ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 405

COCHLEAR PROSTHESES

AN INTERNATIONAL SYMPOSIUM

Edited by Charles W. Parkins and Samuel W. Anderson

The New York Academy of Sciences
New York, New York
1983

LIMITS ON ALTERNATIVE AUDITORY REPRESENTATIONS OF SPEECH*

Michael Studdert-Kennedy

Queens College and Graduate Center
City University of New York
Flushing, New York 11367; and
Haskins Laboratories
New Haven, Connecticut 06510

The history of attempts to construct reading machines for the blind may guide us in our attempts to devise speech-listening aids for the deaf. The goal of the early reading machine work was to construct an acoustic alphabet, that is, to find a set of discrete, discriminable acoustic patterns that might substitute for the visual alphabet. To devise such a set—of tones, chords, bursts of filtered noise, perhaps varying in amplitude or duration—was not difficult, and, if the patterns were presented one at a time in a comfortable test format, listeners readily learned to identify them. However, if patterns were presented in rapid sequence, listener performance dropped precipitously. In fact, no one has yet devised an acoustic alphabet more effective than the dots and dashes of Morse code, with which highly skilled operators may reach a rate of 30–40 words a minute—roughly one-fifth of the rate at which we typically follow spoken language. Not surprisingly, the search for an acoustic alphabet for reading machines has been largely abandoned in favor of synthetic (or compiled) speech.

The failure to devise a viable acoustic alphabet seems all the more surprising when we consider the success of the visual alphabet. On the face of it, one might have expected that transposition into another sensory modality would have been more damaging—certainly more "unnatural"—than remaining within the biologically given modality of sound. But, as we all know, that is far from the case. Even without benefit of special "speed-reading" instruction, reading rates of 300–400 words a minute are commonplace.

What are we to make of this paradox? Why do speech and a visual alphabet succeed where a sound alphabet fails? The answer seems to lie, first, in certain properties of the visual and auditory systems, and second, in the different types of information that speech and alphabets convey.

Although the temporal distribution of light carries important information for the perception of an event as it develops in time, the spatial distribution of light at any instant is sufficient to specify a stationary object. In fact, the eye is particularly sensitive to spatial patterns of contour and it is this sensitivity that writing systems exploit: we read by a series of static fixations during which information from many points in the visual field is gathered simultaneously. By contrast, sound is never static; we perceive auditorily by virtue of the way in which events structure the spectrum over time. Recent studies of the response to speech of auditory nerve fibers in cat [1-3] have demonstrated that the nerve

*This work was supported by Grant HD 01994 from the National Institute of Child Health and Human Development of the National Institutes of Health.

is particularly responsive to abrupt discontinuities in amplitude and frequency: at such instants of discontinuity, there occurs a rapid increase or decrease in firing rate, reflecting an increase or decrease in the number of fibers sampling the input. Such on- and off-responses may serve as a first stage in segmenting the speech signal.[7] Moreover, the spectral structure [4] or shifts in spectral structure [5] in the tens of milliseconds following an abrupt onset may carry important acoustic–phonetic information. Of course, more gradual shifts in spectral structure with transitions into and out of relatively steady-state portions of the signal are also reflected in the temporal structure of cat neural response [2, 6] and, in the human, presumably serve to specify phonetic elements.[7] My point here is simply that the dynamic structure of speech seems nicely tuned to evade (or exploit) the limits imposed by refractory neural response and by neural adaptation and masking. We may suspect then that the failure of acoustic alphabets in reading machines stemmed, in part, from their use of static segments modeled on those of the visual alphabet: the repeated abrupt breaks in acoustic pattern, inevitable in a sequence of arbitrary static elements, ran foul of the refractory, adaptive and masking characteristics of the auditory system. There would seem to be no reason, in principle, why an acoustic alphabet more shrewdly tailored to the requirements of the auditory system should not work. But then, of course, it might be simpler to use a facsimile of the acoustic structure that evolution has prescribed: synthetic speech.

