1091

CLS 33

PAPERS FROM

THE PANELS

ON

Linguistic Ideologies in Contact
Universal Grammar, Parameters and Typology
The Perception of Speech and other Acoustic Signals

APRIL 17-19, 1997

Edited by:

Kora Singer
Randall Eggert
Gregory Anderson

The University of Chicago

CHICAGO LINGUISTIC SOCIETY
CHICAGO, ILLINOIS

What Duplex Perception Tells Us about Speech Perception

D. H. Whalen
Haskins Laboratories

Abstract: The conventional view of speech perception holds that the sounds of speech are treated first as ordinary auditory objects and then matched to speech templates. An alternative view is one which states explicitly that speech perception is performed by a specialized system, often called a module, which treats speech sounds in the context of the vocal tracts that produce them. This view places humans squarely in the line of evolution, since it appears that every higher species has a neurological specialization for communication with its own species. The present paper reviews direct evidence that speech perception is specialized, seen in the phenomenon of duplex perception, in which the speech signal is artificially modified to such an extent that part of it is heard both as speech and as nonspeech simultaneously. Most commonly, this is done by placing a small part of the signal on one ear while the other ear receives the remainder. The speech process is strong enough to overcome this ecologically impossible situation, and it integrates the information from the two sources. The existence of duplex perception, then, immediately shows that speech perception is not dependent on auditory scene analysis, which assigns sounds to sources and sources to locations. Further exploration shows that the discriminability of the two percepts differs radically, indicating that two separable processes are independently at work.

1. Introduction

One of the enduring questions in speech research is whether there is a specialized process devoted to the perception of speech signals or, instead, the perception of speech depends on ordinary auditory processes. There are two main views in the literature. The most common is that speech perception is the phonetic labeling of auditory events are that are otherwise entirely typical in their behavior (Stevens & Blumstein, 1978; Diehl & Kluender, 1989; Schouten & Hessen, 1993). The alternative view is that speech is perceived not as a collection of auditory events but is immediately represented as the events that created those sounds. There are two main variants of this view. In one variant, the modularity view, this process is specific to speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Whalen & Liberman, 1987; Liberman & Mattingly, 1989), while in the "direct realism" variant, all acoustic input (from natural sources) is assumed to be interpreted according to the structures that produced it (Best, 1984; Fowler, 1986; Fowler & Smith, 1986; Best, 1995). In the present paper, I will call the first class of theories "acoustic" and the second "modular," with the understanding that most of the arguments raised with a modular approach can be made in direct realist terms as well, given certain changes.

One of the main difficulties in distinguishing acoustic and modular views is the fact that, at some level, even a specialized speech processor's input will be indistinguishable from that of ordinary auditory events. That is, the speech module does not reach outside the head and transduce the acoustic wave with something other than the ear. The divergence must occur somewhere along the auditory pathway, and the exact location is something that will require a great deal more definition of the problem, and better neuroimaging tools, than we currently have. It is enough for the present discussion that there be the possibility that there is a

CLS 33: The Panels (1997): 435-446.

difference between processing an acoustic signal then passing the result on to a speech label and processing the speech directly.

There is a great deal of evidence that speech and other auditory signals are processed in different ways, and this evidence has been taken by many researchers to indicate that two different processes are involved. I do not have the space to review this literature, but two examples will form the basis of the main part of the paper, one an early example and one quite recent. Before getting on to those cases, though, it is worth taking a look at why acoustic theories are so appealing in spite of all of this evidence.

