THREE LINES OF EVIDENCE FOR DIRECT LINKS BETWEEN PRODUCTION AND PERCEPTION IN SPEECH

D. H. Whalen
Haskins Laboratories, New Haven, CT, USA

ABSTRACT

How perception is guided by production has led to three main theories: no necessary production-perception relationship (eg., automatic speech recognition); speaker adaptations to auditory requirements for speech; or listeners' perceiving linguistically significant actions ("gestures"). Listeners handle acoustic complexity, but is a link with production typical? One link, **individual differences** (where perception matches production), is compatible with theories predicting separate processes. Second, **perceptual parsing,** shows listeners perceptually undo overlapping articulation (vowel-to-vowel coarticulation; F0 in vowel, consonant and intonational distinctions; and duration in vowels and consonants). Third, **vocal imitation,** appears early in infants, but adult listeners can imitate a changed vowel as quickly as a simple (non-choice) response; a tight link seems to persist beyond the learning stage. These lines of research indicate a close link between production and perception. We have yet to explore all the implications, but it has become clear the links exist.

1. INTRODUCTION

The speech signal is structured by the movements of the various articulators. Although this link is uncontentious in itself, the means of testing it directly have, until recently, been somewhat cumbersome and seldom used. The research areas of production and perception, then, have tended to be rather distinct. This is by no means a fatal separation, since much valuable work has been accomplished under this scheme. But certain questions cannot be asked without joining the two approaches. This paper explores three ways in which some of those questions have been addressed.

The production of speech is often viewed as a problem of movement coordination, comparable to other coordinated motor activities, such as reaching or walking. From that standpoint, it is no more necessary to examine the perception of speech than it would be to examine how a particular gait is seen by external observers. And, in fact, the vast majority of production studies have no measurement of the perceptual effects of the utterances produced, except for a check by the experimenter that the category produced was the one intended. While the speech we perceive is normally produced by another human being, the relationship between production and perception has received little direct experimentation. Indeed, some theorists claim that the two processes are separable. Since the movement of the articulators must result in an acoustic signal in order to be perceived, the theory goes, it is the acoustics that are important, not the articulation.

With respect to speech perception, there are three main theoretical perspectives on the relation of perception to production. One view is that there is no necessary connection, and that perception can proceed with no knowledge of production. While this view may not be widely held as a theoretical position, it is by far the main approach in automatic speech recognition (ASR). In ASR, acoustic templates for words are typically the target for the recognition process. While the advances of ASR have been impressive, its limitations have come from its generating its templates from specific recording conditions; if the conditions change, recognition rates plummet. In order to overcome this limitation, it is quite possible that the next generation of ASR will require reference to the production of speech [35]. The second theoretical stance holds that there are certain auditory requirements for speech and that the speaker adjusts the articulators in order to achieve those effects [4, 15, 38]. In this framework, phonetic perception is held to be a relatively simple auditory function, and thus the job of the production mechanism is to ensure that the proper acoustic goals are reached. From this perspective, the study of speech production should uncover strategies for producing optimally perceivable acoustic signals. The third view holds that listeners perceive linguistically significant actions of the vocal tract ("gestures"), because those actions are signaled by their acoustic (and sometimes visual or haptic) consequences [8, 25]. What makes ASR so difficult, on this view, is that the gestures overlap in time and thus affect the articulators in a complex fashion, resulting in complex acoustic consequences. Also, since the acoustic transfer function underlying any instantaneous part of the signal does not uniquely specify a vocal tract shape, we must have additional evidence before recovering the underlying articulation on the basis of the acoustic signal. That listeners handle this complexity in the acoustic signal with ease is apparent from the success that language enjoys. Examining perception in light of production should bring us to a better understanding of the speech mechanism as a whole.

It was not until the mechanical study of speech was possible that scientists began to dissociate the production of speech and its perception. Prior to Edison's phonograph, there was no way of perceiving speech other than by producing it, so the link was too obvious to mention. Once the signal could be examined on its own, however, it began to be possible to describe speech in purely acoustic terms. The visual features that were posited as crucial for speech based on the examination of spectrograms [33] is one example of such an attempt. Other examples can be found in the feature system of Jakobson, Fant and Halle [12], despite the fact that its distinctive features for language were stated both in terms of articulation and in terms of the acoustics, with the two descriptions being otherwise independent. Nonetheless, some of the features, such as "grave," had convincing unitary descriptions only in the acoustic domain.

