Neural Network-assisted Analysis of Tube Vocal Tract Models

Runhui Song, Johan Sjons, Axel Ekström
Department of Linguistics and Philology, Uppsala University,
Centre for Cultural Evolution, Department of Psychology, Stockholm University
runhuisong@163.com, johan.sjons@lingfil.uu.se, axel.ekstrom.su.se

Abstract

We present a pipeline for deep neural network assisted modeling and analysis of the behavior of an acoustic tube. The vocal tract is represented as a series of cylindrical tube segments, each characterized by fixed length and variable cross-sectional area. A large synthetic dataset of such tube configurations is generated, and a circuit theory–based algorithm predicts corresponding formant frequencies. To explore mapping between vocal tract shapes and formant values, the pipeline integrates both linear regression and nonlinear machine learning models —including multilayer perceptrons. Model interpretability is measured using Shapley Additive Explanations (SHAP), which quantifies the contribution of each segment to predicted formant frequencies. The proposed framework enables detailed exploration of the articulatory-acoustic relationships inherent to an acoustic tube and vocal tract simulacrum. We present and describe the pipeline in the context of modeling effects of perturbations on the first three formants for a 16-cm tube, divided into 1 cm segments. Our pipeline can be applied to any method that models predictions of behavior of an acoustic tube, where the tube is conceived as a series of segmented units.

Keywords: speech acoustics, acoustic tube, vocal tract, deep neural network

1. Introduction

For any acoustic resonator, determined by its shape and dimensions, specific frequencies will exist where sound waves reinforce each other, vibrating more strongly. The same physical process is the basis for various phenomena in speech production, where resonance frequencies are termed formants. Essentially, the vocal tract functions as a resonator that dynamically changes shape with the movements of various articulators (lips, tongue tip, tongue body, etc.) (Dunn, 1950; Fant, 1971; Liljencrants and Fant, 1975). For this reason, the vocal tract can be represented as an elongated tube – a series of segments defined by their length and cross-sectional area. This idea is long and well established in phonetic sciences (Fant, 1971; Stevens, 1998; Carré et al., 2017) and its modern implementation goes back some 75 years (Dunn, 1950; Stevens et al., 1953). Today, computational implementations of tube acoustics make possible the investigation of complex theoretical articulatory-acoustic phenomena (Stevens, 1998; Carré et al., 2017; Zhang et al., 2024). We present a pipeline, DeepVocalTube (DVT), for the analysis of large such datasets using deep neural networks (DNNs).

It is long established in phonetic sciences that the first two resonant frequencies (or formants) correspond well to vowel quality, with contributions by the third formant (Fant, 1971; Stevens, 1998; Carré et al., 2017). As such, the single-most crucial sanity test for any large-scale modeling attempt of tube acoustics is the accurate modeling of these three resonants (hereafter F_1 F_2 F_3). Below, we describe a simplistic use case for our pipeline – exploring the relative contributions of each segment in a sequence, to predicted first, second, and third formants.

2. Methods

2.1. Simulating the behavior of an acoustic tube

In theory, any number of algorithms can be used to derive predicted formant frequencies (Henke, 1966; Mermelstein, 1973; Liljencrants and Fant, 1975; Badin and Fant, 1984). The purpose of our pipeline is not empirical investigation *per se*, but methodological innovation. In theory, our pipeline applies to any vocal tract modeling effort, where a model is constructed of segments defined by length and area. As such, in this work, we opted for the approach developed by Fant, for simulating the behavior of a lossless open-to-closed tube (Fant, 1971; Liljencrants and Fant, 1975; Zhang et al., 2024). The algorithm is both simplistic (making few assumptions), computationally efficient, and well established in the literature. Specifically, we here use the computer program presented by Liljencrants and Fant (1975), which simulates digitally behavior observed in analogue circuits (Dunn, 1950; Stevens et al., 1953; Stevens and House, 1955; Fant, 1971), with predicted formant frequencies calculated based on a transfer function from glottis to lips.