Acoustic theories seem to make sense. If you have an acoustic signal, then you would think that you should treat it as an acoustic object and then label the results. This seems as unexceptionable as saying that in vision, we take the optic image and categorize it into optic objects that we then label. But vision is composed of many separate components, including color perception, edge detection and stereoscopic depth perception. All of these processes work off the same sensory input. All are visual in some sense, yet they work independently. Depth perception provides an interesting preview of the duplex perception phenomenon that this paper will focus on. If the images received at the two eyes are experimentally manipulated in the right way, they can elicit a perception of increased stereoscopic depth, as if the two eyes were further apart than they really are. If the disparity is too great, separate images are perceived, resulting in what is normally called double vision. In between these two extremes, however, there is a range where both percepts are available at the same time. That is, increased depth is apparent even though two images are also perceived (Richards, 1971). This is an indication that two separate systems, the one that makes the images at the two eyes cohere and the one that takes retinal differences as indications of depth, are coming to different conclusions about the signal, even though those conclusions are mutually contradictory. That is, it makes no sense for two objects to be perceived at the depth that one object would be at, but this is exactly what happens.

If speech were perceived first as an acoustic object and then classified as speech, the methods of psychophysics should give us the lower limits on what could constitute a speech distinction. For example, subjects are unable to resolve the onset difference between two tones when that difference is less that about 20 ms (Summerfield, 1982), which is quite similar to the VOT boundaries for voicing in many languages (Lisker & Abramson, 1964). But our hearing is by no means limited to those differences we can report. A telling example is the localization of sounds in the horizontal plane based on interaural time differences. These differences reflect the delay involved for the sound in reaching one ear versus the other. These differences are on the order of tenths of a millisecond (Moore, 1982). If our hearing were truly limited to 20 ms differences, then auditory localization would be impossible: This value is too large, by two orders of magnitude, to account for the use in interaural time delays. Instead, localization is an example of another specialization that takes the auditory signal as its input. Basically, we should be no more surprised that there are many specializations working in audition than we are to find many at work in vision.

2. Duplex perception

The phenomenon that I will use to exemplify the radically different treatment of the acoustic signal by the speech and nonspeech systems is duplex perception. To create this phenomenon, we make a small portion of the speech signal (in our case, a third formant transition) stand out from the remainder of the speech by changing

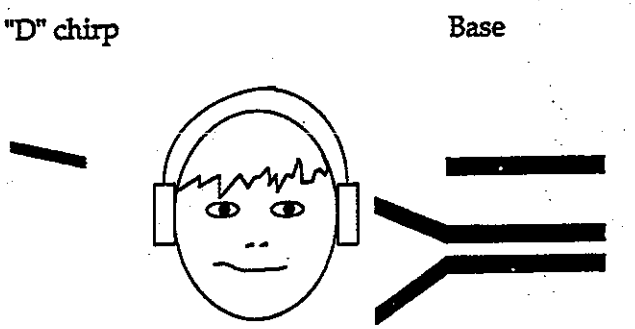

Figure 1: A typical stimulus for a dichotic duplex paradigm.

its location or its composition. In the original version (Rand, 1974), the transition is put on one ear and the remainder of the syllable, or "base," is put on the other (see Figure 1). This elicits a look of wry bemusement on the part of the listener, but results in a speech percept localized to the ear with the base but with the stop information being processed as if the transition had been presented to that ear as well.

The important characteristics of these parts of the signal are these: First, the base by itself sounds like an ambiguous syllable. Although it is as speech-like as any other stylized synthetic syllable, its initial consonant wavers between being perceived as a "d" or a "g." Second, the transition can be made to disambiguate the syllable, in the present case, by specifying "d" or "g." Third, by itself, neither transition sounds like "d" or "g" or, indeed, like speech at all. The two are quite distinct auditorily, but it is not possible to label them correctly as "d" or "g." Fourth, when presented together, the perception is duplex, in that the nonspeech "chirp," as we may call it, is heard in one ear while the disambiguated syllable is heard in the other ear. It is critical that the perception is duplex and not triplex, that is, the ambiguous syllable is not heard separately from the unambiguous syllable. There is one speech percept, and it is the syllable as determined by the transition heard in the other ear. Finally, when we look at qualitative differences between the perception of the transition as chirp and the transition as part of the speech syllable, it is important to remember that these signals are both reaching the ears in exactly the same way for both conditions. Whether the listener is reporting the speech side or the nonspeech side, the auditory stimulation is not merely similar, it is identical in the two cases.