Stetson's Motor Phonetics [37] implied that the movement of the articulators should be the object of perception, at least in comparison with static positions and the acoustic slices associated with them. But it was the Motor Theory of speech perception [19-21, 26] that made the most explicit statement of a direct link

between production and perception. Its revised version states that "the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations" [21, p. 2]. There is no mapping between perception and production because the units involved are the same. This link is assumed to be necessary, because, among other reasons, the treatment of a sound as speech or nonspeech seems to proceed on the basis of whether it *could* form a speech sound or not, not on overt indications or stigmata [21, p. 18] .

However, it is not necessary to adopt the Motor Theory in order to emphasize the link between production and perception, and some gestural theories explicitly do not endorse the Motor Theory. In the most developed of these alternatives, Fowler's "direct realism" [5-8, 10] posits a system in which the gestures of speech are specified directly by the speech signal. There is no need to infer what the articulatory patterns were because the specification is present in the speech signal. Speech, in direct realism, is only another of the many events in the world that are specified by the acoustic signals they cause. What is at issue, then, is the aspects of the acoustic signal that are used by the perceiver and the appropriate descriptors for those aspects. Direct realism predicts that the descriptors will be gestural, while acoustic theories predict they will be auditory.

The findings that prompted the original Motor Theory were largely perceptual [20]. The theory rested primarily on the apparent lack of invariance in the acoustic signal and the evidence that the listener's percept corresponded more closely to articulation than to the acoustics. For example, the second formant transitions of /di/ and /du/ go in opposite directions, while the articulation of the stop in the two syllables is the same gesturally, and the consonants of the two syllables sound the same to listeners. This view was further supported by the wide array of trading relations that were found in speech which make sense in relationship to the productive process but are difficult if not impossible to account for on general auditory principles [for a review, see 34]. It is interesting, then, that the original theory was based on very few motor measurements. Since that time, there have been several studies directly relevant to the production-perception link.

2. INDIVIDUAL DIFFERENCES

One of the most appealing ways of demonstrating a link between production and perception is to establish that certain speakers have variant productions that affect their perception in systematic ways. Some studies of vowel perception have shown such effects. A study by Bell-Berti, Raphael, Pisoni and Sawusch [1] found that there were two patterns in tongue muscle activity for the high front vowels of English. This difference further predicted the size of the effect of an anchoring stimulus in the perception of a synthetic continuum of those same vowels. Subjects seemed to pay more attention to the kind of difference that they themselves used in those vowels. A study by Fox [11] found a relationship between subjects' judgments of similarity in English vowels and their own productions of the three point vowels. Fox's experiment is, perhaps, even more surprising than Bell-Berti et al.'s in that the vowels presented were typical tokens from native speakers, not synthesis. Nonetheless, the listeners seemed to react not just to the acoustic signal as presented but to a more idealized space, correlated with their own behavior. Another study by Paliwal, Lindsay and Ainsworth [30] failed to find a correlation between identification and the speaker's own vowels. However, the statistics used (looking for differences in the magnitude of correlations between the produced and perceived vowels) may have been insensitive to some real differences. Overall, then, there are good indications that some vowel categories are perceived relative to a speaker's own habits. However, since none of these studies was concerned with contrasting articulatory and acoustic theories of vowel perception, none points out that the influence might also be at the level of the acoustic output of a speaker's own vowels. Thus these studies, though interesting, do not appear to distinguish the two kinds of theories.