We come now to a second and deeper reason for the differences between speech and an acoustic alphabet: The information conveyed by speech is quite different from the information conveyed by an alphabet (or by any other secondary mode of representation such as a syllabary or logographic system). Speech conveys information about the gestures that produced it. There is nothing intrinsically linguistic in these gestures. They have, of course, been selected by a long process of evolution and historical development to answer the motoric and perceptual demands of linguistic communication. But the linguistic value of the gestures is available only to the listener (or viewer) who knows something of the language being spoken. The gestures themselves are simply a complex, coordinated pattern of articulatory activity, and the function of the speech signal is to specify this activity. By contrast, the elements of an alphabet specify nothing. Their structure is arbitrary, constrained only by the perceptual capacities of their users. Their structure can be arbitrary precisely because their function is not to specify, but to indicate (or symbolize) something other than themselves, namely, the elements of a language.[8]

What does all this have to do with listening aids for the deaf? Let me begin by justifying the claim that the function of the speech signal is to specify articulation. The value of lip-reading for the deaf is well known, but the implications of this skill for an understanding of normal speech perception have only recently begun to emerge. McGurk and MacDonald [9] discovered that listeners' perception of a spoken utterance can be altered if they simultaneously watch a videotape of a speaker articulating a different utterance. For example, if subjects are presented with four spoken repetitions of the syllable [ba], while watching a synchronized videotape of a speaker articulating the sequence [ba, va, ða, da], it is the latter sequence that they will typically report. Or again, if subjects are presented with audio [ba] in synchrony with video [ga], they will typically report [bga], [gba] or [da]. Here the integrated percepts are either clusters corresponding to both the acoustical and the optical patterns presented, or a merged segment [da], corresponding to neither.[10, 11]

Roberts and Summerfield [12] took advantage of this last effect to demonstrate the distinction between surface acoustic structure and the phonetic percept in a study of audio-visual adaptation. The standard adaptation procedure calls for forced-choice, two-category identification of items drawn at random from a synthetic speech continuum, along which some acoustic correlate of a single phonetic feature has been systematically varied, as from [ba] to [da] or from [da] to [ta]. Subjects make their judgments both before and after being exposed to (that is, adapted with) many repetitions of a good exemplar from one or other of the two endpoint categories. The effect of this adaptation procedure is a significant drop in the frequency with which subjects assign ambiguous items from the continuum to the category from which the adapting syllable was drawn.

Roberts and Summerfield [12] gave a novel twist to this paradigm. On a [bɛ-dɛ] continuum they compared the effect of a standard auditory adaptor, [bɛ], with that of an audio-visual adaptor, auditory [bɛ], visual [gɛ], intended to be perceived as [dɛ]. Of their twelve subjects, six reported the audio-visual adaptor most of the time as [dɛ] or [ðɛ], four as [klɛ], one as [flɛ], and one as [ma]. Not one of the subjects reported the auditory syllable actually presented, namely, [bɛ]. Yet the adaptation effect was purely auditory: every subject displayed a drop in the number of items identified as [bɛ], roughly equal to the drop for the control condition in which auditory [bɛ] was presented alone. Thus, while listeners' auditory systems were normally adapted by the acoustic input, their conscious phonetic percepts were specified by a blend of acoustical and optical information.

The basis of this acoustical–optical blend is not understood. The acoustical information is presumably carried by the familiar pattern of formants, friction noise, plosive release, harmonic variation, and so on; apparently, the optical information is carried by varying configurations of the lips and, perhaps, of the tongue and teeth.[11] But how these qualitatively distinct patterns of light and sound are combined to yield an integrated percept is not obvious. We can certainly point to covarying properties in, for example, the opening of the mouth and a pattern of formant movement, or in the rounding of the lips and a downward shift in spectral structure. But these are mere correlations: They are not directly given in perception—unless we adduce an *ad hoc* theory of learned association. And this seems implausible, in light of the discovery that 6-month-old infants recognize structural correspondences between patterns of speech and facial movement.[13, 14] What we need is some deeper, underlying metric, common to both the light reflected and the sound radiated from mouth and lips.