Note that duplex perception is a situation where the speech module is ignoring scene analysis. Scene analysis is the allocation of auditory objects to sources and sources to locations (Bregman, 1990). The effectiveness of the transition can be reduced (though not to insignificance) by making it part of a stream with more coherence, for example, by repeating the transition in a rhythmic pattern (Ciocca & Bregman, 1989; Ciocca, Bregman, & Capreol, 1992). But this reduction of phonetic effectiveness obscures the more important fact that scene analysis *was* successful, in the first case, in assigning the transition to a separate stream. The speech module, however, ignored that information and took the speech information from wherever it could. This behavior comes about, I believe,

because there is no overt indication of what constitutes a speech signal. Speech is just that which can be attributed to a human vocal tract, and the only way of finding out whether a signal could have been such a signal is to analyze it and see if a successful analysis can be made. If so, then speech is perceived. If not, nonspeech is perceived. In some cases, that is, in duplex perception situations, both succeed at once.

The duplex effect can be obtained with the base on either ear. Given the advantage that speech has when it is presented on the right ear and thus initially processed by the left hemisphere (Shankweiler & Studdert-Kennedy, 1967; Darwin, Howell, & Brady, 1978), there may be some differences in the success rate of duplex perception. However, this issue has not been formally explored.

There are many experiments involving duplex perception, but I am going to focus on just two, one which shows the qualitative difference between the speech and the nonspeech perceptual process, and another which shows that the availability of speech or nonspeech information is radically different.

3. Qualitative differences between speech and nonspeech percepts

The first example of duplex perception comes from Mann and Liberman (1983). Their stimuli consisted of a base which was ambiguous between "da" and "ga" and a continuum of transitions, nine members in all, ranging from a value that unambiguously signaled "da" to one that unambiguously signaled "ga." In between, many of the transitions, when combined with the base, were still ambiguous between the two stops. None of the transitions, by themselves, sounded like speech. The schematic spectrograms can be seen in Figure 2.

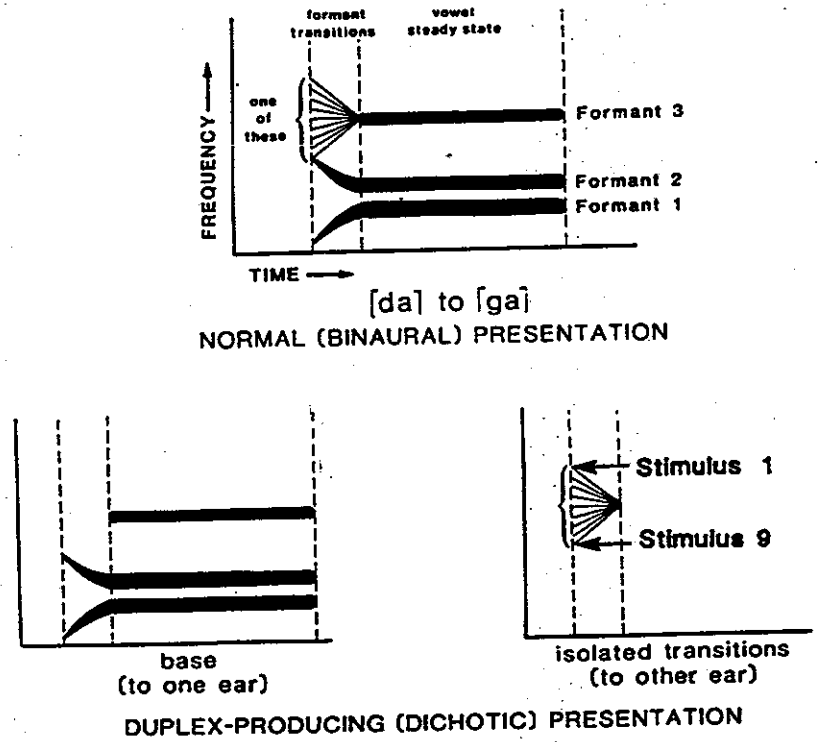

Figure 2: Schematic spectrograms of the stimuli used in Mann and Liberman (1983). © Elsevier Sequoia, reprinted with permission.

