Challenges of analyzing variability in speech from linguistic and motor control perspectives

D. H. Whalen[1,2,3]

[1]*City University of New York (USA),* [2]*Haskins Laboratories (USA),* [3]*Yale University (USA)*
whalen@haskins.yale.edu

Variability is intrinsic to movement in biological systems, including speech. Although such variability has, in the past, been treated as unwanted noise, there is increasing evidence to indicate that variability has uses as well. When learning a new task, variability can lead to faster learning via "exploration" [1]. Further, lack of typical variability can be classified in extreme cases as a disorder [2]. In speech, variability is seen in virtually every measure taken, and it seems to be unavoidable as well [3]. Speakers appear to match the variability in their environment even when it does not match their intrinsic rate [4]. The fact that variability does appear to be normally distributed [5] suggests that the central tendency is indeed the target for speech sounds. Online compensation might indicate that trajectories are corrected during a syllable's production [6], but a more likely alternative is that speakers have a somatosensory indication of the accuracy of a starting position. Changes in starting position do affect intergestural timing [7], consistent with the availability of such information.

Theories of phonetics have dealt with variability in different ways. To the extent that linguistic phonologies are the beginning of a planning process, they leave any implementation of variability to a phonetic level. The Directions into Velocities of Articulators (DIVA) model has mappings between somatosensation locations and acoustic consequences of such configurations [8]. Although this model can accommodate motor equivalence for producing similar acoustic outputs, it is not at all clear how the targets are selected within the target region during production. Articulatory Phonology [9, 10] is implemented via a dynamical system, but the basic theory generates a single, determinate set of parameters for a given context, resulting in a lack of variance. Variability in phonetic measurements has been handled by adding stochastic noise to the model [11]. However, even though the pattern of results can be matched in this way, stochastic noise does not seem to allow for a differentiation between deliberate (exploratory) variability and inadvertent (true noise) variability.

An approach to useful and harmful variability that has been developed in the motor control literature is the Uncontrolled Manifold Analysis (UCM) [12, 13]. Given multiple repetitions of successful movement trajectories, it is possible to see which variants are benign (lying on the uncontrolled manifold) versus destructive (being part of the controlled manifold, i.e., the path to success). Some attempts have been made to apply this model to speech [14, 15], but the nonlinear relations between articulation and acoustics make it difficult to obtain enough tokens to train an appropriate model (note that in reaching studies, the relations were mostly linear). Further, the size of the target changes the manifold itself, and the manifold is the way in which the action achieves the target. Smaller targets lead to more constrained actions. With larger targets, there will be some correct productions that may nonetheless be considered non-ideal ("kind of a success"). The UCM says nothing about this effect. The other major issue missed by focusing only on the target is that there is information that often exists in the trajectory itself. Listeners make use of contextual variability when they "parse" the speech signal for coarticulatory effects [16]. Thus, definitions of the target and the success are both difficult to define for speech.

Where does this leave us? We need to study larger datasets than has been possible in the past. However, because even finding a true measure of the shape of the variability requires about 200 token of the same production [5], direct experimentation is challenging. Analysis of large corpora necessarily excludes a careful analysis of contextual factors (which may contribute to parsing), in addition to relying on inaccurate measures of vocal tract resonances [17]. It would therefore seem that a combination of modeling, production and perception experiments, elaboration of theories, and corpus work is necessary. The overarching issue is what the target of speech sounds is, and how much we make use of the variability intrinsic to the targets and in individual tokens reaching the targets.

References:

1. Wu, H.G., et al., *Temporal structure of motor variability is dynamically regulated and predicts motor learning ability.* Nature Neuroscience, 2014. **17**(2): p. 312-321.
2. Goldberger, A.L., *Fractal variability versus pathologic periodicity: Complexity loss and stereotypy in disease.* Perspectives in Biology and Medicine, 1997. **40**: p. 543-561.
3. Tilsen, S., *Structured nonstationarity in articulatory timing*, in *Proceedings of the 18th International Congress of Phonetic Sciences*, The Scottish Consortium for ICPhS 2015, Editor. 2015, University of Glasgow: Glasgow. p. 1-5.
4. Tang, D.-L., B. Parrell, and C.A. Niziolek, *Movement variability can be modulated in speech production.* Journal of Neurophysiology, 2022. **128**: p. 1469-1482.
5. Whalen, D.H. and W.-R. Chen, *Variability and central tendencies in speech production.* Frontiers in Communication, 2019. **4**(49): p. 1-9.
6. Niziolek, C.A., S.S. Nagarajan, and J.F. Houde, *What does motor efference copy represent? Evidence from speech production.* Journal of Neuroscience, 2013. **33**: p. 16110-16116.
7. Shaw, J.A. and W.-R. Chen, *Spatially conditioned speech timing: Evidence and implications.* Frontiers in Psychology, 2019. **10**(2726).
8. Guenther, F.H., *A neural network model of speech acquisition and motor equivalent speech production.* Biological Cybernetics, 1994. **72**: p. 43-53.
9. Browman, C.P. and L.M. Goldstein, *Towards an articulatory phonology.* Phonology Yearbook, 1986. **3**: p. 219-252.
10. Iskarous, K. and M. Pouplier, *Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology.* Journal of Phonetics, 2022. **95**(101195): p. 1-28.
11. Gafos, A.I., et al., *Stochastic time analysis of syllable-referential intervals and simplex onsets.* Journal of Phonetics, 2014. **44**: p. 152-166.
12. Latash, M.L., *Biomechanics as a window into the neural control of movement.* Journal of Human Kinetics, 2016. **52**(1): p. 7-20.
13. Scholz, J.P. and G. Schöner, *The uncontrolled manifold concept: identifying control variables for a functional task.* Experimental Brain Research, 1999. **126**: p. 289-306.
14. Kang, J., *The effect of speaking rate on vowel variability based on the uncontrolled manifold approach and flow-based invertible neural network modeling*. 2021, City University of New York.
15. Szabados, A. and P. Perrier, *Uncontrolled Manifolds in vowel production: Assessment with a biomechanical model of the tongue*, in *Proceedings of Interspeech 2016*, N. Morgan, Editor. 2016. p. 3579-3583.
16. Fowler, C.A. and M. Smith, *Speech perception as "vector analysis": An approach to the problems of segmentation and invariance*, in *Invariance and variability in speech processes*, J.S. Perkell and D.H. Klatt, Editors. 1986, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 123-136.
17. Whalen, D.H., et al., *Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986).* Journal of the Acoustical Society of America, 2022. **152**: p. 933-941.