Deep neural networks identify sensitive regions of an acoustic tube

Runhui Song[1,2], Johan Sjons[1,3], Axel Ekström[2,3]
[1]Uppsala University, Sweden
[2]KTH Royal Institute of Technology, Sweden
[3]Stockholm University, Sweden

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/023/000683

Abstract

Tube vocal tract modelling is a staple of 20th century phonetics and speech acoustics research. Here, we apply modern data analysis methods – deep neural networks – to derive from tens of thousands possible configurations of an acoustic tube, all possible relationships between tube perturbations and formant frequencies. Here, we demonstrate the validity of the broader methodological framework, and illustrate how our deep neural network pipeline produce high fit for formant predictions made by a computer program simulating the acoustic properties of a close-to-open tube.

Keywords: tube, machine learning, vocal tract, speech production

Introduction

Relationships between speech production and acoustic outcome have long been a staple of articulatory phonetics research (Fant, 1971). Relevant historically influential theories have proposed "stable" regions of the vocal tract predicting the prevalence of vowel qualities in natural languages (Stevens, 1989; Mrayati et al., 1988); 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.

Methods

Predicting the behavior of an acoustic tube

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.

The algorithm was derived from Liljencrants and Fant (1975). Using this method, a transfer determinant is recursively computed throughout the tube sequence (i.e., list of segments). Angular frequency ω is defined as:

$$\omega = 2\pi F$$

where F denotes sound wave frequency in Hertz.

Normalized phase angle of the n^{th} segment is computed

$$\theta_n = \frac{\omega L_n}{c}$$

where c is the speed of sound at 35°C and L_n is the length of the n^{th} segment. The ratio of the area of two subsequent segments is represented as

$$k_n = \frac{A_{n+1}}{A_n}$$

and the formula for deriving the transfer determinant is

$$\begin{cases} \Delta_1 = \cos\theta_1 - \frac{\omega L_0}{c}\sin\theta_1, \\ \Delta_n = d_{n-1,n}\Delta_{n-1} - b_{n-1,n}\Delta_{n-2}, \quad \text{when } n \geq 2, \end{cases}$$

where

$$d_{n-1,n} = \cos\theta_n + k_{n-1}\cos\theta_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}},$$

$$b_{n-1,n} = k_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}}.$$

Finally, a quasi-spectral function is constructed, once the determinant for the final segment Δ_M has been computed

$$Y(F) = \cos^2(\arctan(\Delta_M))$$

In addition, an adjustment was also included specifically for transitions where a subsequent segment $L_{n-1} < 0.16 \cdot L_n$. This correction was derived from Ingård (1953) and is specified as:

$$\delta_i \simeq 0.48 \cdot \sqrt{A}(1 - 1.25\xi)$$

This correction was implemented to control for the atypical behavior of narrow-to-open segment transitions.[1]

Deep neural networks

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 modelling 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) (Shapley, 1953; Lundberg & Lee, 2017) to assess the influence of each input segment in affecting changes to formants.

Results

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).

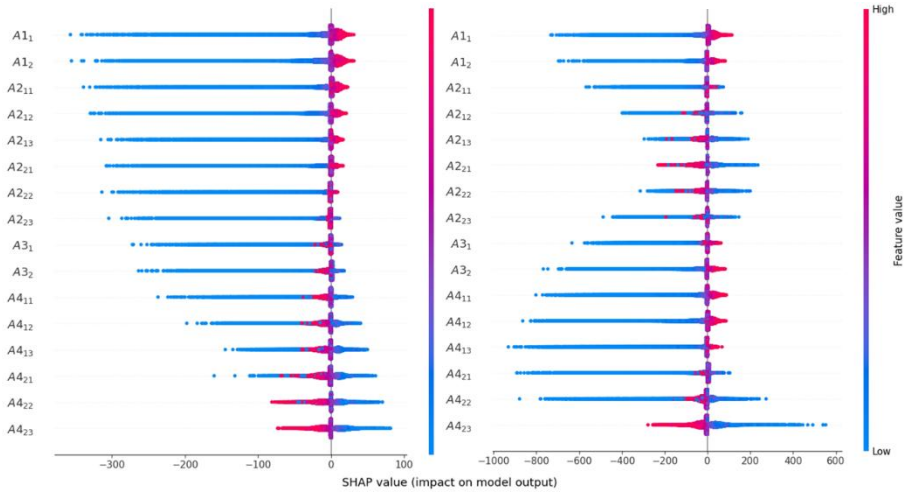

Figure 1. A "sensitive region" can be understood as one where effects of constriction vary significantly.

Discussion

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 (Fant, 1971; Carré et al., 2017). Relationships underpinning speech can be derived from the properties of an acoustic tube.

Notes

1. Details pertaining to this correction are also found in Ingard (1953, p. 1041) and Fant (1971, p. 36). The issue is noted in Liljencrants and Fant (1975) but expressly not included, because natural configurations rarely include such rapid transitions.

Acknowledgements

AE was funded through the Swedish Research Council (2025–00209\VR).

References

Carré, R., Divenyi, P., Mrayati, M. 2017. Speech: A dynamic process. De Gruyter. https://doi.org/10.1515/9781501502019

Fant, G. 1971. Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Walter de Gruyter.

Ingard, U. 1953. On the theory and design of acoustic resonators. The Journal of the Acoustical Society of America, 25(6), 1037-1061. https://doi.org/10.1121/1.1907235

Lundberg, S.M., Lee, S.I. 2017. A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), Advances in neural information processing systems, 30 (NIPS 2017).

Liljencrants, J., Fant, G. 1975. Computer program for VT-resonance frequency calculations. STL-QPSR, 16, 15-21.

Mrayati, M., Carré, R., Guérin, B. 1988. Distinctive regions and modes: a new theory of speech production. Speech Communication, 7(3), 257-286. https://doi.org/10.1016/0167-6393(88)90073-8

Shapley, L.S 1953/1997. A value for n-person games. In H. W. Kuhn (Ed.), Contributions to the theory of games. Princeton University Press, pp. 307–317.

Stevens, K.N. 1989. On the quantal nature of speech. Journal of Phonetics, 17(1-2), 3-45. https://doi.org/10.1016/S0095-4470(19)31520-7