3. PERCEPTUAL PARSING

Gestural overlap is necessary for fluent, intelligible speech. The overlapped "phasing" [3] of one gesture with respect to another gives the speech signal its cohesion. However, it also causes the signal to be highly context-sensitive and to bear a complex relation to the underlying structure of the utterance. Research suggests that uncoarticulated sound sequences, that is, acoustic alphabets, are unintelligible to listeners if presented at the same rate as that at which speech is typically produced [20]. We can also assume that switching instantaneously from one sound source to another (though easy for a computer synthesizer with stored sound samples) would be impossible for the speaker. Still, many acoustic theories of speech perception hold that simple acoustic properties, usually ones common to the mammalian auditory system, are responsible for the phonetic percept at any particular time slice. Therefore, these theories tend to "dispose" of the gestural overlap by assuming that only the most salient aspect of that acoustic time-slice stands out [41].

However, listeners do not to "dispose" of context-sensitive acoustic information: They use anticipatory coarticulatory information in the domain of one phonetic segment as information for the coarticulating source [e.g., 42]. Moreover, there is evidence that listeners "parse" acoustic speech signals along gestural lines [e.g., 9]. For example, the increase in F0 on a high as compared to a low vowel attributable to intrinsic F0 [44] does not make the vowel's pitch higher. Rather, it contributes to the vowel's perceived quality. Accordingly, the F0 during the vowel is "parsed" into a component that contributes to pitch (or an intonational gesture) and a component that, accurately, is ascribed to the (vocal-fold stretching) lingual gesture for the vowel. There is no reason that for an acoustic-signal perceiving system to parse F0 into components. Accordingly, it is not clear how acoustic theories might explain this and other cases of gestural parsing [e.g., 31].

4. IMITATION

Infants imitate speech readily, and what they imitate is transformed to their vocal tracts. That is, they do not produce the formant and F0 patterns in the target utterance, primarily because they do not have the vocal tracts that can do so. But what they do is sensibly related to the patterns presented, if we assume that the acoustic signal has been transformed into an articulatory space in which an /u/ is a labial- and velar-constricted vowel and an /i/ is a palatal-constricted vowel regardless of the speaker. Similarly with intonation, the F0 is not matched; rather, the changes within the infant's range are similar to the changes within the adult's range.

Results from a fast shadowing task also indicate that the perception of the speech signal allows subjects to imitate faster

than they can make a choice in responses. Such a level is compatible with the notion that the production and perception systems are interrelated. Some subjects are able to change from saying one vowel to another just as quickly when they are asked to shadow the stimulus vowel as they are when they are required to produce a single, prespecified vowel whenever the stimulus vowel changes quality [16, 32]. Although these studies show that there is a link between the production and perception of certain vowels, they do not unambiguously show that a link is common for more typical situations. If speech perception can provide both immediate motoric and discrete, symbolic representations, then the symbolic ones may still be computed with respect to the acoustic signal alone. However, the immediacy of the link shown in the shadowing studies is difficult to encompass in a theory that only connects production and perception via abstract linguistic units. Additionally, the stimuli in some of the shadowing studies were synthetic vowels, which presumably did not match all of the speakers' vowel spaces (if any), so some normalization must have taken place for the appropriate vowel to be produced. The immediacy of the response indicates that this representation must have a motoric component.

Another direct link between production and perception can be seen in experiments involving the Verbal Transformation Effect (VTE). This effect, first discovered by Warren [39], involves the repetition of a single word for an extended period of time. Listeners will typically report that, at some point in the stream, the word has changed. Sometimes, a large number of words will be reported, as when the target word "case" gives rise to percepts of "pace," "taste," "paste" and "haste" [40]. In an interesting extension, Lackner [18] asked subjects to repeat each item as it was presented. These subjects reported virtually no transformations. A further study showed that even having subjects mouth the words rather than utter them aloud is sufficient to reduce the VTE [23]. The experimental results indicate that the involvement of the productive system actively affects perception.