The algorithm[1] recursively computes a determinant through tube segments, the value of which is called *transfer determinant* Δ_n, reflecting the

[1] For full modeling considerations, we refer to the original publications (Fant, 1971; Liljencrants and Fant, 1975).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 654–663
11–16 May 2026.

impedance transformation up to the n_{th} tube segment. The angular frequency ω, measured in radians per second (rad/s) is defined as:

$$\omega = 2\pi F \tag{1}$$

where F represents the frequency of the sound wave measured in Hertz (Hz), which corresponds to the specific acoustic frequency being simulated. For example, if the response of the tube to a $500\ Hz$ sound wave is studied, then $F = 500\ Hz$.

The normalized phase angle of the n_{th} tube segment is:

$$\theta_n = \frac{\omega L_n}{c} \tag{2}$$

where $c = 35300$ cm/s 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 connected tube segments $(A_{n+1}\ and\ A_n)$ is represented as:

$$k_n = \frac{A_{n+1}}{A_n} \tag{3}$$

The recursive formula for the transfer determinant is:

$$\begin{cases} \Delta_1 = \cos\theta_1 - \frac{\omega L_0}{c}\sin\theta_1, \\ \Delta_n = d_{n-1,n}\Delta_{n-1} - b_{n-1,n}\Delta_{n-2}, \quad \text{when } n \geq 2, \end{cases} \tag{4}$$

where:

$$\begin{aligned} d_{n-1,n} &= \cos\theta_n + k_{n-1}\cos\theta_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}}, \\ b_{n-1,n} &= k_{n-1} \cdot \frac{\sin\theta_n}{\sin\theta_{n-1}}. \end{aligned} \tag{5}$$

After obtaining the determinant of the final tube segment Δ_M, a quasi-spectral function is constructed:

$$Y(F) = \cos^2\left(\arctan(\Delta_M)\right) \tag{6}$$

DVT also includes an internal end correction for segments where a narrow segment opens into a much wider segment. The original computational approach developed by Liljencrants and Fant (1975, p. 16) did not include such a correction, "since actual [vocal tract] configurations seldom display such discontinuities". However, because DVT models the behavior of an acoustic tube – including but *not* limited to realistic vocal tract configurations – we determined such an end correction was necessary. The correction was implemented based on Ingard (1953, p. 1041) and Fant (1971, p. 36). It determines that where the relationship between a smaller segment A_0 and a more expansive segment A is $A_0 < 0.16A$, the length of A_0 may be adjusted to compensate. This change is expressed as:

$$\delta_i \simeq 0.48 \cdot \sqrt{A}(1 - 1.25\xi) \tag{7}$$

where A denotes the cross-sectional area of the greater-area section, and $\xi = \sqrt{A_0/A}$ is a correction factor indicating the ratio of curvature to area. Note that depending on the purposes of modeling, additional internal end corrections (Dang et al., 1998; Lindblom et al., 2007) may be similarly included in data generation procedures.

2.2. Dataset Generation

It is generally agreed upon that while a small number of tube segments allow for simplistic modeling of influential factors such as opening (jaw and lips), tongue position, and pharyngeal constriction into account (Fant, 1971), additional segments allow for finer discrimination and more accurate representation of vocal tract behavior (Stevens et al., 1953; Fant, 1971; Mrayati et al., 1988; Carré et al., 2017). For example, the "distinctive regions model" developed by Mrayati and colleagues (1988) identifies eight regions necessary to capture naturalistic speech-like behavior of a tube (where each region corresponds to a selectively "sensitive" part of the tract, where changes are disruptive and selectively influential on formant output). In addition, reliability of DNNs is contingent on achieving close fit to data, even when data becomes more complex. For these reasons – to showcase both the appropriateness of the modeling procedure to basic research in speech production, and the applicability of our pipeline as benefiting future such research – we constructed a 16-segment model divided into equal-length segments of 1 cm in length (Fig. 1).[2]

