



OCCAM'S RAZOR IS A DOUBLE-EDGED SWORD: REDUCED INTERACTION IS NOT NECESSARILY REDUCED POWER

D. H. Whalen

Haskins Laboratories, New Haven

ABSTRACT

Although Norris, McQueen and Cutler have provided convincing evidence against the lexicon influencing phonetic processing, their simplification has costs for the processes themselves. While their arrangement may ultimately prove correct, its validity is not due to Occam's Razor. Their statistical accumulation presupposes units of a particular size for those accumulations, including trans-segmental combinations. Listeners parse the signal into the various phonetic units in a fairly complete way, without strict left-to-right processing. If the results of that parsing are the units at which the statistical properties accumulate, then the parsing must be complete before the statistical likelihoods can act, or the statistical likelihoods are part of parsing itself. The first possibility seems to require yet another level between speech and the lexicon; the second possibility requires a far more complex speech processor than previously proposed. At this point, it is only possible to make preliminary suggestions for a resolution.

1. INTRODUCTION

The perception of speech is performed by a biological specialization that has evolved over the last million years or so to enable communication between members of the human species and to make speech researchers look stupid. The process is so automatic and necessary for humans that it disguises the complexity that exists in the way the segments are presented to the ear. The ear does not have enough resolving power to have an acoustic alphabet that would present phoneme-sized sounds one after another and still have an efficient rate of communication [1]. Instead, the sounds we receive are highly overlapped, or coarticulated, so that an efficient stream can be presented. It is only because the gestures that create those sounds and the interpretation of those sounds as gestures coevolved that we are able to speak at typical rates of 10 or 12 phonemes per second [2].

What comes naturally to the human ear (and, of course, the brain attached to it) is difficult for the speech scientist. After 50 years of relatively easy instrumental access to the acoustic structure of speech, we still have not deciphered the basic units that are used perceptually. We have had proposals of features [3], segments [4], diphones [5], triphones [6], syllables [7], acoustic landmarks [8], successive speech spectra [9], gestures [10], and, perhaps, none at all [11]. I do not propose here to decide among these various alternatives. What is important to note is that there is no clearly superior way of analyzing the speech signal that would convince

researchers to agree on the result as happened with, say, the introduction of the periodic table in chemistry. The evidence for such a breakthrough does not yet exist.

There are many levels of linguistic structure that can be shown to affect perception, so it does not appear likely that there is a single perceptual unit. Subjects can monitor for some of these units, including at least phonemes, syllables, bisyllables and words [12-14], though disjoint groups like /b/ and /s/ have been used as well [15]. Monitoring for features is more difficult, since most monitoring tasks are interpreted in terms of letters, and features are not well represented in letters. But listeners are aware of the features, showing the "oddity" response of mismatch negativity in a magnetoencephalography paradigm [16]. But beyond the metalinguistic task of monitoring, it has been found that virtually every acoustic consequence of speech articulation affects perception [17-21]. Further, these levels of description interact, most notably when the constraints of syllable affiliation lead to trading relations within a single phonetic dimension that informs both consonant and vowel judgments [22]. Again, this discussion is not to resolve this issue, but to indicate that there are levels of complexity that have yet to be sorted out.

Every theory assumes some primitive, even the theories that claim not to have them, and these primitives are important for the details of how frequency of occurrence affects the organization of phonology and the lexicon. My own theory [1, 23] has gestures as the primitive. This accounts well for the overlap of gestures in production and the untangling of that overlap in perception. It is compatible with early facility in speech perception [24, 25] and the ability to imitate [26]. One of its biggest challenges is to explain the discrepancy between the lack of clear indications of a segmental structure with the wildly successful nature of orthographies that tap into just that segmental level. It seems very unlikely that such an entity would be created by the association with a letter, though there is some evidence that says just that [27-29]. If phonemes are made up of gestures and gestures are recognized first, gestural theory should have as good an account for lexical access as a theory that posits features which are then, perhaps, gathered into phoneme-sized bundles [3].

As stated above, even the connectionist theories that claim to be neutral as to the levels of analysis inherent in the system make assumptions about the elements in the system. Every connectionist model must begin with an input, which is, in the best case, an unanalyzed set of spectra [30, 31]. Even this reduction counts as establishing units. In the gestural model, for example, these spectral slices are not significant and would not be

possible units at all. If the connectionist models later develop units at various levels that we are familiar with from compositional linguistic analysis, so much the better; but a successful model is under no obligation to do so. The nodes at which frequencies can accumulate are specified at least in number and sometimes in structure before any input is received, presumably in a fashion similar to the availability of neurons in the brain. The models must make assumptions about those elements that we know are wrong. Further, our understanding of what is involved in this case is limited by the danger in interpreting the hidden level nodes. Still, the overlap with structurally determined categories is sufficient to make a case that the lexical activation can influence the perceptual portion of the network.

