional aspects. Their primary function is to achieve a linguistically-significant gesture. In doing that, however, in addition, they compensate for perturbing actions of coarticulating gestures. This second aspect can be elaborated further, thanks to the work on "coarticulation resistance" conducted largely by Daniel Recasens [10-13]. Recasens' work shows that synergies provide selective barriers to some coarticulatory actions. A consonantal gesture that requires a constriction to be made by the tongue dorsum will prevent or considerably reduce coarticulation by vowel gestures that also use the tongue dorsum. A consonantal gesture that does not require the tongue will not block vocalic gestures of the tongue dorsum. In short, synergies prevent coarticulation from being the destructive or distorting force that it has been characterized as being in the literature.

The import of these findings for a realist theory of speech perception has to do with the realist claim that phonological constituents of utterances are public actions of the vocal tract, not covert categories in the mind that those actions imperfectly represent. If we describe vocal-tract actions, not at the level at which we track movements of individual articulators, but at the physiologically real, coarser-grained level at which gestures are achieved (or, for some phonological segments, such as /m/ or /p/, the level at which gestural constellations are achieved), we see the context-independence required of the commutable phonological components of spoken words. Except in the most casual speech, bilabial closure is invariantly achieved when a speaker intends to produce a bilabial stop despite variation in the contributions to closure by the jaw, the lower lip and the upper lip. Coarticulation does not eliminate--indeed, synergies prevent it from eliminating--context-independence at the gestural level of description of vocal-tract actions for speech. These findings illustrate the importance for a theory of speech perception of understanding speech production.

Studying acoustic speech signals and listeners' attention to them

According to the realist theory, perception has a universal function that it must, therefore, serve the speech perceiver as well as the visual, auditory,

haptic, gustatory and olfactory perceiver. The function is to acquaint perceiver/actors with components of their ecological niche. This is an evolved function that natural selection has shaped perceptual systems to serve. Universally, then, perceivers use the structure in media that stimulate their sense organs--light for seeing, air for hearing, etc.--not as *objects* of perception, but as information for the part of the niche that caused the structure. Therefore, the second component of a realist research program is to discover how structure, largely in air, can specify gestures to perceivers.

This component of the research program, currently neglected, has to lag that on speech production, because we can only look for informative structure in the acoustic signal once we have identified the gestures and can determine how they should causally structure the air.

A final component of the realist research program is to study the perceiver's use of the acoustic speech signal. Realist perceivers should betray their use of the signal to recover gestures in two complementary ways. They should "parse" such unitary dimensions of the signal as its fundamental frequency (F0) into parts if distinct linguistic gestures have had converging effects on them. Complementarily, they should count the constellation of sometimes diverse acoustic consequences of a single gesture as a constellation that specifies the gesture.

In fact, listeners exhibit both symptoms of realist perceiving. For example, listeners judge intonational accents on high vowels (with high intrinsic F0) as lower or less prominent than accents, having the same F0, on low vowels (with low intrinsic F0) [14]. That is, they behave as if they have parsed F0 due to vowel intrinsic F0 from that due to production of the intonational melody. In turn, listeners hear the F0 that they ascribe to vowel intrinsic F0, not as pitch on the vowel, but as vowel height [15]. That is, vowel intrinsic F0 serves as part of the constellation of acoustic consequences of vowel production that provides information for the vowel.

Listeners' use of constellations to specify gestures is further indexed by their very poor discrimination, under some conditions, of syllables that differ in two ways as compared to their discrimination of syllables that differ in just one of those two ways [16]. This occurs, for example, when two syllables differ in the duration of a silent interval between [s] and [lIt] and/or in the presence or absence of labial transitions before the [lIt] sequence. Syllables with transitions are identified as "split" if the duration of silence is sufficiently long and as "'slit" if it is not. In the absence of transitions, more silence is required to shift the percept from "slit" to "split." If one syllable of a pair has a long silent interval, but no transitions whereas the second syllable of the pair has a shorter silence and transitions, listeners can fail to discriminate the pair members even though syllables differing only in presence or absence of transitions are highly discriminable. Out of context, of course, an interval of silence and transitions are highly discriminable. However, when they specify the same gesture (in the example, labial closure) in the context of a syllable, perceivers of gestures discriminate them poorly.

