Estimating "Good" Variability in Speech Production using Invertible Neural Networks

Jaekoo Kang[1,2], Hosung Nam[2,4] and D. H. Whalen[1,2,3]

[1]The Graduate Center, CUNY; [2]Haskins Laboratories; [3]Yale University; [4]Korea University
jkang@gradcenter.cuny.edu; dwhalen@gc.cuny.edu; hnam@korea.ac.kr

Redundancy is an important component of skilled human motor movements, allowing flexibility in execution. Computational approaches to assessing redundancy (e.g., the Uncontrolled Manifold approach or UCM; Latash, 2012; Scholz & Schöner, 1999) can allow us better insight into motor control. Redundancy can be seen mathematically as comprising the "null space." It is an algebraically defined concept; in a task space, any nonzero solutions that belong to the null space of the input (X as n-dimensional vectors) are mapped to zeros in the output (Y as m-dimensional vectors, where n is larger than m; Strang, 2009). The null space indicates redundancy in an input-to-output system. This concept has been widely studied in the field of robotics (Berenson et al., 2011; Huang et al., 2018; Siciliano, 1990), especially focusing on removing or constraining the null space of a motor system because redundancy is generally not preferred in precise robotic movements.

Computing the null space (or "good" variability) for speech is not tractable algorithmically: First, the articulation-to-acoustics mapping is often unknown and has to be accurately estimated if a null space is to be computed. Second, computing the null space of the forward mapping is impossible without a mandatory mismatch in the number of dimensions of inputs and outputs ($n > m$). Third, the null space alone does not show how variability at the acoustic output is structured although exploring variability is important to understand the articulation-to-acoustics stream (Whalen et al., 2018).

To address these issues, the current project explores a method of modeling the null space directly from data using flow-based Invertible Neural Networks (INNs; Ardizzone et al., 2018), a machine-learning technique using artificial neural networks (Bishop, 2006). Advantages of INNs for the null space modeling are as follows. First, the articulation-to-acoustics mapping can be accurately estimated using the technique of 'normalizing flow' which gradually transforms a probability density $p(X_{articulation})$ into the desired density $p(Y_{acoustics})$ forcing interpretability and invertibility of the transformation (Tabak & Turner, 2013; Tabak & Vanden-Eijnden, 2010). Second, dimensional mismatches are no longer required, as the data can be made compatible by padding zeros in the input and adding multivariate Gaussians in the output dimension as learnable latent variables. Third, using the estimated null space, variability at the acoustic output can be explained, and the exact inverse mapping becomes possible because information loss is minimized in the forward-inverse mapping, which can be further combined with the Goal Equivalent Manifold approach to computing input-output variability/sensitivity index (Cusumano & Cesari, 2006).

Articulatory and acoustic recordings of 32 speakers in the X-ray microbeam database (Westbury, 1994) are selected focusing on nine English vowels (/u, ʊ, æ, ɑ, ʌ, ɔ, e, ɪ, i/) following Whalen et al. (2018). Articulatory kinematic data are pre-processed into five major components using a principal component analysis. Corresponding formant frequencies (F1, F2, F3) are also extracted at five-equidistant time intervals. INNs are trained on the standardized and pre-processed data and compared with linear regression and simple neural networks. The results

are discussed in terms of the performance of INNs, interpretation of the null space and the structure of articulatory and acoustic variability in speech production as skilled action.

References

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., & Köthe, U. (2018). Analyzing inverse problems with invertible neural networks. *ArXiv Preprint*, 1–20.

Berenson, D., Srinivasa, S., & Kuffner, J. (2011). Task space regions: a framework for pose-constrained manipulation planning. *The International Journal of Robotics Research*, *30*(12), 1435–1460.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. (M. Jordan, J. Kleinberg, & B. Schölkopf, Eds.). New York: Springer-Verlag.

Cusumano, J. P., & Cesari, P. (2006). Body-goal variability mapping in an aiming task. *Biological Cybernetics*, *94*(5), 367–379.

Huang, Y., Silverio, J., Rozo, L., & Caldwell, D. G. (2018). Hybrid probabilistic trajectory optimization using null-space exploration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 7226–7232). IEEE.

Latash, M. L. (2012). The bliss (not the problem) of motor abundance (not redundancy). *Experimental Brain Research*, *217*(1), 1–5.

Scholz, J., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research*, *126*(3), 289–306.

Siciliano, B. (1990). Kinematic control of redundant robot manipulators: A tutorial. *Journal of Intelligent and Robotic Systems*, *3*(3), 201–212.

Strang, G. (2009). *Introduction to linear algebra* (4th Ed.). Wellesley, MA: Wellesley - Cambridge Press.

Tabak, E. G., & Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, *66*(2), 145–164.

Tabak, E. G., & Vanden-Eijnden, E. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, *8*(1), 217–233.

Westbury, J. (1994). *X-ray microbeam speech production database*. Madison, WI: Waisman Center, University of Wisconsin.

Whalen, D. H., Chen, W.-R., Tiede, M. K., & Nam, H. (2018). Variability of articulator positions and formants across nine English vowels. *Journal of Phonetics*, *68*, 1–14.