For training and evaluating the multilayer perceptron (MLP), we generated 40,000 sample data, corresponding to random permutations of the 16-segment tube sequence, where each segment was varied between 0.1 cm^2 and 10 cm^2. To validate our choice of 40,000 samples, we conducted a sensitivity analysis of dataset size by using different amount of data (1,000, 5,000, 10,000, 20,000, and 40,000). A larger sample is possible (in theory, up to the total number of possible permutations), provided no computational limitations exist. However, data samples cover a wide range of tube model configurations, enabling the models to learn the relationship between tube shape and formant frequencies. Since the length of each tube segment was kept constant in the modeling, only the area of tube segments was used as an input feature in the subsequent modeling.[3]

[2] The pipeline in theory applies to any tube model, where predictions are based on parametrized segments.

[3] Note, however, that the pipeline may be employed to study effects of e.g., elongation of segments.

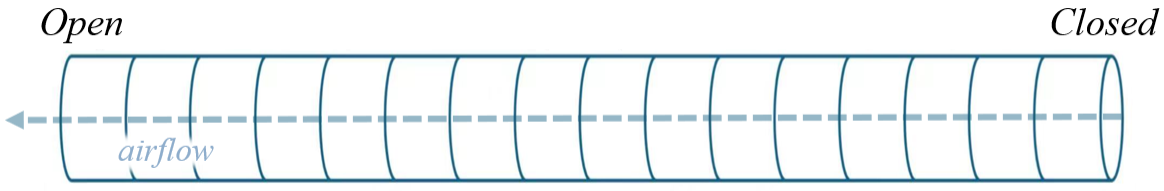

Figure 1: An open-to-closed acoustic tube, segmented into 16 equal-length segments. For comparisons with speech production, the initial "closed" section corresponds to the laryngeal opening, while the "open" segment corresponds to the termination of the oral tract, i.e., the lips.

2.3. Data Analysis Method

2.3.1. Linear Regression as Diagnostic Baseline

To probe the degree of linear dependence between tube configuration and acoustic output, we fitted a multiple linear regression model using the nth formant frequency as the outcome variable and the cross-sectional areas of tube segments as predictor variables. The purpose was to assess whether a simple linear mapping could approximate the articulatory-acoustic relationship. However, given the physical properties of acoustic resonance in non-uniform tubes, the linear model provided a poor fit to the data, with R^2 values below 0.32 across all three formants. Residual analysis further indicated substantial nonlinear structure. We did not apply nonlinear feature transformations (e.g., logarithmic scaling or polynomial expansions). Because of the clearly nonlinear relationship between tube dimensions and resonance frequencies, we proceeded to nonlinear modeling approaches.

2.3.2. Machine Learning Approaches

Considering the evidently limited ability of linear models, we used a commonly used nonlinear modeling method, the multilayer perceptron (MLP). This machine learning model is able to effectively capture the complex nonlinear relationship between tube configuration and formant frequencies, and provide a certain degree of interpretability with SHAP. Test iterations tested XGBoost (Chen and Guestrin, 2016) – an integrated learning algorithm based on Gradient Boosting (Friedman, 2001). The basic idea of Gradient Boosting is to construct a series of decision trees (Breiman et al., 2017), where each new tree is used to correct the residuals of the previous set of models, thus gradually approximating the target output. However, earlier studies showed that XGBoost was consistently outperformed by MLP (Song, 2025). In addition, XGBoost has a limitation of presenting gradient outputs in continuous space. In comparison, MLPs perform better at approximating smooth nonlinear functions, and thus show better strength in modeling the complex mapping relationship between tube segments and formant frequencies.