2. REPORTED RESULTS

The most convincing case of lexical input was that of Elman and McClelland [32], until Pitt and McQueen [33] provided contrary evidence. A segment whose major acoustic realization is replaced by noise can be expected to be perceptually filled in [34, 35], but that the restored segment should have a coarticulatory influence is unexpected. What seems even more unexpected is Pitt and McQueen's finding that the lexicon is not filling in the gap, but phoneme probabilities are. Although these authors were more concerned with the issue as defined by the earlier research, their result opens new possibilities that have yet to be fully explored. They show that the speech perception mechanism utilizes a parsing strategy [19, 36, 37], but that this parsing is sensitive to likelihoods in cases where positive evidence is missing. Here is where the issue of the units becomes critical.

Frequencies can only accumulate on an element in a system, so the elements that are posited to explain the coarticulatory results of Pitt and McQueen must be perceptually relevant. It may be the case that they are not primary, since we have already seen that multiple levels of description are needed for speech, but those levels must exist. Thus whenever a nonlexical frequency effect is found in the phonetic system (as in Pitt and McQueen), the system must have that combination of elements present to attach the frequency to. Considering that every combination of words seems to be a potential site for the effects that Pitt and McQueen found, this means that there must now be a vastly increased number of units in the perceptual system. It is no longer possible to derive the combined frequencies by rule from the lexical frequencies involved: The strings of phonemes must be represented without the interference of the lexicon.

3. CONCLUSION

The upshot is that the architecture is simpler in that the lexicon does not directly influence the phonetic processing, but the phonetic processing is greatly increased in complexity. Any time there is a string of phonemes, it appears, they will generate a sequence that

begins to accumulate a frequency count. The effects that Pitt and McQueen discuss go from one segment to another across a word boundary, but, it seems likely that there will be larger components that will be represented as well. Coarticulation spreads beyond the syllable [38-40], and the perceptual parsing mechanism takes these larger influences into account. The size needed may extend up even to well-memorized passages, at least for individuals. All of these are hierarchically related, but predicting them from the lexicon would seem to be, in terms of storage requirements at least, simpler.

The added complexity in the phonetic component could well be necessary, however. The fact that Pitt and McQueen obtained their results with nonwords as well as words indicates that there is a level of description that transcends the lexicon. Using the lexicon to make "analogous" probabilities for nonwords may make for more complications than the addition of superordinate elements in the phonetic component. At this stage of our theorizing, it seems premature to assign complexity values to any description. Our theories must stand on whether they explain how things work, not on how simply they do it.

In closing, let me give an example of a lexical influence that Cathi Best, Julia Irwin and I found recently [41] and which does not immediately seem to avoid having the lexicon influence perception. We wanted to explore infants' perception of allophones in their own language, to see at what stage they became sensitive to whether the allophones were in the typical syllable position or not. In the course of trying to validate our stimuli with adults, we discovered that *adults* were not sensitive to the difference: They could distinguish the aspirated /p/ from the unaspirated /p/, but they preferred the aspirate even in the syllable position where it was less appropriate. Even when we asked our listeners to detect a foreign accent (which seemed like a reasonable possibility with allophones), they had the same preference. All of these tests were with nonwords, which we wanted to use with the infants so that overall frequency of occurrence would be (essentially) zero. But when we put those same allophones in words, the listeners preferred the appropriate allophone. We are following this study up with work on allophones of /l/, in which we will manipulate word frequency as well, but the remarkable thing is that there is an effect of lexicality at all. Certainly we would have expected, from our own theoretical stance and the results of Pitt and McQueen, that lexicality would have no effect here.

This is the kind of puzzle that is revealed by the second edge of Occam's razor, opening up new depths as it slices on the upswing as well as the down. We may have severed the downward links from the lexicon into the speech perception mechanism, but we have exposed that mechanism as a far more statistically inclined machine than seemed necessary if the lexicon could handle the probabilities. What we need to find out now is how general these statistical properties are, how they develop in language acquisition, and whether the statistical properties can be seen in on-line processing

via neural imaging. The result will be a far better understanding of the perceptual processes that allow us to communicate with each other.