Insight into the nature of this metric may come from the phenomenon of imitation. We are not surprised that if someone pouts her lips, someone else can pout in imitation, because the capacity to imitate seems "natural," so natural, in fact, that even 36-hour-old infants can do it.[15, 16] Nonetheless, there is a puzzle here. How does the imitator transpose a pattern of light into a pattern of movement? Or, as Gibson's followers ask, how do we get light into a muscle? What seems to be required is that we perceive in a movement not simply its surface kinematics (such as displacements, velocities, and accelerations), but also the biophysical forces that define the structure to be moved and that produce the movement, in other words, the underlying dynamics (mass, elasticity, force, and momentum, for example). Notice that we require an equivalent (and, surely, no less biologically specialized) perceptual capacity to

account for our ability to repeat a spoken utterance. The "kinematics" of the utterance, that is, the changes and rates of change in sound source and spectral structure, as illustrated in the spectrogram, are incommensurate with the muscular controls that produce them. How do we get sound into a muscle? We are led, it seems, by the facts of imitation to suppose that we perceive in both the sound and the sight of speech not only the surface kinematics, but also the underlying dynamics of articulation.[17-19]

For present purposes, the import of this brief discussion is simply that the forces controlling articulation are not revealed in the static spectral sections of the spectrograph. Therefore, if the speech percept is, indeed, properly specified in a ". . . metric closely related to that of articulatory dynamics," [11] as the facts of imitation seem to require, then any viable alternative auditory representation of speech must, first and foremost, preserve the information that makes these dynamics available to the listener: not the static spectral patterns putatively corresponding to the static elements of phonetic description, but the subtle patterns of changing spectral structure.

The point is vividly illustrated by the phenomenon of sinewave speech.[20] Here the spectral structure is drastically reduced. The speech signal is whittled down to a temporal pattern of change in three (or even two) sinewaves, roughly following the center frequencies of the first three (or two) formants. Yet the speech remains surprisingly intelligible: Many listeners (perhaps most listeners, with practice) can understand unknown sentences—even, incidentally, when all amplitude changes corresponding to syllabic rise and fall have also been eliminated.[21] In effect, the sinewave transformation removes all information concerning the sound source (voicing, frication, plosive release), but retains (sharply reduced) information concerning instants of silence (for example, stop closure) and the partially reciprocal variations in size and shape of the front and back cavities of the vocal tract.[22] However, what is crucial here is that the information preserved is not simply some trace of the formant structure, sufficient to specify instantaneous cavity relations, but the temporal patterns of spectral change that specify the forces controlling the movements by which cavity shapes are determined.

The nature of these temporal patterns may be clarified if we consider another, closely related study. Risberg and Agelfors [23] report the results of an experiment in which the speech signal was reduced to a single sinewave modulated in frequency to match the center of gravity, and in amplitude to match the total energy, of the spectrum. With practice, subjects were able to identify up to nearly 90% of a closed set of twelve Swedish spondee words and up to 50% of the Swedish numerals from 13–99. The authors suggest that, while their minimal spectral display evidently preserved enough information for the identification of fricatives and certain vowels, subjects may have relied largely on the timing of word or syllable onsets and offsets. It would seem that these timing patterns may have sufficed to distinguish among the elements of a closed set, very much as some arbitrary, unique property of a word may serve for artificial machine recognition.

My point here is not that a single sinewave representation might not, by judicious weighting of its spectral determinants, be made to specify the phonetic structure of unknown sentences. That might very well be the case. My point is rather to distinguish between temporal patterns of syllable onsets and offsets, which may serve to distinguish among a closed set of words or word sequences, and a temporal pattern of spectral change that specifies the

dynamics of articulation. The latter entails rates of change in overall structure, as well perhaps as delicate shifts in the temporal alignment of formant movement that reflect interarticulator timing. It is something of this that the earlier sinewave studies [19] seem to have preserved, if we may judge from their subjects' success in identifying unknown phonetic sequences.

In conclusion, the movements of speech are not disembodied. They are the movements of a biologically given articulatory system with characteristic configurations. At any instant, the configuration may be described in terms of articulator placement (as in traditional phonetics) and may be specified, in principle, by the static morphology of lips and mouth or by the spectral structure of sound. But speech only emerges from systematic changes in articulator placement and interarticulator organization. It is to change that the auditory system is particularly sensitive, and it is in spectral change that the prepared observer finds the dynamic patterns that specify the phonetic event. The limits on an alternative auditory representation of speech would seem then to be that it preserve some minimal, perhaps grossly transposed, spectral structure, sufficient to specify instantaneous articulator placement, and that this structure be such that it can be modulated over time to specify the underlying articulatory dynamics.

Acknowledgments

My thanks to Robert Remez and Philip Rubin for useful discussion.