Here, we implemented a MLP regressor using scikit-learn. The structure of the MLP network was set as (1) an input layer containing 16 elements (length and cross-sectional areas), (2) two hidden layers containing 64 and 32 neurons respectively; (3) an activation function *ReLU*, applied in each hidden layer to introduce nonlinearity; (4) an output layer, a neuron for predicting the continuous variable N corresponding to $N \in \{1, 2, 3\}$ formants.

All input variables were standardized before model training by *StandardScaler*, a common data preprocessing technique that transforms each input feature into a distribution with a mean of 0 and a standard deviation of 1 (Goodfellow et al., 2016). This process aims to remove the differences in the numerical scales between features and ensuring that the contribution of each dimension to the loss function is on the same scale. For neural networks, which are based on gradient optimization, a large difference in the scales of the input features may lead to an imbalance in the parameter update process, which may slow down the convergence speed of the model or even cause training instability.

The model training process employed the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001 and a batch size of 200. Adam is an optimized method that combines momentum and adaptive learning rate, and is able to enhance the stability of training while improving the convergence speed. The momentum mechanism utilizes historical gradient information to accelerate parameter updates, while the adaptive learning rate automatically adjusts the learning step size according to the gradient history of different parameters. Adam is widely used in deep learning tasks for its robustness.

The loss function was mean squared error (MSE) with L2 regularization $\alpha = 1 \times 10^{-4}$, and the random seed was fixed at 42 to ensure reproducibility. Unless otherwise specified, all other parameters were kept at their default values in scikit-learn's MLPRegressor.

This study set 500 epochs for the 16-segment models. Increasing the number of epoch helped

the models to adapt to the increasing complexity of the high-dimensional inputs, and to ensure that they have sufficient learning ability. The implemented model made a trade-off between interpretability and expressiveness by adopting a shallow MLP architecture with two hidden layers. It has been shown that shallow networks can effectively model structured nonlinear relationships with moderate dimensional inputs. In contrast, deeper networks are likely to suffer from training instability and degradation of interpretability, although they may in theory be more powerful (Ba and Caruana, 2014).

Several model tuning strategies were tested, including increasing the number of neurons in the hidden layers and adding additional hidden layers, as well as adjusting various hyperparameters. However, these modifications did not lead to noticeable performance improvement, instead, they considerably increased the training time. Therefore, the original model configuration was retained for analysis.

2.3.3. K-Fold Cross-Validation

In order to increase the reliability of the evaluation of the model's performance, we employed a *k*-fold cross validation, avoiding overfitting (Fushiki, 2011), by dividing the dataset into training and test set. Specifically, the data were divided equally into k non-overlapping subsets (folds); in each iteration, the $k-1$ subsets were used for training, and the remaining one for testing, looping k times such that each subset was used as a test set once. The final model performance was then averaged over all folds of evaluation metrics. For the 16-segment experiment (40,000 samples), a five-fold cross-validation was used. In other words, 32,000 data points were used for training and the remaining 8,000 for testing. The reason for using five folds was to increase the model's ability to generalize (Arlot and Celisse, 2009), while still retaining a sufficiently large training set in each split. For the smaller datasets (i.e., in the 4- and 8-segment experiments), we used two folds for the same reason – to avoid training on too little data.

2.3.4. SHAP for Model Interpretability

Neural models are typically "black boxes", precluding insights into how predictions are made (e.g., Samek et al., 2021). As a way around this problem, and, more pointedly, to ease the interpretability of the model output, the pipeline employs the Shapley Additive Explanation (SHAP) (Lundberg and Lee, 2017) framework to uniformly measure the contribution of each input feature to the model prediction results – a method derived from game theory (Shapley, 1953). SHAP measures the average contribution of a given input feature to the overall prediction across all possible sequences of input features. SHAP treats features as "participants" and model predictions as "cooperative benefits", calculating the average contribution of each feature to the output to explain the behavior of the model.

