Published in final edited form as:

Lang Cogn Process. 2014 January 1; 29(1): 44–45. doi:10.1080/01690965.2013.852230.

Auditory feedback control is involved at even sub-phonemic levels of speech production

Prof. Frank H. Guenther

Departments of Speech, Language, & Hearing Sciences and Biomedical Engineering, Boston University, Boston, MA 02115, guenther@bu.edu

The target article describes two neural models that link the brain's lexical representations with the neural circuitry responsible for speech motor control. These models form an important contribution to the speech and language research literature, which contains very few models that link linguistic entities with speech motor control mechanisms, and even fewer that incorporate both neuroanatomical and computational treatments. As a long-time believer in the importance of neurocomputational models for elucidating brain function, I find this work exciting and important. However, both the SFC and HSFC models are at an early stage of development, and certain aspects of the models will likely need to be modified. Because my primary expertise is in speech motor control, I will focus the remainder of this commentary on the motor control components of the more recent of the two models, the HSFC model.

At the motor control level, the HFSC model shares many traits with the DIVA model of speech production, which provides the most detailed and thoroughly tested account of the neural computations underlying speech motor control (see Tourville and Guenther, 2011 for the most recent version of the DIVA model). One of the biggest differences between the HSFC model and the DIVA model, aside from the inclusion of interactions with lexical representations in the former, concerns the placement of auditory feedback control at a higher processing level than somatosensory feedback control in the HSFC model. Specifically, auditory feedback control operates at the syllable level in the HSFC model, whereas somatosensory feedback control operates at the phoneme level.

Hickok's argument for this distinction is as follows. First, he asserts that the linguistic conception of the phoneme as a bundle of distinctive features (e.g., Chomsky and Halle, 1968) implies that some phonemes cannot be coded in auditory terms. He further asserts that it is straightforward to interpret phoneme features as somatosensory entities, leading to his claim that phonemes are coded in somatosensory rather than auditory terms.

One problem with this argument is the assumption that distinctive feature bundles adequately describe phonemes. Hickok asserts that the features defining a voiceless stop consonant such as /t/, which include [–continuant, –voiced, +coronal], involve only non-acoustic events. He brushes off the criticism of an earlier reviewer who noted that /t/ involves a burst of acoustic energy as the tongue releases its constriction, stating that [+release] is not part of the linguistic definition of /t/. Here Hickok seems to miss a fundamental point: the conception of a phoneme as a bundle of features is a convenience for the linguist interested in higher-level aspects of language, not an accurate picture of the sensorimotor processes underlying articulation of the phoneme. In Hickok's view, producing a string of phonemes is a matter of moving through a set of static phonemic targets, each defined by a bundle of distinctive features. These targets correspond to the endpoints of opening movements (for vowels) and closing movements (for consonants). The view of the phoneme as a static bundle of features has been convincingly refuted as an account of speech motor control by Fowler (1980), and clear counterexamples abound. For example,

diphthongs inherently involve time-varying formant frequency patterns and corresponding time-varying tongue positions, even when produced in isolation; they cannot be adequately represented by a static bundle of features. The production of voiced versus voiceless consonants depends on voice onset time (VOT), an obviously time-dependent phenomenon. These examples indicate that, generally speaking, phonemes are not static vocal tract configurations but instead are inherently time-varying entities. A voiceless stop consonant is more than just a closed vocal tract with no vocal fold vibration; it includes movement into and out of the closure, along with a corresponding sequence of acoustic events.

Hickok goes on to assert that phoneme features are best defined in somatosensory rather than auditory terms. Again, many counterexamples exist. One of the clearest is the American English phoneme /r/, which is produced with widely varying vocal tract gestures by the same speaker in different phonetic contexts. However, the auditory signal corresponding to /r/ remains remarkably constant across contexts despite this articulatory variability (Guenther et al., 1999). In other words, /r/ is much better characterized by its acoustic or auditory properties than its articulatory or somatosensory properties. We have hypothesized elsewhere (e.g., Ghosh et al., 2010) that all phonemes involve both auditory and somatosensory targets, but the relative importance of these targets varies by phoneme, with consonants generally relying more heavily on somatosensory targets and vowels relying more heavily on auditory targets.

Hickok concludes his case for auditory feedback control at the syllable rather than phoneme level by suggesting that the auditory system is better suited for processing sound sequences, such as syllables, than individual phonemes. However, clear evidence that auditory feedback control operates at the phoneme level and below can be found in the auditory perturbation literature. For example, Niziolek and Guenther (2013) found that speakers compensated to unexpected real-time perturbations of their first formant frequency while they were producing one-syllable words even when the perturbation did not change the perceived vowel identity, though compensation was greater for perturbations that pushed the sound across a phonemic boundary.

It thus appears that both auditory and somatosensory feedback control are involved in the lowest levels of speech production, in contrast with Hickok's proposal of a hierarchical relationship in which somatosensory feedback control acts at the phoneme level and auditory feedback control at the syllable level. In addition to operating on phonemic and sub-phonemic levels, I believe auditory and somatosensory feedback control are both involved at the syllable level as well, particularly for well-learned syllables. Although Hickok is correct that the first version of the DIVA model (Guenther, 1994, 1995) used static phonemes as the sound units of speech production, this was considered to be a simplification to make initial simulations tractable. Later versions of the model utilize time-varying auditory and somatosensory targets for speech sounds, and they do not assume that the phoneme is the primary unit of representation for speech production, instead positing the existence of syllable-sized motor programs (with associated time-varying auditory and somatosensory targets) in addition to phoneme motor programs (e.g., Guenther et al., 2006).

Despite these quibbles, I believe the SFC and HSFC models represent important initial steps at linking word representations in the brain with the motor control circuits that produce these words. I am sure these models will generate testable predictions that will lead to an increasingly refined account of the lexical-motor interface, and I look forward to seeing this work evolve in the coming years.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R01DC002852. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

References

- Chomsky, N.; Halle, M. The sound pattern of English. Harper & Row; New York: 1968.
- Fowler CA. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*. 1980; 8:113–133.
- Ghosh SS, Matthies ML, Maas E, Hanson A, Tiede M, Menard L, Guenther FH, Lane H, Perkell JS. An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America*. 2010; 128:3079–3087. [PubMed: 21110603]
- Guenther FH. A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*. 1994; 72:43–53. [PubMed: 7880914]
- Guenther FH. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*. 1995; 102:594–621. [PubMed: 7624456]
- Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, Perkell JS. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America*. 1999; 105:2854–2865. [PubMed: 10335635]
- Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*. 2006; 96:280–301. [PubMed: 16040108]
- Niziolek C, Guenther FH. Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *Journal of Neuroscience*. 2013; 33:12090–12098. [PubMed: 23864694]
- Tourville JA, Guenther FH. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*. 2011; 26:952–981. [PubMed: 23667281]