5. GENERAL DISCUSSION

Other work has pointed to apparent discrepancies between production and perception. Ohala [27, 28] has detailed several instances in which the ear is fooled about a sound's source. For example, spectral zeroes added by the open state of the glottis during /h/ may explain the development of nasalization in some languages [24]. These effects should be acoustic since the articulation can be presumed not to include any opening of the velopharyngeal port. Yet this dependence on the acoustics does not, as Ohala has sometimes claimed [29], indicate that speech perception is unrelated to articulation. It is precisely because the historical change results in a change in articulation that these occasional perceptual mistakes indicate a production/perception link. If the acoustic results were paramount, we would expect to find these languages introducing spectral zeroes in a variety of ways, including both the presence of a glottal fricative and the use of a velar gesture. Instead, we have a consistent reanalysis of the (acoustic) zero as (productive) nasality. Listeners apparently could not help perceiving a nasal gesture, even when none was present. A similar case is the "intrusive nasal" of the Eastern Algonquian languages [43]. If the acoustics were sufficient for maintenance of the contrast, we would not expect an articulatory change. The articulation does change, strengthening our supposition that production and perception are intricately connected. Such cases of gestural ambiguity in the acoustic signal are rare enough that the examples are rather hard to find, and worthy of note when they do. But saying that there are occasional ambiguities is not the same as saying there is no relationship.

In articulatory accounts of speech perception, the acoustic structure of speech is not ignored. The earliest formulations of the Motor Theory described gestures as needing to have noticeable acoustic consequences [20]. Certainly, the direct realist account asserts that the gestures are recovered from the acoustic signal. In some sense, this means that it must be possible to derive a correct theory of speech perception that uses only the acoustic signal, though the objects recovered would (on the direct realist account) be gestural ones. This congruence makes distinguishing acoustic and articulatory accounts difficult. However, an example of differing predictions is seen in Williams [46], which builds upon Lindblom and Studdert-Kennedy [22]. The coarticulation of vowels with consonants results in "undershoot" or the failure of the vowel to reach the formant values available in isolation. Williams showed that the perceptual adjustment for undershoot is a phonetic one, since there is a psychoacoustic effect in the *opposite* direction for tones which mimic the trajectories of the formants. Thus it is only when interpreted as speech that perceptual adjustment for the undershoot occurs, and it appears for articulatorily coherent reasons. If we were to rely on psychoacoustics alone, we would expect the effect to go in the other direction.

Articulatory accounts assume that the speaker has an appropriate level of understanding of the relationship between articulatory configurations and the acoustic signals they will produce. This understanding is either biologically determined (in the Motor Theory) or provided by the structure of the acoustic signal itself (in the direct realist approach). The acoustic account must assume that the speaker's knowledge of how to produce a particular formant pattern is a form of self-monitoring. It must also assume that the areas of functional equivalence can be determined as well, leading to the alternate production strategies previously found [13]. The need for self-monitoring is particularly clear in a framework such as Jusczyk's [14] WRAPSA model, in which the infant begins without distinguishing speech from nonspeech sounds. Knowing how her own vocal tract will affect the air waves, then, must be a matter of exploration. There are some difficulties with this assumption. First, the most complete form of vocal exploration occurs during babbling, but infants typically favor front vowels over back vowels [2, 45], reducing the range of relationships explored. More importantly, the formants that an infant's vocal tract produces are very far from those of the adult models available and indeed of the infant's own, future formants of adulthood, yet even 12 week old infants can imitate them [17]. By the time that the vocal tract is reasonably adult-like (after puberty for boys, perhaps earlier for girls), there is, presumably, very little in the way of exploration, at least as seen in variability across age [36].

Any successful theory of speech must explain how it is that listeners can relate an articulation to an acoustic output, to account for both parsing and imitation. Thus any theory that attributes the targets to articulation or to acoustics will correctly describe many phenomena. The ease and rapidity of imitation and the completeness of parsing the acoustic signal into articulatory sources are two factors that are difficult to describe in purely acoustic terms. I would argue, then, that the link between production and perception is strong, immediate, and unavoidable in human speech.

ACKNOWLEDGMENTS

This work was supported by NIH grant DC-02717 to D. H. Whalen and Haskins Laboratories.

REFERENCES

[1] Bell-Berti, F., L.J. Raphael, D.B. Pisoni, and J.R. Sawusch. (1979), "Some relationships between speech production and perception." *Phonetica*, vol. 36, pp. 373-383.

[2] Boysson-Bardies, B.d., P. Halle, L. Sagart, and C. Durand. (1989), "A crosslinguistic investigation of vowel formants in babbling." *Journal of Child Language*, vol. 16, pp. 1-17.