Mann and Liberman presented these stimuli dichotically, with the transition on the right ear and the base on the left. They used an AXB discrimination paradigm, in which the first and third members of a triad were always different in transition (by three steps, in this case), while the middle stimulus always matched either the first or the third. Such a forced choice discrimination task allows us to compare the perceptual performance on stimuli even if there are no category labels, as with the nonspeech chirps. Subjects were run under two sets of instructions. In one, they were to attend to the ear receiving the speech, while in the other, they were to attend to the nonspeech ear. They were not informed that the stimuli in the two cases were identical.

The results can be seen in Figure 3. For the speech, we see a typical categorical perception result, in which the endpoints of the stimulus range are discriminated at chance while the middle elicits quite high discrimination. The identification results do, in fact, predict this pattern quite well, a necessary test that this is categorical discrimination. For the nonspeech discrimination, we see a totally different pattern of results. Stimuli that elicited chance discrimination in the

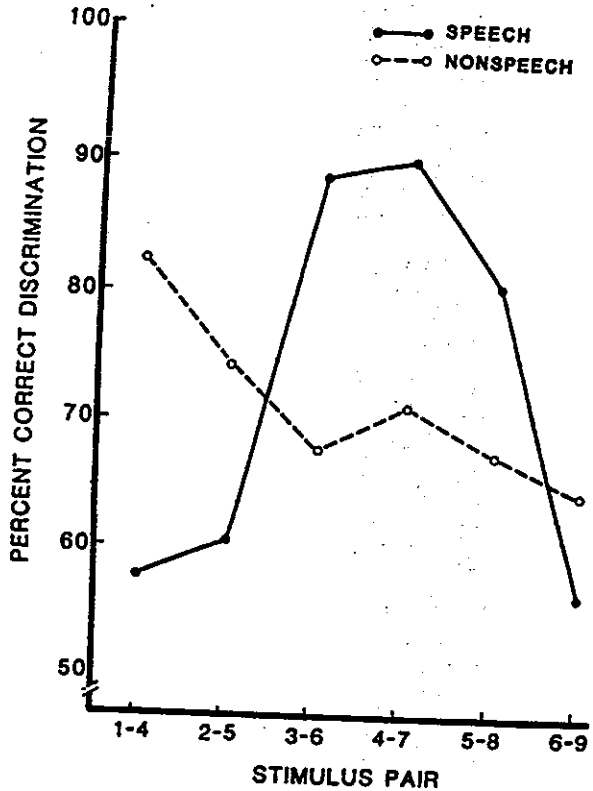

Figure 3: Results of the discrimination task in Mann and Liberman (1983). © Elsevier Sequoia, reprinted with permission.

speech condition are now relatively easy to discriminate. Performance varies somewhat across the continuum, but is always above chance. Performance is better than the speech task at the endpoints and worse in the middle.

If speech sounds are processed by typical auditory processes, why should there be such a difference? The AXB paradigm makes as few demands as possible on the labeling system of the subject, and yet there is a complete lack of correspondence between the two discrimination functions. If the nonspeech chirp, which is, after all, perceived as a separate object at the same time, were processed first and then combined with the speech base, it would seem that subjects should be able to make use of that fact and thus perform at higher levels at the endpoints. Instead, we find that the speech system makes use of the transitions to inform it about which linguistically significant gestures were present, and the degree to which these are apparent determines the degree to which the stimuli can be told apart.

