In search of the Unicorn: Where is the invariance in speech?

Steven Greenberg
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704, USA
steveng@icsi.berkeley.edu

Commentary on Sussman, H. Fruchter, D., Hilbert, J. and Sirosh, J. (1998) Linear correlates in the speech signal: the orderly output constraint. Brain and Behavioral Sciences, 24: 241-299.

Abstract

Understanding spoken language involves far more than decoding a linear sequence of phonetic elements. In view of the inherent variability of the acoustic signal in spontaneous speech, it is not entirely clear that the sort of representation derived from locus equations is sufficient to account for the robustness of spoken language understanding under real-world conditions. An alternative representation, based on the low-frequency modulation spectrum, provides a more plausible neural foundation for spoken language processing.

Classical models of speech perception presume that the essence of meaning can be distilled from a linear (or quasi-linear) sequence of linguistic elements. At the acoustic level these elements are most commonly associated with phonetic segments (or "phones"), through whose sequential association larger, more abstract units such as the syllable, word and phrase are derived. In this traditional view the phone functions as the minimal linguistic unit capable of distinguishing among lexical entities. And in turn, each phone is composed of distinctive (articulatory or acoustic) features which, when bound together, yield a specific phonetic element. Within this framework each phone is commissioned to play a specific and important role in the systematic conversion of sound into meaning. Any misstep along the way potentially jeopardizes the speech

decoding process and hence it is crucial for each phonetic segment to be accurately and faithfully represented.

The locus equations so elegantly derived by Sussman and colleagues in their target paper provide a neat, compact means by which to derive the requisite invariant representations from the underlying acoustic signal within this traditional theoretical framework. Unfortunately, it is not entirely clear that speech understanding necessarily entails such a linear decoding process or that neuronal mechanisms exist capable of extracting the feature patterns required to functionally simulate the representational equivalence effected by locus equations.

Detailed phonetic transcription of spontaneous spoken English (four hours of informal, conversational dialogs systematically sampled from the Switchboard corpus [Godfrey et al., 1992]) indicate that it is often difficult to associate much of the acoustic signal with specific phonetic symbols (Greenberg et al., 1996). Phone elements are frequently deleted or are significantly transformed during the process of spoken discourse, so that words are rarely characterizable as a linear sequence of phonetic elements. Even trained phoneticians frequently have difficulty identifying a significant proportion of speech sounds contained in the Switchboard corpus. However, with few exceptions, these conversations are perfectly understandable. Furthermore, the phonetic variability occasioned by dialectal, idiolectal and entropic factors is enormous. Many of the most common words are phonetically realized in dozens of different ways (Greenberg, 1997). Often, the most reliable cues to phonetic identify are temporal, rather than spectral in nature (Greenberg et al., 1996; Greenberg, 1997).

In addition to these speaker and linguistic sources of phonetic variability, environmental factors such as reverberation and background acoustic interference cause a significant alteration of the spectro-temporal properties of the speech signal reaching the listener's ears (Greenberg and Shire, 1997; Kingsbury et al., 1997). Thus, it is not entirely clear what sort of "invariance" should be sought in the signal given the nature of acoustic-phonetic variability commonly found in informal, spontaneous speech.

And yet it is tempting to search for some form of invariant representation given the robustness of speech under such a wide range of environmental and speaker conditions. Some property (or combination of properties) of the speech signal must be responsible for the hardiness of spoken communication. Locus equations, to the extent they are associated with specific formant trajectories in the signal, are unlikely to yield the sort of invariant representation required to account for the intelligibility of speech in the real world, as they require a relatively faithful transduction of the acoustic signal in the auditory pathway. Unfortunately, auditory neurons are unlikely to provide sufficient precision of coding (at least at the level of the auditory cortex - see Schreiner's commentary on the target paper) to accommodate the sort of neuronal processing implied by locus equations (at least in mammalian species other than bats).

A more likely means of providing a quasi-invariant representation of the speech signal is through neural computation of the low-frequency (<25 Hz) modulation spectrum. The magnitude of the modulation spectrum at any given frequency is derived from the modulation pattern of the speech waveform over a predefined bandwidth (typically 1/4 to 1-octave wide). Preservation of this modulation information, distributed across frequency channels is sufficient to encode natural sounding, intelligible speech (Dudley, 1939). The modulation transfer function of neurons in primary auditory cortex (Schreiner and Urbas, 1986) matches precisely the modulation spectrum of spontaneous speech (English - Greenberg et al., 1996; Japanese - Arai and Greenberg, 1997), as well as the temporal transfer function of the vocal apparatus during speech production (Maeda, personal communication; Smith et al., 1993). An extension of the modulation spectrum, the "modulation spectrogram" (which embeds the modulation spectral information into a spectrographic format) has been successfully used in automatic speech recognition systems to preserve linguistic features otherwise degraded by acoustic interference (Greenberg and Kingsbury, 1997; Kingsbury et al., 1997).

References

Arai, T. and Greenberg, S. (1997) The temporal properties of spoken Japanese are similar to those of English, in Proceedings of Eurospeech, Rhodes, Greece, in press.

Dudley, H. (1939) Remaking speech. Journal of the Acoustical Society of America 11: 169-177.

Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development, in ICASSP-92 IEEE International Conference on Acoustics, Speech and Signal Processing, 1, pp. 517-520.

Greenberg, S., Hollenback, J. and Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus, in Proceedings of the Fourth International Conference on Spoken Language (ICSLP), Philadelphia, S24-27.

Greenberg, S. (1997) On the origins of speech intelligibility in the real world, in Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 23-32.

Greenberg, S. and Kingsbury, B. (1997) The modulation spectrogram: In pursuit of an invariant representation of speech, in ICASSP-97 IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, pp. 1647-1650.

Greenberg, S. and Shire, M. (1997) Temporal factors in speech perception, in CSRE-based Teaching Modules for Courses in Speech and Hearing Sciences. London, Ontario: AVAAZ Innovations, pp. 91-106.

Kingsbury, B., Morgan, N. and Greenberg, S. (1997) Improving ASR performance for reverberant speech, in Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 87-90.

Schreiner, C. E. and Urbas, J. V. (1986) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hearing Research 21: 227-241.

Smith, C., Browman, C., McGowan, R. and Kay, B. (1993) Extracting dynamic parameters from speech movement data. Journal of the Acoustical Society of America 93: 1580-1588.

[Note: papers by Greenberg and colleagues available in electronic versions (PDF and PS) from: http://www.icsi.berkeley.edu/~steveng]