Studying the relation of speaking and listening to learning

There is a kind of symmetry in communicative events at the phonetic or phonological level of description: To speak is to engage in a kind of activity having linguistic significance that speakers share with members of their language community, and to listen is to perceive that activity by a speaker and to detect its significance. To listen successfully, then, is to achieve "parity" [17] in communication.

In this final section of the paper, I will suggest tentatively that some phonetic learning, which happens throughout a speaker's life, occurs due to this tight coupling between speaking and listening, which engenders a dispositon of listeners to imitate the speech they hear.

There is a striking set of findings in the literature that the discovers of the findings and I agree implies that speech listeners hear the actions of a speaker's vocal tract [18-20]. I am referring to findings by Kozhevnikov and colleagues and by Porter and colleagues regarding the latency with which listeners can imitate speakers. I will use these findings as foundations for drawing some inferences about a possible role for imitation in speech learning.

In general, in the literature on reaction time, it is well-known that "simple" reaction times are shorter than "choice" reaction times by 100 ms or more [21]. A simple reaction time procedure involves detection. For example, a subject might be instructed to press a button any time that a light flashes whether the light is red or blue or green. The subject merely has to detect the light and hit his or her one response button. In a comparable choice response task, the subject must hit a different button depending on the color of the light that flashes. Accordingly, the choice task involves not only detection, but also a mapping between the color of the light and the appropriate response button. It is not surprising that choice response times are longer than simple response times.

However, they are not always significantly longer. They were longer by a statistically nonsignificant 12 ms in the research of Porter and Lubker. In that research, the simple response task was to shift from producing the vowel [a] to another vowel [o] whenever a model speaker's vowel shifted from [a] to any of three vowels including [o]. In the choice task, listeners shifted from [a] to whatever vowel the model speaker had shifted to. Average simple response times for an [a] to [o] shift were 168 ms; corresponding average choice times were 180 ms, a nonsignificant difference.

I draw two inferences from these findings. One, following Porter and colleagues, I infer that listeners perceive vocal tract actions. The choice task involves almost no choice at all if listeners perceive the gestures of the talker, because perceiving the talker's gestures is, essentially, receiving instructions for the required response. If, instead, listeners hear the acoustic signal, the task is still a choice task: Listeners have to determine which gestures of their vocal tract will produce an acoustic signal corresponding to the one they heard.

The other inference is not warranted by these findings alone, but it is, I think, suggested by these findings considered in the context of relevant others. The inference is that imitation is dispositional, and this disposition to imitate, I propose, leads to some speech learning.

Consider two findings, one on infants and one on adults that suggest a disposition to imitate. Meltzoff and colleagues [22, 23] find that newborn infants imitate the facial expressions of adults. For example, infants will protrude their own tongue in the presence an adult protruding his or her tongue. This, of course, is especially remarkable, because the newborn cannot see its own imitation. The tendency to imitate does not go away. McHugo and colleagues [24] recorded from muscles of the face of subjects viewing a videotape of Ronald Reagan on the presidential campaign trail. Regardless of the viewers' opinion of Reagan, pro or con, when Reagan frowned on film, the corrugator muscle of the forehead, which is associated with frowning, was active. When Reagan smiled, the zygomaticus muscle at the lips was active.

Are listeners likewise disposed to imitate speakers? There is, to my knowledge, no strong evidence on this point. There is, however, evidence of vocal "accommodation"[25] whereby people speaking together may converge in their vocal intensity, speaking rate or frequency of pausing.

Recent research by Michele Sancier and me is in its preliminary stages. However, it has provided a striking outcome that suggests dispositional vocal imitation leading to speech learning in a speaker well past the critical stage of language acquisiton.