4. Early availability of speech percepts compared with nonspeech percepts

Another version of duplex perception was first reported in Whalen and Liberman (1987). In this version, duplexity is attained not by splitting the signal between the two ears but rather by changing the amplitude relationship between one part and another. The stimuli again manipulated the F3 transitions that signaled the difference between "da" and "ga." In this case, the disparity between the base and the transition was made by using a frequency-varying sinusoid as the transition. In isolation, this sinusoid sounded like a whistle of changing pitch, one going up and one going down. The overall estimate of pitch was higher for the "da" whistle than for the "ga" whistle.

When combined with the base at extremely low intensities, these transitions have no effect on perception. At slightly higher levels, they distinguish "da" from "ga" even though the whistles are not perceptible. At higher levels still, the whistle begins to become apparent as a separate percept, but the speech is still perceived as appropriate to the category indicated by the transition. Thus we can obtain duplex perception even in a case where not only is the physical stimulation identical for both the speech and nonspeech percept, but the stimulation at the two ears is identical as well. Further, the speech is perceived at levels for which no whistle is even perceived, much less identified.

In our earlier work, we selected the levels at which to present the stimuli based on subjects' reports of when they began to hear the whistles. This methodology has been criticized as not finding the appropriate level of detection (Bailey & Herrmann, 1993). This criticism misses the point that if there is any level at which the speech can be accurately heard and the nonspeech cannot, there is evidence that the speech process is taking priority. Another way of emphasizing the same point, however, is to present an extremely broad range of intensities of the transition. If the range extends far enough to have stimuli at levels below the detectability of the whistle as well as above, then we should be able to see whether subjects really do make use of the transition for speech before they are able to make use of it for nonspeech.

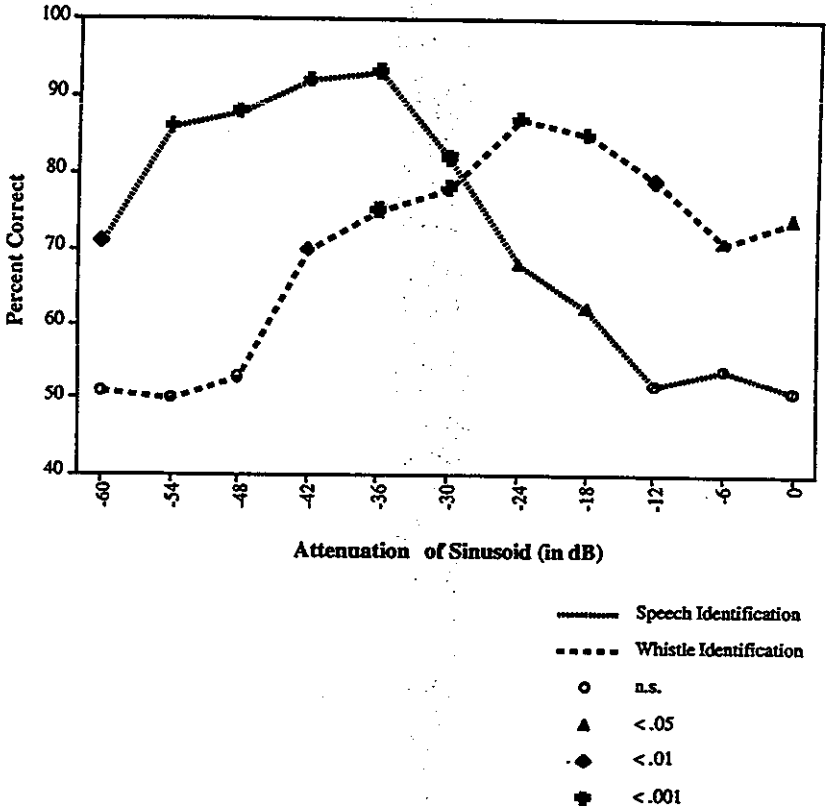

Figure 4: Results of identification tasks for a broad range of intensity values for the transition. The symbols indicate how reliably different from chance (50%) the values are. Adapted from Xu, Liberman and Whalen (in press).

