Journal of Phonetics (1986) **14**, 75–82

Against the direct realist view of speech perception

John J. Ohala

Department of Linguistics, University of California, Berkeley, California 94720, U.S.A.

1. Introduction

In this paper I will offer comments on two of the claims Fowler makes regarding speech perception in the featured paper of this volume:

(1) that the "distal event" which is perceived by the listener is "the articulating vocal tract"; and
(2) that this event is (or it can be or is typically) recovered from the acoustic signal in a way that involves no mediation by any cognitive processes, including inferences, hypothesis testing, and the like.

It should be noted at the outset that arguments can only be given for the *plausibility* or lack of it of various claims on these issues since, as far as I can tell, no crucial experiment has been proposed, let alone conducted, which would differentiate Fowler's position from others (although more on this later).

Also, although my remarks on the direct realist position will be critical, there is much about it that I agree with and find admirable: the criticism of the naïve mentalist view of speech perception, the desire to squeeze as much useful information out of the speech signal as possible, and the boldness and explicitness of the theoretical position argued for.

2. What is the distal event in speech perception?

In his analysis of visual perception, Gibson (1966) identified the object which reflects light as the distal event which is perceived. In drawing an analogy between visual perception and speech perception, Fowler identifies "the moving vocal tract" as the distal event which is perceived in the case of listening to speech. But even if one accepts the Gibsonian view as it applies to visual perception, it is not so simple to extend it to speech. In vision, the identification of the distal object—the thing behind the immediate sensory stimulus—is fairly obvious. The simplicity of the case of visual perception is due to there being nothing "behind" the object for the perceiver to have to know about. In speech it is not clear that an *a priori* analysis will so easily reveal what the distal event is. Certainly the moving vocal tract is *one* of the events which lies upstream of the acoustic signal, but why focus exclusively on that level? What about the muscular contractions which lead to the vocal-tract movements, the neuronal excitations which lead to the muscular contractions, or even the mental events in the brain of the speaker

0095–4470/86/010075 + 08 $03.00/0

which precede these events—the latter, in some sense, being the ultimate cause of what the speaker pronounces?

Furthermore, there is a sense in which social conventions—which include what we call the "phonology" of the language—dictate what the ultimate target or object of speech perception is. The word-initial glottal stop in the Cairo Arabic word [ʔax] "brother" functions differently from that in the English word [ʔajs] "ice". Recognizing such a difference is surely included in what we commonly mean by "speech perception".

An analogy from the domain of vision may clarify this. Imagine looking at one of the more abstract of Henry Moore's sculptures entitled "reclining figure". Assuming, for the sake of argument, that the perception of the bronze or marble mass itself were done as Gibson would suggest, that would still not enable us to see the "human figure" that is represented by it. The viewer must rely on a great deal of social and artistic knowledge to figure out what the artist intended to express by making the object. Speech perception is similar: the listener must perceive not simply the physical object that is produced when a speaker makes noises with his vocal tract; he must also figure out the linguistic function of those noises, i.e. as with the Moore sculpture, what the object "stands for". As has been repeatedly argued by structural linguistics, part of the communicative function of speech sounds resides in their being different from all other speech sounds. Although this point may have been overemphasized at times (it cannot be seriously maintained, for example, that the ciphers in the speech code have *no* physical reality; see Ohala, 1979), there is, nevertheless, much truth in it. Speakers implement these differences between speech sounds in different ways but, nevertheless, they function similarly. For these reasons, although the perception of the physical events in speech—whether acoustic or articulatory—is a necessary first step in speech perception, this by no means completes the process; the linguistic or communicative function of these events must also be established and this, it seems clear, can only be done by inference, hypothesis testing, and the like.

I grant that it is necessary in general for listeners to attempt to figure out how to produce the speech sounds they hear in others' speech. This is because listeners are also speakers and can only learn what sort of pronunciation is expected of them from listening to others (Ohala, 1981). That is, they must reconstruct the approximate form of the articulation because they have to function as speakers, not because they are listeners. Even so, there seems to be evidence that it is more the *sound* of words that must be accurately reproduced, not necessarily the vocal-tract movements.