As such, SHAP straightforwardly illustrates feature contributions to prediction. Specifically, a positive SHAP value indicates that the feature contributes to increasing the predicted value compared to the model's average output, and a negative SHAP value indicates that the feature contributes to decreasing the prediction. Here, we employed two types of SHAP visualizations, (1) the Summary Plot. which aggregates SHAP values across all samples and features, highlighting feature importance, as well as how high and low values of each feature affect the prediction; and (2) the Waterfall Plot, which illustrates a single prediction by showing how each feature changes the output from a baseline value to the final prediction. In sum, (1) allows for visual inspection of the contributions of every segment to formant output, as observed in the dataset, while (2) visualizes the contribution of segments to formant changes within a single tube model. To facilitate interpretation of the MLP "black box", we also adopted KernelExplainer (Lundberg and Lee, 2017), which approximates the marginal contributions of features by perturbing model inputs and constructing a local linear model. SHAP provides both local and global interpretability, as well as uniform and robust interpretation results for different types of models, which is particularly suitable for the needs of multi-model comparison and structural analysis (Molnar, 2020).

2.4. Evaluation Metrics

We employed a total of five common regression metrics to evaluate model performance: the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE). Each metric offers complementary data pertaining to model performance, and together they provide a coherent summary of model quality. MAE, MAPE, MSE, and RMSE show how far off the predictions are on average. R^2 shows how well the model explains the data.

2.4.1. Coefficient of Determination

R^2 is an estimate of how much of the variance in the target variable is explained by the model. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

where y_i corresponds to true values, $\hat{y}_i$ to predicted values, $\bar{y}$ to the mean of y_i, and n to the number of data points. An R^2 value ranges from 0 to 1, where 1 means perfect prediction, and 0 means the model cannot explain any of the variance in the data (only predict the average). For our purposes, a higher R^2 suggests that a model better captures the relationship between tube configurations and formant frequencies.

2.4.2. Mean Absolute Error

MAE is the mean absolute difference between the predicted values and the actual values in a dataset. It is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

The lower the MAE, the better a model fits a dataset. A MAE value of 0 means the model predicts perfectly. MAE treats all errors equally, without giving extra weight to larger ones, so it provides an average measurement of the model's performance.

2.4.3. Mean Absolute Percentage Error

MAPE is the computed average of the absolute relative errors between the predicted and true values, expressed as a percentage:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{10}$$

The smaller the value of MAPE, the more accurate a prediction is. For example, a MAPE value of 11.5% means an average difference between predicted and actual value of 11.5%.

2.4.4. Mean Square Error

MSE is the average squared difference between the predicted values and the actual values in a dataset. It is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11}$$

Unlike MAE and MAPE, MSE is more sensitive to observations that are more derived from the mean, thus making it useful for identifying poor predictions.

2.4.5. Root Mean Square Error

RMSE is the square root of MSE:

$$\text{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

and allows for the observations of model deviations from true labels

3. Results

3.1. Data Sensitivity Analysis

To determine the required data volume for the experiment, we generated eight datasets of increasing sizes: 1,000, 5,000, 10,000, 20,000, 40,000, 80,000, 100,000, and 120,000 samples. We then used these datasets to evaluate the effectiveness of MLP using F_3 – which is more sensitive to minor changes than lower-frequency resonances. The evaluation results for this experiment are presented in Table 1 below.

Dataset *N*	R^2	MAE	MAPE	MSE	RMSE
1,000	0.3262	285.02	15.39%	134478.90	366.42
5,000	0.4642	251.33	13.37%	108494.82	329.30
10,000	0.7072	184.48	9.88%	60558.82	245.64
20,000	0.8877	108.49	5.75%	22757.70	150.79
40,000	0.9360	78.08	4.13%	13016.64	113.96
80,000	0.9521	66.27	3.51%	9677.87	98.37
100,000	0.9543	64.68	3.43%	9316.56	96.46
120,000	0.9537	64.34	3.42%	9477.86	97.27

Table 1: Evaluations of model fit for different sizes of dataset.