Our research was inspired both by the foregoing evidence that humans are disposed to imitate and anecdotal evidence that geographically dislocated adults may show dialectal drift toward the ambient speech of their new language community. A sample anecdote involves a young woman from Tennessee who attended a college in New England. Returning from Tennessee after the Christmas break of her freshman year, she announced to her

New England friends, in what sounded to them as a marked southern accent, that her family and friends back home had told her that she had lost her southern accent. Another example is of a colleague of Michele and mine at Haskins Laboratories who is a native British English speaker. He reported to me that his relatives in English consider him to speak with a "ghastly American accent." However, in reporting this to me, he pronounced the adjective [gastli], using a vowel that sounded British to my ears, rather than the American [ae]. In both of these cases, and many others that colleagues have reported to us, it is apparent that some drift toward the ambient speech of the language community must be occurring well after the end of the ostensible critical period for language acquisition.

When this drift occurs in speakers of a different dialect of a common language, there may be more than one source of the drift. It may occur, as we suppose, because listeners are disposed to imitate the gestures they hear. Or, instead or in addition, it may occur for social reasons. Fitting in vocally may facilitate fitting in socially.

To avoid that second possible source of gestural drift, we have been looking at the speech of a bilingual speaker. She is a native speaker of Brazilian Portuguese who attends the University of Connecticut. She is a fluent, but accented, speaker of English. She spends the academic year in Connecticut, where she speaks English almost exclusively and the summer and occasional Christmas breaks in Brazil where she speaks Portuguese almost exclusively. If this individual shows drift of her gestures in *Portuguese* toward the gestures of her English-speaking language community when she is in Connecticut, the reason is unlikely to be social. It is no social advangate for her to produce American-accented Portuguese. Accordingly, we interpret crosslinguistic gestural drift as evidence of speech learning based on a disposition of listeners to imitate perceived gestures.

The speaker had her own anecdote that led us to focus first (and, so far, only) on the voice onset times (VOTs) of her voiceless consonants. When she goes home to Brazil, her father accuses her (not in these words) of producing excessively aspirated voiceless stops. If her father is correct, then her unaspirated Portuguese voiceless stops are drifting in the direction of the aspirated voiceless stops of English.

To look for evidence of drift, we recorded the speaker on six occasions. First, we recorded her twice, approximately 24 hours apart, after she had been in Connecticut for five months and just before she left for a visit to Brazil. Next, we recorded her in two sessions one within hours of her return to Connecticut after her two month stay in Brazil and a second session one day after that. Finally, we recorded her in two sessions after she had been in Connecticut for four months.

In each session, a speaker of English read 12 sentences to her. After each sentence, our subject produced its Portuguese translation. This procedure was repeated four times so that four translations of each sentence were obtained. Compatibly, a speaker of Portuguese read 12 sentences four times each to our subject, who provided their English translations. Figure 1 provides the data on her Portuguese and English /p/s and /t/s. In the displays, findings are collapsed over each pair of sessions recorded 24 hours apart. (Generally, these sessions did not provide statistically distinguishable measures of VOT.)

Two findings are notable. First, analogous to some findings of other studies of the speech of bilinguals [26], when the speaker produces English voiceless stops, her VOTs are longer than those of her Portuguese stops. (Notice that the vertical scales for the graphs of the speaker's Portuguese and English speech both span 30 ms of VOT; however, the graph of Portuguese speech displays VOTs between 0 and 30 ms in duration, whereas that of English speech displays VOTs between 35 and 65 ms in duration.)