[3] Browman, C.P. and L. Goldstein. (1989), "Articulatory gestures as phonological units." *Phonology*, vol. 6, pp. 201-251.

[4] Diehl, R.L. (1991), "The role of phonetics with the study of language." *Phonetica*, vol. 48, pp. 120-134.

[5] Fowler, C.A. (1980), "Coarticulation and theories of extrinsic timing control." *Journal of Phonetics*, vol. 8, pp. 113-133.

[6] Fowler, C.A. (1981), "Production and perception of coarticulation among stressed and unstressed vowels." *Journal of Speech and Hearing Research*, vol. 46, pp. 127-139.

[7] Fowler, C.A. (1984), "Segmentation of coarticulated speech in perception." *Perception and Psychophysics*, vol. 36, pp. 359-368.

[8] Fowler, C.A. (1986), "An event approach to the study of speech perception from a direct-realist perspective." *Journal of Phonetics*, vol. 14, pp. 143-28.

[9] Fowler, C.A. and J.M. Brown. (1997), "Intrinsic f0 differences in spoken and sung vowels and their perception by listeners." *Perception and Psychophysics*, vol. 59, pp. 729-738.

[10] Fowler, C.A., P.E. Rubin, R. Remez, and M. Turvey. (1980). "Implications for speech production of a general theory of action", in *Language production*, B. Butterworth, Editor, pp. 373-420. Academic Press: New York.

[11] Fox, R.A. (1982), "Individual variation in the perception of vowels: Implications for a perception-production link." *Phonetica*, vol. 39, pp. 1-22.

[12] Jakobson, R., G. Fant, and M. Halle. (1951), *Preliminaries to speech analysis*. 1951, Cambridge, MA: MIT Press.

[13] Johnson, K., P. Ladefoged, and M. Lindau. (1993), "Individual differences in vowel production." *Journal of the Acoustical Society of America*, vol. 94, pp. 701-714.

[14] Jusczyk, P.W. (1993), "From general to language-specific capacities: the WRAPSA Model of how speech perception develops." *Journal of Phonetics*, vol. 21, pp. 3-28.

[15] Kluender, K.R. and A.J. Lotto. (1994), "Effects of first formant onset frequency on [-voice] judgments result from auditory processes not specific to humans." *Journal of the Acoustical Society of America*, vol. 95, pp. 1044-1052.

[16] Kozhevnikov, V.A. and L.A. Chistovich. (1966), *Rech, artikulyatsiya, i vospriyatiye, [Speech: Articulation and perception]*. 1966, Washington, DC: Joint Publications Res. Service.

[17] Kuhl, P.K. and A.N. Meltzoff. (1996), "Infant vocalizations in response to speech: Vocal imitation and developmental change." *Journal of the Acoustical Society of America*, vol. 100, pp. 2425-2438.

[18] Lackner, J. (1974), "Speech production: Evidence for corollary discharge stabilization of perceptual mechanisms." *Perceptual and Motor Skills*, vol. 39, pp. 899-902.

[19] Liberman, A.M. (1996), *Speech: A special code*. 1996, Cambridge, MA: MIT Press.

[20] Liberman, A.M., F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy. (1967), "Perception of the speech code." *Psychological Review*, vol. 74, pp. 431-461.

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

[22] Lindblom, B. and M. Studdert-Kennedy. (1967), "On the role of formant-transitions in vowel recognition." *Journal of the Acoustical Society of America*, vol. 42, pp. 830-843.

[23] MacKay, D.G., G. Wulf, C. Yin, and L. Abrams. (1993), "Relations between word perception and productions: New theory and data on the verbal transformation effect." *Journal of Memory and Language*, vol. 32, pp. 624-646.

[24] Matisoff, J. (1975). "Rhinoglottophilia: The mysterious connection between nasality and glottality", in *Nasalfest: Papers from a symposium on nasals and nasalization*, C.A. Ferguson, L.M. Hyman, andJ.J. Ohala, Editor, pp. 265-287. Language Universals Project: Stanford.