For example, Ladefoged, DeClerk, Lindau & Papçun (1972) have shown that different speakers use somewhat different articulations to produce the "same" vowel. Two different vocal tract configurations for the American English [ɔ˞] have been documented (Delattre, 1971; Uldall, 1958) although both, so far as we know, sound virtually alike. Part of the ventriloquist's art, as is well known, is the ability to exploit the many-to-one relationship between articulation and sound. For that matter, the same could be said of talking birds (Greenwalt, 1968; Klatt & Stefanski, 1974). In both cases we perceive "speech" produced with quite atypical vocal-tract movements. Certain sound changes can be explained by noting the acoustic–auditory similarity of many different articulations (Ohala, 1974, 1979; Ohala and Lorentz, 1977; Sweet, 1874). Atal, Chang, Mathews & Tukey (1976) have provided a theoretical basis for this phenomenon. Riordan (1978) reviews a considerable amount of literature as well as her own study on what may be called "compensatory articulations" (where some degree of acoustic constancy in speech sounds can be had by trade-offs in the position of different articulators). A good portion

of this literature reports case histories of speakers with some speech pathology spontaneously discovering alternative articulations which enable them to more closely approximate normal speech (see also Ohala, 1980).

In view of this, there is still much of value to be gleaned from Jakobson's dictum "We speak in order to be heard in order to be understood." The ultimate goal is to convey a message and we do this by making sounds, and we make sounds by moving our articulators. To the extent that we can satisfy a more ultimate goal in this task, precise or consistent fulfillment of the more proximate goals is less necessary. If this is accepted then the "event" which we perceive in speech is the acoustic signal, or perhaps, more accurately, the acoustic signal the speaker intends to produce.

An additional problem with applying the Gibsonian analysis to speech is this: what if the speaker—unlike the reflected object—chooses to emit a vague signal? That is, what if, as Lindblom (1983) has suggested, the speaker, who is aware of the listener's ability to reconstruct the intended pronunciation given only a few choice clues, decides to put only as much energy into the task of speaking as is necessary?

3. Listeners do make errors

Fowler characterizes the "direct realist" view of perception as one in which recovery of the distal event is necessarily "unmediated by cognitive processes of inference or hypothesis testing, which introduce the possibility of error" (p. 4). The fact is, errors *are* made. Sound change, the change in pronunciation of words over the centuries, reveals this: speech perception studies in the laboratory, under quite ordinary listening conditions, also show this and, moreover, yield sound substitutions which are quite similar to those found in sound change. (For this reason it is clear that sound change is in part caused by perceptual misapprehensions.) We can draw two conclusions from this.

3.1. *The speech signal is inherently ambiguous*

First, it must not be the case, as Fowler would claim, that the acoustic signal is such a faithful conveyer of the vocal-tract movements which lie behind the acoustic signal. It is true that, considering the vast amount of spoken exchanges between speakers and hearers, only a very small fraction actually give rise to what linguists study as sound change. However, it seems clear (based on laboratory studies) that when context or other sources of information on the "target" pronunciation are not available, listeners mishear a considerably greater percentage of heard speech. For some sounds or sequences of sounds, this percentage may be on the order of 33% or more (Winitz, Scheib & Reeds, 1972).

Fowler might reply that errors occur when the listener, like the bat mentioned in the study she cites, decides to ignore the stimulus array and instead operates on "automatic pilot". In other words, the fault occurs because the listener does not really perceive speech which, if it had been fully examined, would have provided sufficient information for the error-free apprehension of speech. There are three counter-arguments to this. First, if it turns out that listeners can, and normally do, operate on "automatic pilot" —and Fowler herself reviews some of the extensive evidence for this—what value is there in restricting the term "speech perception" only to the minority of cases, if any, where this is not done? Secondly, it is evident that errors in speech perception are not simply the fault of the listener due to, for example, "limited resolving power" in his auditory