Just such an experiment is reported in Xu, Liberman and Whalen (in press). For this experiment, we used an identification task rather than discrimination. Since there were only two transitions, one for "da" and one for "ga," it was easy to train subjects to identify them has "high" and "low" pitches, respectively. The syllables, of course, were identified as "da" or "ga." We then combined the base with the sinusoidal formants at eleven different intensity levels, six dB apart. Thus the difference between the most intense sinusoid and the least was 60 dB. The use of a different response type allows us to be sure that the results of duplex perception are general across perception and not restricted to the AXB discrimination paradigm.

Subjects were presented with the base with one of the transitions, whose amplitude varied through the levels specified, in random order. In one condition, they identified the whistle as "high" or "low." In another condition, they identified the speech as "da" or "ga." The results are shown in Figure 4. At the lowest

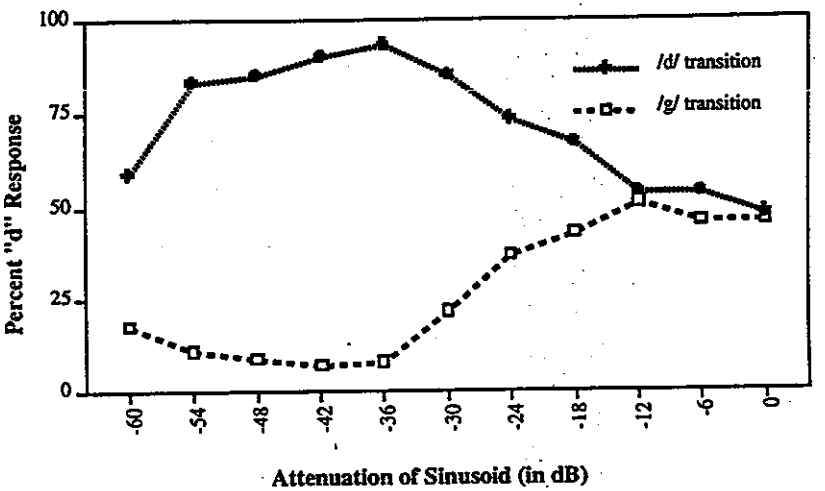

Figure 5: Identification of /d/ and /g/ transitions, combining with the base, as "da." Since "da" is the correct response for /d/ transitions, these rise up toward 100%. For /g/ transitions, 0% would be perfect performance. Adapted from Xu, Liberman and Whalen (in press).

intensities of the transition, the speech was already significantly above chance. It gained in reliability for the next several increases of intensity, and then began a decline which ultimately reduced the perception to chance levels with the most intense transitions. By contrast, the nonspeech identification began at chance levels at the lowest intensities, and did not become reliably better than chance until at least 18 dB higher than was necessary for the speech. Performance then improves for several amplitude levels, then declines slightly at the highest intensities.

Over a thirty decibel range, then, perception of these stimuli is duplex. Not only is the speech accurately heard, so is the whistle. The picture changes at both ends of the continuum. At the lowest intensities of the transition, only the speech is heard accurately. Indeed, many subjects complained about having to label the whistle when they did not, in fact, hear a whistle at all. This is the very definition of impenetrability in Fodor's (1983) theory of modularity: The information is clearly being used and is therefore audible, but the acoustic components cannot be brought to consciousness.

The two response functions are largely independent. While there is some visual appearance of reciprocity in the functions, the numbers argue for independence. The correlation between the two sets of values, for example, is -.39, which is not significant. Additionally, for five of the steps between members of the continuum, one of the functions goes up while the other goes down. However, this means that for five of those pairings, both functions go in the same direction. We can conclude, then, that these identification functions are independent. It seems difficult to explain this result without positing independent perceptual processes at work as well.