[25] Mattingly, I.G. (1991), "In defense of the Motor Theory." *PERILUS*, vol. 14, pp. 167-172.

[26] Mattingly, I.G. and A.M. Liberman. (1969). "The speech code and the physiology of language", in *Information processing in the nervous system*, K.N. Leibovic, Editor, pp. 97-114. Springer-Verlag: Berlin.

[27] Ohala, J.J. (1981). "The listener as a source of sound change", in *Papers from the parasession on language behavior*, M.F. Miller, Editor, pp. 178-203. Chicago Linguistic Assn: Chicago.

[28] Ohala, J.J. (1990), *What's cognitive, what's not in sound change*. 1990, Duisburg: L.A.U.D.

[29] Ohala, J.J. (1994), "Speech perception is hearing sounds, not tongues." *Journal of the Acoustical Society of America*, vol. 94, pp. 2849(A).

[30] Paliwal, K.K., D. Lindsay, and W.A. Ainsworth. (1983), "Correlation between production and perception of English vowels." *Journal of Phonetics*, vol. 11, pp. 77-83.

[31] Pardo, J.S. and C.A. Fowler. (1997), "Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch." *Perception and Psychophysics*, vol. 59, pp. 1141-1152.

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

[33] Potter, R.K., G.A. Kopp, and H.G. Green. (1947), *Visible speech*. 1947, New York: Van Nostrand.

[34] Repp, B.H. (1982), "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception." *Psychological Bulletin*, vol. 92, pp. 81-110.

[35] Rose, R.C., J. Schroeter, M.M. Sondhi, and O. Ghitza. (1994), "Speech production models in automatic speech recognition--Forming a lasting marriage between speech science and speech technology." *Journal of the Acoustical Society of America*, vol. 95, pp. 2848(A).

[36] Stathopoulos, E.T. (1995), "Variability revisited: an acoustic, aerodynamic, and respiratory kinematic comparison of children and adults during speech." *Journal of Phonetics*, vol. 23, pp. 67-80.

[37] Stetson, R.H. (1951), *Motor phonetics: A study of speech movements in action*. 1951, Amsterdam: North Holland.

[38] Stevens, K.N., S.J. Keyser, and H. Kawasaki. (1986). "Toward a phonetic and phonological theory of redundant features", in *Invariance and variability in speech processes*, J. Perkell and D.H. Klatt, Editor, pp. 426-449. Lawrence Erlbaum Associates: Hillsdale, NJ.

[39] Warren, R.M. (1968), "Verbal transformation effect and auditory perceptual mechanisms." *Psychological Bulletin*, vol. 70, pp. 261-270.

[40] Warren, R.M. and M.D. Meyers. (1987), "Effects of listening to repeated syllables: Category boundary shifts versus verbal transformations." *Journal of Phonetics*, vol. 15, pp. 169-181.

[41] Whalen, D.H. (1983). "Integrating and disposing theories of speech perception", in *Abstracts of the Tenth International Congress of Phonetic Sciences*, A. Cohen and M.P.R.v.d. Broecke, Editor, pp. 526. Foris: Dordrecht.

[42] Whalen, D.H. (1984), "Subcategorical phonetic mismatches slow phonetic judgments." *Perception and Psychophysics*, vol. 35, pp. 49-64.

[43] Whalen, D.H. and P.S. Beddor. (1989), "Connections between nasality and vowel duration and height: Elucidation of the Eastern Algonquian intrusive nasal." *Language*, vol. 65, pp. 457-486.

[44] Whalen, D.H. and A.G. Levitt. (1995), "The universality of intrinsic F0 of vowels." *Journal of Phonetics*, vol. 23, pp. 349-366.

[45] Whalen, D.H., A.G. Levitt, P.-L. Hsiao, and I. Smorodinsky. (1995), "Intrinsic F0 of vowels in the babbling of 6-, 9-, and 12-month-old French- and English-learning infants." *Journal of the Acoustical Society of America*, vol. 97, pp. 2533-2539.

[46] Williams, D.R. (1986), *Role of dynamic information in the perception of coarticulated vowels*. Unpublished Ph.D. thesis, University of Connecticut, Storrs.