system or not paying full attention to the speech signal. Rather, the speech signal is sometimes *inherently ambiguous*. Winitz *et al.* (1972) found that for certain CV sequences, e.g. [p^hi], presented with other such sequences to listeners who had to identify the consonant, the error rate was greater when 100 ms following the burst was presented (thus including burst plus aspiration plus some fraction of the voiced formant transitions) than when only the burst was presented. Paradoxically then, for a few specific CV sequences, the greater the amount of information presented, the greater was the error. The reason for this is not hard to identify: as some recent research has indicated, the burst is often a much more reliable cue to stop place of articulation than are the transitions (Blumstein & Stevens, 1979). The transitions for [p^hi] are very similar to those for [t^hi] (and this was the most common error for [p^hi]) (Ohala, 1978, 1980, 1983*a*, *b*). Other similar cases abound. Thirdly, the character of many sound changes is such that they do not yield to an analysis either in terms of limited resolution of the auditory apparatus, failure to factor out two or more phonemes' different coarticulated gestures, or failure to attend fully to the acoustic speech signal. For example, it seems clear that sound changes of the sort which transformed "actual" from [ækt + juəl] to [ækʧuəl] occurred because listeners misconstrued as purposeful, i.e. linguistic, the noise which was produced accidentally when the stop was released before a palatal glide (which has such a narrow passage for the escaping air to flow through that audible turbulence is created). In other words, listeners have to form a hypothesis about the function of this noise, and sometimes they accept the wrong hypothesis. Many other examples of sound change also lend themsleves to such an interpretation of faulty hypothesizing on the part of the listener, e.g. the development of "epenthetic" glides intervocalically (e.g. the French dialectal variants [lui] ~ [luvi] "Louis", [tye] ~ [tywa] ~ [tyba] "to kill" (Grammont, 1933, p. 234)).

I agree with Fowler that listeners are generally capable of factoring out the influence of one segment upon another (Ohala, 1981) and have presented evidence for this based on certain types of sound change, namely *dissimilation* (Ohala, 1981, 1983*a*). These sound changes reveal that, occasionally, this factoring out is done inappropriately and thus constitutes a perceptual error. Fowler interprets these as "failures to segment the signal precisely along coarticulatory lines". She therefore lays the blame for this error in perception on the listener, whereas in the account I give for this phenomenon I emphasize that the speech signal is also to blame because it contains inherent ambiguities which would lead to this type of error. The evidence, I believe, is on my side: if listeners were to blame entirely, then such "segmentation failures" (or whatever one wants to call them) ought to be fairly random, i.e. evenly spread out over all types of sound sequences. The fact is, they are not random; there is a great deal of "structure" to these errors. Among other things, they involve quite specific types of features, those which are spread out in time, such as aspiration, labialization, and the like, and not features whose primary cues are realized in a relatively short time window, e.g. stop, affricate. Moreover, by definition, they occur when two sounds sharing these features occur in the same word (Ohala, 1981, 1983*a*, and in press). We must conclude that all cues in the speech signal are not of equal clarity or salience and some of them are notoriously ambiguous.

Dissimilation is only one type of sound change of course, but it is generally the case, as I have argued in several publications, that sound changes—and thus the misapprehensions that give rise to them—are far from random as to the kind of speech sound sequences they involve, both in the case of the original segments and the segments they change into.

3.2. *Listeners must be aware of the speech signal's unreliability*

The second conclusion we can draw from the fact that speech perception is subject to errors is that the listener must be aware of this fact too. The natural strategy—the only strategy, in fact—to adopt when one recognizes that one may be dealing with unreliable data (as a source of information about something) is to make a *guess* about what the data imply ("abduction" or "hypothesis", to use polysyllabic equivalents). This guess must be tested, i.e. one evaluates it by seeing what its consequences are. Maximally efficient guessing is not random; it is based on prior experience.

There is a considerable literature which shows that speech perception can be greatly aided by supplying information which aids hypothesis formation. First, there is the phoneme restoration effect (Warren, 1970), where speech sounds which have been spliced out of a recorded utterance are "filled in" by the listener—presumably by the application of higher-order knowledge of the redundancies in the message. Secondly, there is the effect whereby connected speech imbedded in noise suddenly becomes "clear" when a single keyword describing its contents is provided to the listener (Bruce, 1956), etc. (see also Miller & Nicely, 1955, p. 352). Fowler's (1983) suggestion that linguistic competence serves "to direct the listener's attention to the segments, which are fully specified in the signal" (p. 309), would not apply to these cases because in one, namely phoneme restoration, the segments which are "heard" in the signal are not physically present and in the second, because the key word does not change the "competence" of the listener —as this term is usually used.

All of this is quite incompatible with a "direct realist" view of speech perception.

4. A proposal for a crucial experiment

It might be objected that phoneme restoration and the like involve mechanisms which are of a higher order than those directly used in speech perception *per se*, i.e. phonotactic, syntactic, and semantic redundancies. I therefore suggest an acoustic phonetic analogue to the phoneme restoration effect.