These results were presaged in a similar study by Vorperian, Ochs and Grantham (1995). They found extremely similar results with similar stimuli. The main difference was in the behavior of the speech identification at the highest intensities. Vorperian *et al.* had found that the highest intensities of the transitions elicited almost only "da" responses. They attributed this to the auditory "splatter" effect, in which intense low frequencies will spread upwards at high intensities. In our results, by contrast, the responses were evenly split between "da" and "ga" when the syllables were not accurately identified. This can be seen in Figure 5, which plots the percentage of "da" responses separately for the /d/ transition and the /g/ transition. The top function shows that the /d/ transition elicits significantly many "da" responses up until the final three intensity levels. The /g/ transition, by eliciting "ga" responses, is significantly below chance when stated in terms of "da" responses, but it also elicits as many "da" as "ga" responses for the three most intense levels. We are at a loss to explain the difference between these results, but the amount of agreement is impressive: Both studies show again that the speech perception system makes use of speech information in a way that is different from the nonspeech perception and at intensity levels at which the nonspeech percept is not available.

5. Conclusion

The treatment of the acoustic signal is radically different in the speech mode than in the nonspeech mode. Speech perception treats the acoustic signal as information about the source, which is interpreted in terms of the gestures produced by the human vocal tract. This conclusion has been supported over the years in many different experiments using diverse techniques. One of those techniques is duplex perception, and two of those findings were highlighted here (Mann & Liberman, 1983; Xu, et al., in press).

Nonspeech psychoacoustics does not account for this data. There is significant contribution of the transition even when its nonspeech character is inaudible. The level at which this nonspeech percept is available, then, has no bearing on when it will provide phonetic information. Acoustic theories that hold that psychophysics will reveal the lower limits on what can form a speech percept are thus called into question. This doubt is similar to that found in the existence of auditory localization, since the localization system uses timing differences that are two orders of magnitude smaller than those detectable in psychophysical tasks. The completely different response functions for speech and nonspeech in the studies reported on here make it more clear that psychophysics of nonspeech analogs is largely orthogonal to the study of speech perception.

These results are accommodated most easily if we posit a biological specialization for speech. The profound impact that speech has on our lives is evident not only in the great use we make of it as adults but also in the extreme conditions required for children *not* to acquire spoken language. For hearing children with unimpaired articulation, it practically takes a complete absence of language input for the child to avoid developing speech. The biological details remain to be worked out, of course. For example, despite the unique characteristics of human language, we can expect to find overlap and correspondences to other signalling systems throughout the animal kingdom. No other species, though, has the overarching need to perceive speech as speech. Duplex perception shows that humans are so determined to hear speech that they will do so even when ecological reality tells them they shouldn't. This fact, far from being a bizarre footnote in

psychophysics, tells us that the speech system takes precedence in the perception of the acoustic stream.

Acknowledgments

The writing of this paper was supported by NIH grants HD-01994 and DC-02717 to Haskins Laboratories. Thanks go to Alvin M. Liberman and Yi Xu for permission to reprint their figures. Helpful comments were provided by Alvin M. Liberman, Carol A. Fowler, Michael Turvey, Stephen R. Anderson, Jennifer Pardo, Bruno H. Repp, and Harriet Magen.

References

Bailey, Peter J., and Paul Herrmann. 1993. A Reexamination of Duplex Perception Evoked by Intensity Differences. Perception and Psychophysics, 54.20-32.

Best, Catherine T. 1984. Discovering Messages in the Medium: Speech and the Prelinguistic Infant, Advances in Pediatric Psychology , ed. by H. E. Fitzgerald, B. Lester, & M. Yogman, 97-145. New York: Plenum Press.

Best, Catherine T. 1995. A Direct Realist Perspective on Cross-Language Speech Perception, Cross-Language Speech Perception , ed. by W. Strange & J. J. Jenkins, 171-204. Timonium, MD: York Press.

Bregman, Albert S. 1990. Auditory Scene Analysis. Cambridge, MA: MIT Press.

Ciocca, Valter, and Albert S. Bregman. 1989. The Effects of Auditory Streaming on Duplex Perception. Perception and Psychophysics, 46.39-48.

Ciocca, Valter; Albert S. Bregman; and Kathleen L. Capreol. 1992. The Phonetic Integration of Speech and Non-Speech Sounds: Effects of Perceived Location. Quarterly Journal of Experimental Psychology, 44A.577-593.

