Dimension-based Attention in Learning and Understanding Spoken Language

Frederic K Dick (f.dick@bbk.ac.uk)
Birkbeck/UCL Centre for NeuroImaging, 26 Bedford Way,
London WC1H 0DS

Lori L Holt (loriholt@cmu.edu)
Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213

Howard Nusbaum (hcnusbaum@uchicago.edu)
Department of Psychology, The University of Chicago, 5848 South University Avenue
Chicago, Illinois 60637

Neeraj Sharma (neerajww@gmail.com)
Department of Electrical Communication Engineering, Indian Institute of Science,
Bangalore, 560 012

Barbara Shinn-Cunningham (shinn@cns.bu.edu)
Department of Biomedical Engineering, Boston University, 610 Commonwealth Ave.
Boston, MA 02215

Keywords: attention; learning; computational modeling; speech communication.

Background

The acoustic world is variable and messy. Signals that allow listeners to identify and/or disambiguate important auditory objects (an unusual rustle of leaves in a darkening forest, your winning bingo number) will likely share many acoustic characteristics with other sounds in the environment. What is more, the most diagnostic auditory dimensions for detecting the ominously rustling leaves or categorizing a crucial syllable may be obscured, distorted, or missing. Just as often, the diagnostic acoustic dimensions may change completely across contexts.

These challenges are ubiquitous in speech comprehension. People talk with different accents, different speaking rates, different vocal tracts, and often at the same time. A talker may turn away from you in the middle of a conversation, and then decide to whisper the most intriguing part of their sentence, so as not to be overheard by the person they are talking about. Or, worse yet, the talker may imitate the very different and distinct voice of the object of conversation. Each of these cases dramatically changes the character, usefulness, and significance of the acoustic dimensions available for speech comprehension.

This variability and context-dependence in the mapping between acoustic dimensions and the identity or significance of a sound segment means that an 'acoustical fingerprint' computational approach that has been so successful in recognizing recorded music even in degraded recordings and noisy backgrounds (Wang, 2003) may not be very successful for humans or machines trying to perceive and understand natural speech in everyday environments.

Thus, a basic problem is discovering the acoustic dimensions that allow for speech decoding. In particular, the listener must track parameters of speech-relevant acoustic dimensions, so that the information they provide for speech decoding can be appropriately weighted. This information includes 1) a dimension's *validity* (how strong is the mapping between variability across an auditory dimension, and sound identity or category?); 2) its *availability* (how often is that dimension encountered); 3) its *redundancy* (how much informational overlap is there between this dimension and others?); 4) its *salience* (to what degree do the basic acoustics of the dimension attract exogenous attention); and 5) its *cost* (how difficult is it to perceive the dimension, or extract its underlying statistics?).

Crucially, the values of these parameters vary widely as a function of acoustic and linguistic context. For example, whereas the pitch of some speech segments is a secondary cue to in quiet environments, its role is much greater in noise. Therefore, the listener must dynamically reweight the informative acoustic dimensions, taking into account the context-specific modulation of these parameters to maximize the likelihood of recovering the talker's intended message.

Viewed this way, speech and non-speech auditory recognition and categorization can be thought to involve dynamic perceptual reweighting of multiple auditory dimensions across different timescales (Francis & Nusbaum, 2002; Heald & Nusbaum, 2014; Idemaru & Holt, 2014). By extension, proficient speech comprehension can be modelled - at least in part - as becoming skilled in modulating the attentional gain to specific auditory dimensions, and doing so in a context-sensitive way. This theoretical approach takes advantage of advances in understanding the acquisition of expert visual skills, and also relies on the decades of research in both human and non-human animals that listeners adjust

their reliance on different dimensions on a moment-by-moment basis (Shamma & Fritz, 2014). It is also complementary to, and informed by, a number of theoretical models (functionalist, probabilistic, Bayesian, cue-integration) that emphasize the dynamic and probabilistic nature of language processing and speech perception (e.g., Bates & MacWhinney, 1989; Kleinschmidt & Jaeger, 2015; McClelland & Elman, 1986; McMurray & Jongman, 2011;

In this symposium, we sketch out a framework for approaching speech perception as a process of dimension-based auditory attention.

First, Lori Holt reviews the evidence for the ubiquity of context-informed perceptual reweighting in multiple speech domains, showing that listeners are not only sensitive to the validity, availability, redundancy, salience, and cost of auditory dimensions that are diagnostic for sound identity or category, but that they adjust their use of these dimensions based on multiple, dimension-specific contextual cues on a moment-to-moment basis.

Second, Howard Nusbaum discusses how the acquisition of perceptual expertise is shaped by context and cognitive mechanisms of working memory and attention. To this point, conceptually framing auditory processing as a skill rather than a native ability (as speech perception is often characterized) opens the way to investigate how feedback from the environment about success in achieving perceptual goals can work to shape attention dynamically in and for different contexts.

Third, Fred Dick reviews ongoing behavioral and neuroimaging research from our laboratories that examines both associative and causal relationships between general auditory attentional and speech comprehension skills. In particular, he focuses on what candidate mechanisms might be involved at different stages of learning and transfer between non-speech and speech, and how short- or long-term shifts in the informational parameters of a given dimension (in particular *validity* and *salience*) might drive representational change.

Fourth, Neeraj Sharma reports on how current machine perception models for speech recognition reweight acoustic dimensions, whether they use similar strategies as reported for humans, and what mechanisms or metaphors machine and human perceptual approaches might profitably borrow from each other.

Finally, Barbara Shinn-Cunningham lays out key similarities and differences in the way that object- and dimension-selective attention appear to play out in audition and vision.

References

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. *The crosslinguistic study of sentence processing*, *3*, 73-112.

Brookshire, G., Lu, J., Nusbaum, H., Goldin-Meadow, S., & Casasanto, D. (2017). Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences*, *114*, 6352-6537.

Dick, F. K., Lehet, M. I., Callaghan, M. F., Keller, T. A., Sereno, M. I., & Holt, L. L. (2017). Extensive Tonotopic Mapping across Auditory Cortex Is Recapitulated by Spectrally Directed Attention and Systematically Related to Cortical Myeloarchitecture. *Journal of Neuroscience*, *37*(50), 12187–12201.

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349.

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental Auditory Category Learning. *Journal of Experimental Psychology Human Perception and Performance*.

Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35.

Heald, S. L. M., Hedger, S. C., & Nusbaum, H. C. (2017). Understanding sound: Auditory skill acquisition. To appear in B. Ross (Ed)., *Psychology of Learning and Motivation*, 67, 53-93.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148-203.

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

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, *12*(5), 182-186.

Shinn-Cunningham, B., Best, V., & Lee, A. K. (2017). Auditory Object Formation and Selection. In *The Auditory System at the Cocktail Party* (pp. 7-40). Springer, Cham.

Shamma, S., & Fritz, J. (2014). Adaptive auditory computations. *Current opinion in neurobiology*, *25*, 164-168.

Van Hedger, S. C., Heald, S. L. M., Koch, R., & Nusbaum, H. C. (2015). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition, 140*, 95-110.

Wang, A. (2003, October). An Industrial Strength Audio Search Algorithm. In *Ismir* (Vol. 2003, pp. 7-13).