Deep neural networks identify sensitive regions of an acoustic tube

Relationships between speech production and acoustic outcome have long been a staple of articulatory phonetics research. Relevant historically influential theories have proposed “stable” regions of the vocal tract predicting the prevalence of vowel qualities in natural languages; and “distinctive regions”, where perturbations of some vocal tract areas are more influential on formant frequencies than perturbations of others. However, investigations targeting such relationships have to date never been performed incorporating recent advancements in big data machine learning techniques.

We designed an experiment where a computational acoustic tube model was set to randomly perturb area function increments, holding lengths of segments constant. The length of the total section was held constant at 16 cm, while the number and areas of segments were varied systematically across three experiments: (1) a four-tube model, (2) an eight-tube model, and (3) a 16-tube model. To model area function-formant relationships and appropriately model the complex, non-linear dependencies between input features (area segments of the vocal tract) and the target output (formants), we employed multi-layer perceptrons (MLPs). The network architecture consisted of two hidden layers with 64 and 32 neurons, respectively. To our knowledge, this is the first such modeling effort.

The neural networks were trained on synthetic datasets generated through a tube model of the vocal tract. The model simulated speech by varying cross-sectional areas across different segments of the tract. Each dataset consisted of thousands of data points to ensure a representative sample of possible configurations. To validate the robustness of the models, k-fold cross-validation was applied (2-fold for the first experiment, 4-fold for the subsequent ones), which helped mitigate overfitting and ensured generalizability. To address that neural networks are notoriously opaque in terms of interpretability, we used SHapley Additive exPlanations (SHAP) to assess the influence of each input segment in affecting changes to formants. Our analyses reaffirm several key assumptions about speech production and acoustics, for that opening (i.e., lips) or anterior constrictions (i.e., oral cavity) had dominant roles in shaping F3, in ways that are consistent with both lip rounding and rhoticity. In addition, taken in sum, our observed SHAP values match perfectly previously reported “sensitivity functions” for segments observed for each of F1, F2, and F3 - effectively serving as a sanity check on the appropriateness of our methodology. However, our results also highlight several often under-recognized relationships. For example, our models consistently show a stark influence of constriction on F1 and F2 in the posterior-most segment (corresponding to the glottis or larynx opening).

Our work serves the dual purpose of adding to available methods for investigating a long-standing question in phonetic sciences - “how do perturbations of an acoustic tube correspond to speech output?”; and builds on, reaffirms, and nuances earlier attempts to answer the same. Our methodology is blind to any biases possibly imposed by pre-existing theory; yet, it reiterates a basis of phonetic and phonological theory, drawn purely from acoustic theory. Relationships underpinning speech can be derived from the properties of an acoustic tube.