Assume, first, that certain pairs of speech sounds, symbolized as E and F, are structurally similar to one another except that one of them has a certain distinguishing feature (the "foot" of the "E") that the other lacks. We then present these and other such pairs to listeners under conditions where the stimuli are systematically distorted in a way which progressively and incrementally obscures the distinguishing feature. The conditions of presentation should be such that the listener has ample opportunity to notice the distortion. The rate of confusions should increase as the degree of distortion increases, of course. I would predict that under these circumstances there would also be a *disproportionate increase* in the type of error where F was confused with E, i.e. where the distinguishing feature missing from the F stimuli is induced by listeners. This result would not be predicted by a direct realist account of speech perception because it would imply that listeners guessed about details of the speech signal which could not be directly detected.

Although this experiment has yet to be run in the form just described, I believe there are experiments reported in the literature which may provide some of the data required. In Table I, I summarize some relevant data from two studies: the ratio of confusions of high back rounded vowels for high front unrounded vowels (for example, the confusion of [u] with [i]) to the total confusions for the same vowels. The first study, which

TABLE I. Percent confusion of back rounded vowels with front unrounded having similar height

Study	Vowels involved	Condition	Percent confusion
Peterson & Barney (1952)	u > i U > I ɔ > ɛ	Ordinary listening conditions	0.0
	i > u I > U ɛ > ɔ	Ordinary listening conditions	0.0
Pickett (1957, Table 1)	u > i U > I	Low-frequency noise	37
	o > e ɔ > ɛ	Flat noise	65
		High-frequency noise ($S/N = -30$ dB)	76
		High-frequency noise ($S/N = -40$ dB)	60
	i > u I > U	Low-frequency noise	22
	e > o ɛ > ɔ	Flat noise	42
		High-frequency noise ($S/N = -30$ dB)	66
		High-frequency noise ($S/N = -40$ dB)	37

represents the condition of no distortion of the stimuli, presents data from Peterson & Barney (1952). (Many similar studies could be used, e.g. Stumpf, 1926, Ch. 3.) The second study (Pickett, 1957) reports the case where noise obscured parts of the spectrum: low-frequency noise which would obscure primarily the first formant, flat noise which, given the typical roll-off in the high frequencies of speech spectra, would obscure high second formants somewhat more than the first formant, and high-frequency noise which would obscure primarily high second formants. Of course, the two studies are not strictly comparable since they involved different listening conditions and different stimulus vowels: however, the point is that under normal listening conditions where, presumably, the listener does not assume any systematic distortion of the stimuli, there is practically no confusion of back vowels with front vowels (or vice-versa). However, when the vowels are heard under noise, especially noise that might obscure the high second formant, listeners give evidence that they "fill in" the missing higher formant and do this more often than they fail to detect the higher formant when it is present in the heard stimulus (compare rows 3–6 with rows 7–10 in Table I).

Results such as these need to be examined carefully. On the surface, however, they are not at all compatible with the direct realist view of speech perception. The hypothesizing listeners apparently engage in during these experiments is done at such an early stage in the process of decoding speech that those defending the direct realist position could not dismiss this as lying outside of speech perception proper—at least, not without arbitrarily redefining the meaning of the phrase "speech perception".

5. Conclusion

My argument against the direct realist view of speech perception is that: (1) it is not clear that the moving vocal tract constitutes the sole object of speech perception; and (2) listeners make perceptual errors which appear to be the result of incorrect hypothesizing. I propose a crucial experiment to reveal whether listeners guess at the character of portions of the speech spectra which have been obliterated. Analysis of previously reported studies suggest that they do.

There is one more plausibility argument that can be offered. Obviously, science itself is not conducted by direct perception of the causes underlying the phenomena we wish to understand. If it were, we would not be having this dispute about mechanisms of perception. We proceed by making hypotheses and evaluating them by tests. It seems unlikely and unnecessary that there should be two ways of acquiring knowledge—one way for speech and another in science in general.

References

Atal, B., Chang, J. J., Mathews, M. V. & Tukey, J. W. (1976). Articulatory compensation: A study of ambiguities in the acoustic-articulatory mapping, *Journal of the Acoustical Society of America,* **60**, S77.

Blumstein, S. & Stevens, K. N. (1979). Acoustic invariance in speech production: evidence from measurement of the spectral characteristics of stop consonants, *Journal of the Acoustical Society of America,* **66**, 1001–1017.

