A syllable-centric framework for the evolution of spoken language

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

Commentary on MacNeilage, P. (1998) The Frame/Content Theory of Evolution of Speech Production, Brain and Behavioral Sciences, 21: 499-546.

Abstract

The cyclic nature of speech production, as manifested in the syllabic organization of spoken language, is likely to reflect general properties of sensori-motor integration rather than merely a phylogenetic progression from mastication, teeth chattering and lipsmacks. The temporal properties of spontaneous speech reflect the entropy of its underlying constituents and are optimized for rapid transmission and decoding of linguistic information conveyed by a complex constellation of acoustic and visual cues, suggesting that the dawn of human language may have arisen when the articulatory cycle was efficiently yoked to the temporal dynamics of sensory coding and rapid retrieval from referential memory.

The syllable is an important representational unit that has largely been neglected in models of speech perception/production and spoken-language understanding. In many ways the syllable serves as the interface between sound and meaning (Greenberg, 1996; 1997) and it is refreshing for an evolutionary perspective, such as MacNeilage's, to afford a central role to this important unit of linguistic organization. This commentary focuses on the evolutionary origins of the syllabic cycle in speech production and the importance of "information" and "time" for sculpting the contours of this modulatory activity.

MacNeilage suggests that the origins of speech production may be linked to mastication, which bears a motoric similarity to the open and closing phases of the articulatory cycle associated with syllabic elements of spoken language. Although there may, indeed, be some evolutionary relation between this non-linguistic, motoric behavior and speech, an alternative perspective, based on the temporal properties of sensori-motor function and integration, provides a potentially more comprehensive and explanatory framework with which to investigate the evolutionary conditions under which spoken language arose.

The time interval corresponding to the average length of a syllable - 165 to 200 ms (Arai and Greenberg, 1997; Greenberg et al., 1996) - is ubiquitous with respect to neurological function, corresponding to the time constant for energy integration in both audition (e.g., Eddins and Green, 1995) and vision (e.g., Regan and Tyler, 1971), as well as to the minimum response time for motoric activity (e.g., Meijers and Eijkman,1974). This interval also corresponds to the time required by many regions of the cortex to classify and evaluate sensory events (e.g., Rohrbaugh et al., 1977) and to retrieve pattern-relevant information from memory (John, 1967). The temporal properties of the articulatory cycle are likely to reflect this general sensori-motor and information-retrieval integration time constant .

The syllabic structure of spoken language is more complex and heterogeneous than MacNeilage's characterization implies. A syllable in English can assume one of ca. 15 different segmental variations with respect to consonant-vowel (CV) composition and order (Greenberg, 1997). Although the CV form favored by MacNeilage is the most common variant (34% of the phonological forms, 47.2% of the phonetically realized instances), other syllabic patterns, such as CVC (31.6%, 22.1%), VC (11.7%, 4.8%) and V (6.3%, 11.2%), occur quite frequently (ibid). Together, these four syllabic forms comprise 83.6% of the phonologically defined (and 85.3% of the phonetically realized) syllables in a corpus of spontaneous (American English) discourse (Switchboard, cf. Godfrey et al., 1992). The remaining 16.4% (14.7%) of the syllables reflect more "complex" forms containing consonant clusters at either onset, coda or both. Although these

complex syllables comprise less than a sixth of the corpus, their importance should not be underestimated. Most of these forms are associated with low-frequency, content nouns (such as "strength" [CCCVCC] or "flasks" [CCVCCC]) which provide much of the informational detail characteristic of spoken language. This heterogeneity in phonetic composition is reflected in the variability of syllabic durations. Although the mean duration of a syllable is 165-200 ms, the standard deviation of this distribution is high (ca. 100 ms, both for English and Japanese, indicating that 85% of the syllables vary in length between 100 and 300 ms in length [Greenberg et al., 1996; Arai and Greenberg, 1997]), reflecting the heterogeneous segmental composition of the syllabic elements. This variability in syllabic duration is significant for understanding the neurological bases for information coding in spoken language. Commonly occurring words (over 80% of which are monosyllabic), largely predictable from context (e.g., "function" words), tend to be pronounced in a "reduced" fashion closer to the canonical CV structure, than low-frequency, highly informative "content" words. Deviation from this canonical pattern appears to be one means of linguistically marking elements invested with unusually high entropy.

Thus, the information associated with any specific linguistic element is likely to be reflected in its duration, and therefore the temporal properties of speech production potentially provide a window onto the neurological mechanisms mediating the lower- and higher-levels of spoken language. The distribution of syllabic durations (both in English and Japanese) matches the low-frequency modulation spectrum (defined as the magnitude of energy in the speech signal low-pass filtered below 20 Hz, cf. Greenberg and Kingsbury, 1997), with a peak at ca. 5 Hz (reflecting the mean syllabic duration of 200 ms) and substantial energy distributed between 3 and 10 Hz (Greenberg et al., 1996; Arai and Greenberg, 1997). This modulation spectrum corresponds closely to the temporal transfer function of neurons in the AI region of primary auditory cortex (Schreiner and Urbas, 1988) and the pattern of vocal movements during continuous speech (Smith et al., 1993; Bouabana and Maeda, 1998).

Together, these data suggest that the temporal properties of spoken language may not merely reflect constraints imposed by the inertial characteristics of a biomechanical system descended from a phylogenetically more basic (masticatory) function, but also represent the integration of the articulatory apparatus into an intricately woven web of sensori-motor function optimized for rapid retrieval of stored information and which underlies the brain's capability of constructing a stable representation of the external world under the wide range of environmental conditions typical of the real world.

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, pp. 1011-1114.

Bouabana, S. and Maeda, S. (1998) Multipulse LPC modeling of articulatory movements, Speech Communication, 24: 227-248.

Eddins, D. A. and Green, D. M. (1995) Temporal integration and temporal resolution, in Hearing. Handbook of Perception and Cognition, 2nd ed., B. C. J. Moore (Ed.), San Diego: Academic Press, pp. 207-242.

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, pp. S24-27.

Greenberg, S. (1996) Understanding speech understanding - towards a unified theory of speech perception. Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Keele, England, pp. 1-8.

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.

John, E. R. (1967) Mechanisms of Learning and Memory. New York: Academic Press.

Meijers, L. M. and Eijkman, E. G. (1974) The motor system in simple reaction time experiments. Acta Psychologica, 38, 367-377.

Regan, D. and Tyler, C. W. (1971) Temporal summation and its limit for wavelength changes: An analog of Bloch's law for color vision. Journal of the Optical Society of America, 61, 1414-1421.

Rohrbaugh, J. W., Donchin, E. and Eriksen, C. W. (1974) Decision making and the P300 component of the cortical evoked response. Perception & Psychophysics, 15, 368-374.

Schreiner, C. E. and Urbas, J. V. (1988) 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]