Darwin, Christopher J.; Peter Howell; and Susan A. Brady. 1978. Laterality and Localization: A Right Ear Advantage for Speech Heard on the Left, Attention and Performance VII , ed. by J. Requin, 261-278. Hillsdale, NJ: Lawrence Erlbaum Associates.

Diehl, Randy L., and Keith R. Kluender. 1989. On the Objects of Speech Perception. Ecological Psychology, 1.121-144.

Fodor, Jerry A. 1983. The Modularity of Mind. Cambridge, MA: MIT Press.

Fowler, Carol A. 1986. An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective. Journal of Phonetics, 14.3-28.

Fowler, Carol A., and Mary R. Smith. 1986. Speech Perception as "Vector Analysis": An approach to the Problems of Invariance and Segmentation, Invariance and Variability in Speech Processes , ed. by J. S. Perkell & D. H. Klatt, 123-136. Hillsdale, NJ: Lawrence Erlbaum Associates.

Liberman, Alvin M.; Franklin S. Cooper; Donald P. Shankweiler; and Michael Studdert-Kennedy. 1967. Perception of the Speech Code. Psychological Review, 74.431-461.

Liberman, Alvin M., and Ignatius G. Mattingly. 1985. The Motor Theory of Speech Perception Revised. Cognition, 21.1-36.

Liberman, Alvin M., and Ignatius G. Mattingly. 1989. A Specialization for Speech Perception. Science, 243.489-494.

Lisker, Leigh, and Arthur S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. Word, 20.384-422.

Mann, V. A., and A. M. Liberman. 1983. Some Differences between Phonetic and Auditory Modes of Perception. Cognition, 14.211-235.

Moore, Brian C. J. 1982. An Introduction to the Psychology of Hearing (2nd edition ed.). New York: Academic Press.

Rand, Timothy C. 1974. Dichotic Release from Masking for Speech. Journal of the Acoustical Society of America, 55.678-680.

Richards, Whitman. 1971. Anomalous Stereoscopic Depth Perception. Journal of the Optical Society of America, 61.410-414.

Schouten, M. E. H., and A. J. van Hessen. 1993. Modeling Phoneme Perception. I: Categorical Perception. Journal of the Acoustical Society of America, 92.1841-1855.

Shankweiler, Donald P., and Michael Studdert-Kennedy. 1967. Identification of Consonants and Vowels Presented to Left and Right Ears. Quarterly Journal of Experimental Psychology, 19.59-63.
Stevens, Kenneth N., and Sheila E. Blumstein. 1978. Invariant Cues for Place of Articulation in Stop Consonants. Journal of the Acoustical Society of America, 64.1358-1368.
Summerfield, Quentin. 1982. Differences Between Spectral Dependencies in Auditory and Phonetic Temporal Processing: Relevance to the Perception of Voicing in Initial Stops. Journal of the Acoustical Society of America, 72.51-61.
Vorperian, Houri K.; Marleen T. Ochs; and D. Wesley Grantham. 1995. Stimulus Intensity and Fundamental Frequency Effects on Duplex Perception. Journal of the Acoustical Society of America, 98.735-744.
Whalen, D. H., and Alvin M. Liberman. 1987. Speech Perception Takes Precedence over Nonspeech Perception. Science, 237.169-171.
Xu, Yi; Alvin M. Liberman; and D. H. Whalen. in press. On the Immediacy of Phonetic Perception. Psychological Science, in press.

CLS 33: The Panels
(The Proceedings from the Panels of the Chicago Linguistic Society's Thirty-third Meeting. Volume 33-2)

Library of Congress Catalog Card Number 76-27943

ISSN 0577-7240
ISBN 0-914203-53-3

For information on other publications by the Chicago Linguistic Society contact CLS, 1010 E. 59th Street, Chicago, IL 60637. (773) 702-8529. cls@tuna.uchicago.edu