Bruce, D. (1956). The effect of context upon the intelligibility of heard speech. In: C. Cherry (ed.), *Information theory*, pp. 245–252. London: Butterworth.

Delattre, P. (1971). Pharyngeal features in the consonants of Arabic, German, Spanish, French, and American English, *Phonetica,* **23**, 129–155.

Fowler, C. A. (1983). Realism and unrealism: a reply, *Journal of Phonetics,* **11**, 303–322.

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

Grammont, M. (1933). *Traité de phonétique.* Paris: Librairie Delagrave.

Greenwalt, C. H. (1968). *Bird song: Acoustics and physiology.* Washington, DC: Smithsonian Institute Press.

Klatt, D. H. & Stefanski, R. A. (1974). How does a mynah bird imitate human speech? *Journal of the Acoustical Society of America,* **55**, 822–832.

Ladefoged, P., DeClerk, J., Lindau, M. & Papçun, G. (1972). An auditory-motor theory of speech production, *Working Papers in Phonetics (UCLA),* **22**, 48–75.

Lindblom, B. (1983). Economy of speech gestures. In: P. F. MacNeilage (ed.), *The production of speech*, pp. 217–245. New York: Springer-Verlag.

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants, *Journal of Acoustical Society of America,* **27**, 338–352.

Ohala, J. J. (1974). Experimental historical phonology. In: J. M. Anderson & C. Jones (eds.), *Historical linguistics II: Theory and description in phonology*, pp. 353–389. Amsterdam: North Holland.

Ohala, J. J. (1978). Southern Bantu vs. the world: the case of palatalization of labials, *Berkeley Linguistics Society, Proceedings,* **4**, 370–386.

Ohala, J. J. (1979). Universals of labial velars and de Saussure's chess analogy. *Proceedings of the Ninth International Congress of Phonetic Sciences.* Vol. 2, pp. 41–47. Copenhagen: Institute of Phonetics.

Ohala, J. J. (1980). The application of phonological universals in speech pathology. In: N. J. Lass (ed.), *Speech and language: advances in basic research and practice.* Vol. 3, pp. 75–97. New York: Academic Press.

Ohala, J. J. (1981). The listener as a source of sound change. In: C. S. Masek, R. A. Hendrick & M. F. Miller (eds.), *Papers from the parasession on language and behavior*, pp. 178–203. Chicago: Chicago Linguistic Society.

Ohala, J. J. (1983*a*). The direction of sound change. In: M. P. R. van den Broecke & A. Cohen (eds.), *Abstracts of the Tenth International Congress of Phonetics Sciences,* pp. 253–258. Dordrecht: Foris.

Ohala, J. J. (1983*b*). The phonological end justifies any means. In: S. Hattori & K. Inoue (eds.), *Proceedings of the XIIIth International Congress of Linguist, August 29–September 4, 1982, Tokyo,* pp. 232–243. Tokyo: ICL Editorial Committee.

Ohala, J. J. (in press). Phonological evidence for top-down processing in speech perception. In: J. S. Perkell *et al.* (eds.), *Invariance and variability of speech processes.* Hillsdale, NJ: Lawrence Erlbaum.

Ohala, J. J. & Kawasaki, H. (1984). Prosodic phonology and phonetics, *Phonology Yearbook,* **1**, 113–127.

Ohala, J. J. & Lorentz, J. (1977). The story of [w]: an exercise in the phonetic explanation for sound patterns, *Berkeley Linguistics Society, Proceedings,* **3**, 577–599.

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels, *Journal of the Acoustical Society of America,* **24**, 175–184.
Pickett, J. M. (1957). Perception of vowels heard in noises of various spectra, *Journal of Acoustical Society of America,* **29**, 613–620.
Riordan, C. J. (1978). Acoustic aspects of speech production, Doctoral dissertation, University of Essex.
Stumpf, C. (1926). *Die Sprachlaute.* Berlin: Julius Springer.
Sweet, H. (1874). *History of English sounds.* London: Trübner.
Uldall, E. T. (1958). American 'molar' *r* and 'flapped' *r*. *Revista do Laboratorio de Fonetica Experimental, Coimbra,* **4**, 103–106.
Warren, R. (1970). Perceptual restoration of missing speech sounds, *Science,* **167**, 392–393.
Winitz, H., Scheib, M. E. & Reeds, J. A. (1972). Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conservational speech, *Journal of the Acoustical Society of America,* **51**, 1309–1317.