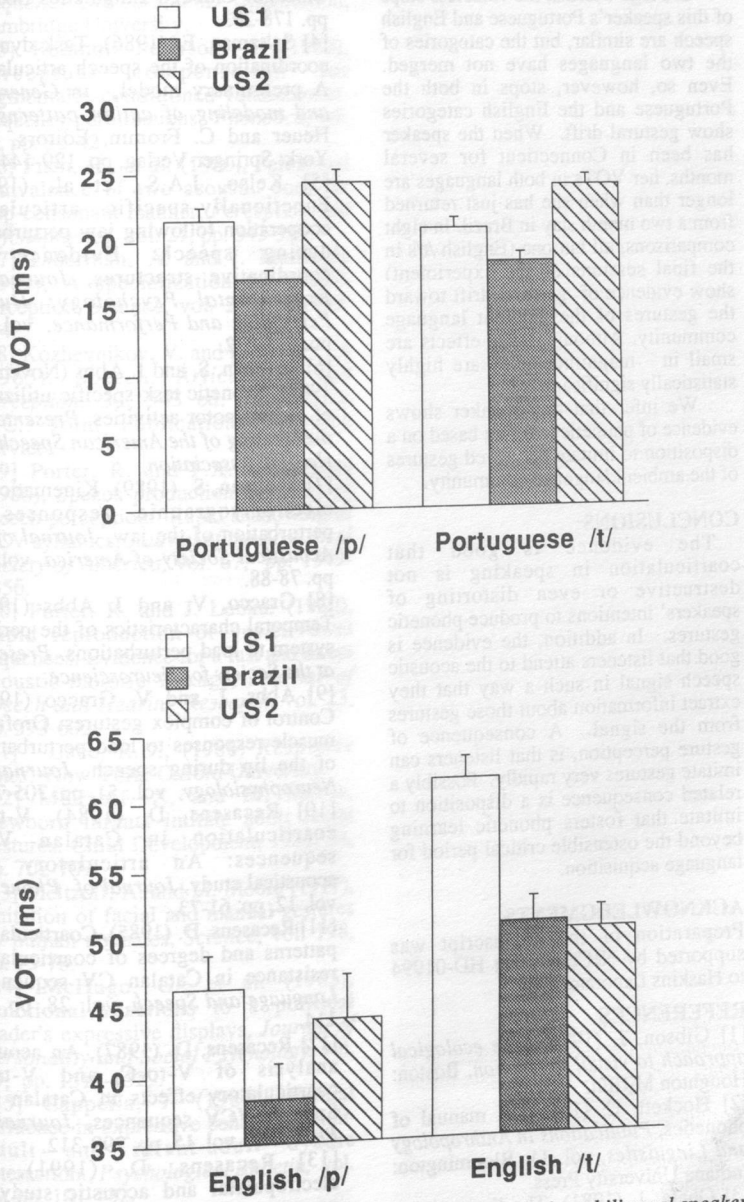

Figure 1: Voice onset times (VOTs) and standard error bars for a bilingual speaker of Portuguese and English measured at three points in time: After a several month stay in the United States (US1), after two months in Brazil, and again after several months in the US.

In Flege's terms, the voiceless stops of this speaker's Portuguese and English speech are similar, but the categories of the two languages have not merged. Even so, however, stops in both the Portuguese and the English categories show gestural drift. When the speaker has been in Connecticut for several months, her VOTs in both languages are longer than when she has just returned from a two month stay in Brazil. In eight comparisons, all but one (English /t/s in the final sessions of the experiment) show evidence of gestural drift toward the gestures of the ambient language community. Although these effects are small in magnitude, they are highly statistically significant.

We infer that this speaker shows evidence of phonetic learning based on a disposition to imitate perceived gestures of the ambient language community.

CONCLUSIONS

The evidence is good that coarticulation in speaking is not destructive or even distorting of speakers' intentions to produce phonetic gestures. In addition, the evidence is good that listeners attend to the acoustic speech signal in such a way that they extract information about those gestures from the signal. A consequence of gesture perception, is that listeners can imitate gestures very rapidly. Possibly a related consequence is a disposition to imitate that fosters phonetic learning beyond the ostensible critical period for language acquisition.

ACKNOWLEDGMENTS

Preparation of the manuscript was supported by NICHD grant HD-01994 to Haskins Laboratories.

REFERENCES

[1] Gibson, J. (1979), *The ecological approach to visual perception.* Boston: Houghton Mifflin.