4. ACKNOWLEDGMENTS

The writing of this paper was supported by NIH grants HD-01994, DC-02717, and DC-00403 to Haskins Laboratories. I thank Carol A. Fowler, Ken Pugh and Matthew Richardson for helpful comments. Mailing address: D. H. Whalen, Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA. Email: whalen@haskins.yale.edu.

5. REFERENCES

- [1] Liberman, A.M. & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, **4**, 187-196.
- [2] Maclay, H. & Osgood, C.E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, **15**, 19-44.
- [3] Jakobson, R., Fant, G. & Halle, M. (1951). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press
- [4] Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.
- [5] Ghitza, O. & Sondhi, M.M. (1997). On the perceptual distance between speech segments. *Journal of the Acoustical Society of America*, **101**, 522-529.
- [6] Wickelgren, W.A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, **76**, 1-15.
- [7] Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychological enquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates
- [8] Stevens, K.N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press
- [9] Klatt, D.H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In Cole, R.A. (Ed.), *Perception and production of fluent speech* (pp. 243-288). Hillsdale, NJ: Lawrence Erlbaum Associates
- [10] Fowler, C.A. & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In Perkell, J. & Klatt, D. (Ed.), *Invariance and variability in speech processes* (pp. 123-136). Hillsdale, NJ: Lawrence Erlbaum Associates
- [11] Marslen-Wilson, W. & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, **101**, 653-675.
- [12] Foss, D.J. & Swinney, D.A. (1973). On the psychological reality of the phoneme: Perception, identification and consciousness. *Journal of Verbal Learning and Verbal Behavior*, **12**, 246-257.
- [13] Savin, H.B. & Bever, T.G. (1970). The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, **9**, 295-302.
- [14] Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121.
- [15] Rubin, P.E., Turvey, M.T. & van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception and Psychophysics*, **19**, 394-398.
- [16] Phillips, C., Marantz, A., McGinnis, M., Pesetsky, D., Wexler, K., et al. (1995). Brain mechanisms of speech perception: A preliminary report. *MIT Working Papers in Linguistics*, **26**, 153-191.
- [17] Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech*, **29**, 3-11.
- [18] Whalen, D.H. (1991). Perception of the English /s/-/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *Journal of the Acoustical Society of America*, **90**, 1776-1785.
- [19] Whalen, D.H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, **35**, 49-64.
- [20] Liberman, A.M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press
- [21] Best, C.T., Morrongoello, B. & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, **29**, 191-211.
- [22] Whalen, D.H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception and Psychophysics*, **46**, 284-292.
- [23] Whalen, D.H. (1999). Three lines of evidence for direct links between production and perception in speech. In Ohala, J.J. et al. (Ed.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1257-1260). San Francisco: University of California, Berkeley
- [24] Eimas, P.D. & Miller, J.D. (1992). Organization in the perception of speech by young infants. *Psychological Science*, **3**, 340-345.
- [25] Werker, J.F. (1989). Becoming a native listener. *American Scientist*, **77**, 54-59.
- [26] Studdert-Kennedy, M. (in press). Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function. In Knight, C., Studdert-Kennedy, M. & Hurford, J.R. (Ed.), *The emergence of language: Social function and the origins of linguistic form*, Cambridge: Cambridge University Press
- [27] Morais, J., Content, A., Bertelson, P., Cary, L. & Kolinsky, R. (1988). Is there a critical period for the acquisition of segmental analysis? *Cognitive Neuropsychology*, **5**, 347-352.
- [28] Morais, J., Cary, L., Alegria, J. & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, **7**, 323-331.
- [29] Castro-Caldas, A., Petersson, K.M., Reis, A., Stone-Elander, S. & Ingvar, M. (1998). The illiterate brain: Learning to read and write during childhood

influences the functional organization of the adult brain. *Brain*, **121**, 1053-1063.

[30] Elman, J.L. & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, **83**, 1615-1626.

[31] McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

[32] Elman, J.L. & McClelland, J.L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, **27**, 143-165.

[33] Pitt, M.A. & McQueen, J.M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, **39**, 347-370.

[34] Samuel, A.G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, **110**, 474-494.

[35] Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, **167**, 392-393.

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

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

[38] Magen, H.S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, **25**, 187-205.

[39] Manuel, S.Y. (1987). *Acoustic and perceptual consequences of vowel-to-vowel coarticulation in three Bantu languages*. Unpublished Ph.D. thesis, Yale University.

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

[41] Whalen, D.H., Best, C.T., & Irwin, J.R. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics*, **25**, 501-528.