References

1. Delgutte, B. 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J. Acoust. Soc. Am. 68: 843–857.
2. Delgutte, B. 1982. Some correlates of phonetic distinctions at the level of the auditory nerve. *In* The Representation of Speech in the Peripheral Auditory System. R. Carlson and B. Granstrom, Eds.: 131–149. Elsevier Biomedical Press. New York, NY.
3. Kiang, N. Y. S. 1980. Processing of speech by the auditory nervous system. J. Acoust. Soc. Am. 68: 830–835.
4. Stevens, K. N. & S. E. Blumstein. 1978. Invariant cues for place of articulation. J. Acoust. Soc. Am. 64: 1358–1368.
5. Kewley-Port, D. 1980. Representations of spectral change as cues to place of articulation in stop consonants. Unpublished Ph.D. dissertation, City University of New York, New York, NY.
6. Sachs, M. B., E. D. Young & M. I. Miller. 1982. Encoding of speech features in the auditory nerve. *In* The Representation of Speech in the Peripheral Auditory System. R. Carlson and B. Granstrom, Eds.: 115–130. Elsevier Biomedical Press. New York, NY.
7. Chistovich, L. A., V. V. Lublinskaya, T. G. Malinnikova, E. A. Ogorodnikova, E. I. Stoljarova & S. Ja. Zhukov. 1982. Temporal processing of peripheral auditory patterns of speech. *In* The Representation of Speech in the Peripheral Auditory System. R. Carlson and B. Granstrom, Eds.: 165–180. Elsevier Biomedial Press. New York, NY.
8. Carello, C., M. T. Turvey, P. N. Kugler & R. E. Shaw. Inadequacies of the computer metaphor. *In* Handbook of Cognitive Neuroscience. M. S. Gazzaniga, Ed. Plenum Press. New York, NY. In press.

9. McGurk, H. & J. MacDonald. 1976. Hearing lips and seeing voices. Nature **264:** 746–748.
10. MacDonald, J & H. McGurk. 1978. Visual influences on speech perception processes. Percept. Psychophys. **24:** 253–257.
11. Summerfield, Q. 1979. Use of visual information for phonetic perception. Phonetica **36:** 314–331.
12. Roberts, M. & Q. Summerfield. 1981. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Percept. Psychophys. **30:** 309–314.
13. Kuhl, P. K. & A. N. Meltzoff. 1982. The bimodal perception of speech in infancy. Science **218:** 1138–1141.
14. MacKain, K., M. Studdert-Kennedy, S. Spieker & D. Stern. Infant intermodal speech perception is a left hemisphere function. Science. In press.
15. Field, T. M., R. Woodson, R. Greenberg & D. Cohen. 1982. Discrimination and imitation of facial expressions by neonates. Science **218:** 179–181.
16. Meltzoff, A. N. & M. K. Moore. 1977. Imitation of facial and manual gestures by human neonates. Science **198:** 75–76.
17. Fowler, C. A., P. Rubin, R. E. Remez & M. T. Turvey. Implications for speech production of a general theory of action. *In* Language Production. B. Butterworth, Ed.: 373–420. Academic Press. New York, NY.
18. Runeson, S. & G. Frykholm. 1981. Visual perception of lifted weight. J. Exp. Psychol.: Hum. Percep. & Perform. **7:** 733–740.
19. Summerfield, Q. 1980. The structuring of language by the requirements of motor control and perception: Group Report. *In* Signed and Spoken Language: Biological Constraints on Linguistic Form. U. Bellugi & M. Studdert-Kennedy, Eds.: 89–114. Verlag Chemie. Deerfield Beach, FL.
20. Remez, R. E., P. E. Rubin & D. B. Pisoni. 1983. Coding of the speech spectrum in three time-varying sinusoids. This volume.
21. Remez, R. E., P. E. Rubin & T. D. Carrell. 1981. Phonetic perception of sinusoidal signals: effects of amplitude variation. J. Acoust. Soc. Am. **69**(S1): 114(A).
22. Kuhn, G. M. 1975. On the front cavity resonance and its possible role in speech perception. J. Acoust. Soc. Am. **58:** 428–433.
23. Risberg, A. & E. Agelfors. 1982. Speech perception based on non-speech signals. *In* The Representation of Speech in the Peripheral Auditory System. R. Carlson and B. Granstrom. Eds.: 209–215. Elsevier Biomedical Press. New York, NY.