[2] Hockett, C. (1955), A manual of phonetics, *Publications in Anthropology and Linguistics*, vol. 11, Bloomington: Indiana University Press.

[3] Ohala, J. (1981), The listener as a source of sound change, in *Papers from the parasession on language and behavior*, C. Masek, et al., Editor. Chicago: Chicago Linguistics Society, pp. 178-203.

[4] Saltzman, E. (1986), Task dynamic coordination of the speech articulators: A preliminary model, in *Generation and modeling of action patterns*, H. Heuer and C. Fromm, Editors. New York: Springer-Verlag, pp. 129-144.

[5] Kelso, J.A.S., et al. (1984), Functionally-specific articulatory cooperation following jaw perturbation during speech: Evidence for coordinative structures, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, pp. 812-832.

[6] Shaiman, S. and J. Abbs (November, 1987), Phonetic task-specific utilization of sensorimotor activities. *Presented at the Meeting of the American Speech and Hearing Association.*

[7] Shaiman, S. (1989), Kinematic and electromyographic responses to perturbation of the jaw, *Journal of the Acoustical Society of America*, vol. 86, pp. 78-88.

[8] Gracco, V. and J. Abbs (1982), Temporal characteristics of the perioral system to load perturbations. *Presented at the Society for Neuroscience.*

[9] Abbs, J. and V. Gracco (1984), Control of complex gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology*, vol.. 51, pp. 705-723.

[10] Recasens, D. (1984), V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study, *Journal of Phonetics*, vol. 12, pp. 61-73.

[11] Recasens, D. (1985), Coarticulatory patterns and degrees of coarticulation resistance in Catalan CV sequences, *Language and Speech*, vol. 28, pp. 97-114.

[12] Recasens, D. (1987), An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences, *Journal of Phonetics*, vol. 15, pp. 299-312.

[13] Recasens, D. (1991), An electropalatal and acoustic study of consonant-to-vowel coarticulation, *Journal of Phonetics*, vol. 19, pp. 177-196.

[14] Silverman, K. (1987), *The structure and processing of fundamental frequency contours*, PhD Dissertation, Cambridge University:

[15] Reinholt Peterson, N. (1986), Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations, *Phonetica*, vol. 43, pp. 31-42.

[16] Fitch, H., et al. (1980), Perceptual equivalence of two acoustic cues for stop-consonant manner, *Perception and Psychophysics*, vol. 27, pp. 343-350.

[17] Liberman, A. and I. Mattingly (1989), A specialization for speech perception, *Science*, vol. 243, pp. 489-494.

[18] Kozhevnikov, V. and L. Chistovich (1965), *Speech: Articulation and Perception.* 30, 543 ed. Washington D.C.: Joint Publications Research Service.

[19] Porter, R. and F.X. Castellanos (1980), Speech production measures of speech perception: Rapid shadowing of VCV syllables, *Journal of the Acoustical Society of America*, vol. 67, pp. 1349-1356.

[20] Porter, R. and J. Lubker (1980), Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage, *Journal of Speech and Hearing Research*, vol. 23, pp. 593-602.

[21] Luce, R.D., (1986) *Response times.* New York: Oxford University.

[22] Meltzoff, A. and M. Moore, Newborn infants imitate adult facial gestures. Child Development, 1983. 54, pp. 702-709.

[23] Meltzoff, A. and M. Moore (1977), Imitation of facial and manual gestures by human neonates, *Science*, vol. 198, pp. 75-78.

[24] McHugo, G., et al. (1985), Emotional reactions to a political leader's expressive displays, *Journal of Personality and Social Psychology*, vol. 49, pp. 1513-1529.

[25] Cappella, J. (1981), Mutual influence in expressive behavior: Adult-adult and infant-adult dyadic interaction., *Psychological Bulletin*, vol. 89, pp. 101-132.

[26] Flege, J.E. (1987), The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification, *Journal of Phonetics*, vol. 15, pp. 47-65.