SIMPLE AGENTS ARE ABLE TO REPLICATE SPEECH SOUNDS USING 3D VOCAL TRACT MODEL

RICK JANSSEN, SCOTT R. MOISIK, DAN DEDIU

Max Planck Institute for Psycholinguistics,
Nijmegen, the Netherlands
rick.janssen@mpi.nl, scott.moisik@mpi.nl, dan.dediu@mpi.nl

Many factors have been proposed to explain why groups of people use different speech sounds in their language. These range from cultural, cognitive, environmental (e.g., Everett, et al., 2015) to anatomical (e.g., vocal tract (VT) morphology) properties. How could such anatomical factors have led to the similarities and differences in speech sound distributions between human languages (see Janssen & Dediu (in press) for a theoretical background)?

It is known that hard palate profile variation can induce different articulatory strategies in speakers (e.g., Brunner et al., 2009). That is, different hard palate profiles might induce a kind of *bias* on speech sound production, easing some types of sounds while impeding others. In a population of speakers (with a proportion of individuals) that share certain anatomical properties, even subtle VT biases might become expressed at a population-level (through e.g., *bias amplification*, Kirby et al., 2007). However, before we look into population-level effects, we first have to consider within-individual anatomical factors. For that, we have developed a computer-simulated analogue for a human speaker: an *agent.* Our agent is designed to replicate speech sounds (frequency-domain vowels) using a *production* and *cognition* module in a computationally tractable manner.

Previous agent models have often used more abstract (e.g., symbolic) signals. (e.g., Kirby et al., 2007). We have equipped our agent with a three-dimensional model of the VT (the *production* module, based on Birkholz, 2005) to which we made numerous adjustments. Specifically, we used a 4^{th}-order Bezier curve that is able to capture hard palate variation on the mid-sagittal plane. Using an evolutionary algorithm, we were able to fit the model to human hard palate MRI tracings (see http://www.mpi.nl/artivark for our data-collection project), yielding high accuracy fits and using as little as two parameters (Janssen et al., 2015). We can thus use this procedure to import palate measurements into our agent's production module to investigate the effects on acoustics. Furthermore, we also

show that our model's fits are comparable to PCA, but without the reliance on an empirical induction step when *generating* hard palates (Janssen et al., submitted). In effect, we can thus exaggerate/introduce novel biases in order to investigate their effect in the agent model.

Our agent is able to control the VT model using the *cognition* module. Previous research has focused on detailed neurocomputation (e.g., Kröger et al., 2014) that highlights e.g., neurobiological principles, speech recognition performance or time-domain acoustics. However, neither the brain nor temporal dynamics in acoustics are the focus of our current study. Furthermore, present-day computing throughput does not allow for large-scale deployment of these architectures, as required by the population model we are developing. Thus, the question whether a very simple cognition module is able to replicate sounds in a computationally tractable manner, and even generalize over novel stimuli, is one worthy of attention in its own right.

Our agent's cognition module is based on running an evolutionary algorithm on a large population of feed-forward neural networks (NNs). As such, (anatomical) bias strength can be thought of as an attractor basin area within the parameter-space the agent has to explore. The NN we used consists of a triple-layered (fully-connected), directed graph. The input layer (three neurons) receives the formants frequencies of a target-sound. The output layer (12 neurons) projects to the articulators in the production module. A hidden layer (seven neurons) enables the network to deal with nonlinear dependencies. The Euclidean distance (first three formants) between target and replication is used as fitness measure. Results show that sound replication is indeed possible, with Euclidean distance quickly approaching a close-to-zero asymptote.

Statistical analysis should reveal if the agent can also: a) Generalize: Can it replicate sounds not exposed to during learning? b) Replicate consistently: Do different, isolated agents always converge on the same sounds? c) Deal with consolidation: Can it still learn new sounds after an extended learning phase ('infancy') has been terminated? Answering these questions forms the foundation of the investigation of anatomical biases on a population level.

References

Birkholz P. (2005). *3D-Artikulatorische Sprachsynthese*. Logos Verlag, Berlin.

Brunner, J., Fuchs, S., & Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *J. Acoust. Soc. Am*., 125 (6), 3936-49.

Janssen, R., Dediu, D. (in press). Genetic biases in language: Computer models and experimental approaches in (Villavicencio, A., Poibeau, T.) *Language, Cognition and Computational Models*. Cambridge University Press.

Janssen, R., Moisik, S.R., Dediu (2015). Bézier modelling and high accuracy curve fitting to capture hard palate variation. In *Proc. of the 18th International*

Congress of Phonetic Sciences. Glasgow, UK.Janssen, R., Moisik, S.R., Dediu, D (submitted). Modelling Human Hard Palate Shape with Bézier Curves. *PLoS ONE.*

Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. U.S.A.*, 112, 1322-1327.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *PNAS*, *104*, 5241-5245.

Kröger, B.J., Kannampuzha, J., & Kaufmann, E. (2014). Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception. *EPJ Nonlinear Biomedical Physics*, 2(1), 1-8.