As the dataset size increased, the performance of MLP improved, with R^2 values increasing and all error metrics decreasing. However, the performance gain showed a clear diminishing return: while the increase from 1,000 to 80,000 samples led to large stepwise improvements with each increase, additional benefits of expanding the dataset beyond 80,000 were marginal. For example, the difference in R^2 between 80,000 and 120,000 samples was less than 0.01, and some evaluation metrics results based on 120,000 samples were slightly worse than those based on 100,000 samples. The computational cost also increased sharply: generating 120,000 samples took about 220 seconds compared to only a few seconds for 1,000 samples. The SHAP analysis procedure finished within 1 minute with 1,000 samples, and required nearly an hour for 80,000. Considering both accuracy and efficiency, we chose 80,000 samples for subsequent illustration and investigation, as this size captured the main performance improvements without incurring excessive computational burden.

3.2. Formants

In the following experiment, we investigated how well the MLP model predicts (F_1, F_2, F_3) and how individual tube segments contribute to these predictions. In general, the model achieved high accuracy across all formants, confirming an ability to reliably capture parameter–formant mappings (Tab. 2). To interpret the model, we applied SHAP

to analyze the contribution of each tube segment. As an illustrative case, we considered an example tube segment constellation corresponding to a vocal tract configuration for open front unrounded vowel [a], split into 16 segments (Fig. 2).

	R^2	MAE	MAPE	MSE	RMSE
F1	0.9849	8.73	2.64%	144.19	12.00
F2	0.9768	31.37	3.02%	1894.12	42.52
F3	0.9521	66.27	3.51%	9677.87	98.37

Table 2: Evaluations of models fit for each formant.

Waterfall plots (Fig. 3) illustrate how SHAP explains the prediction for a single configuration. The horizontal axis shows frequency (Hz). The baseline value $E[f(X)]$ represents the dataset average for a given formant ($F_1 = 520$ Hz, $F_2 = 1569$ Hz, $F_3 = 2638$ Hz). Each bar shows the contribution of one segment (Ai) relative to this baseline: red bars increase the prediction, blue bars decrease it. The final value $f(X)$ at the right end represents the MLP prediction for this example ($[705, 1300, 2709]$ Hz), which closely matches the reference prediction from the example ([722, 1290, 2710]). Segments are ordered by contribution magnitude, so the most influential segments appear at the top.

The ordering of features in each plot reflects their relative importance. For example, F_1 is strongly influenced by the posterior segments A13, A15, and A16, corresponding to a constricted pharynx raising F_1 in natural speech. F_2 is decreased by constrictions around the mid-back oral cavity A11-A13, while smaller posterior areas A15-A16 increase it. F_3 shows opposing effects between adjacent segments A10-A11 and A12-A13. In contrast to the single-case waterfall plots, the beeswarm summary plots in Figure 4 show the sensitivity of each formant to tube segment variations across the entire dataset.

For F_1, front sections (corresponding to the lips and the front oral cavity) A1-A5 generally have positive contributions, indicating that widening these regions raises F_1, consistent with earlier studies (Fant, 1971; Stevens, 1998). In contrast, larger posterior areas (corresponding to pharyngeal regions) near the pharynx) A13-A16 tend to lower F_1.

For F_2, effects of segment changes are more complex: Expansion of the anteriormost sections (corresponding to lip opening) A1-A2 raise F_2. Areas corresponding to the oral cavity A3-A8 often show negative SHAP values, where expansion lowers F_2 (Ladefoged and Maddieson, 1996). Expansion in more posterior sections (roughly corresponding to the tongue root area and the front part of the pharynx) A9-A13 again show positive effects, consistent with tongue retraction and "bunching" lowering F_2 in natural speech. Areas corresponding to the lower pharynx A14-A16 have negative effects, such that expansion reduces F_2.

For F_3, several strong effects are observed. The anteriormost section (again, corresponding to the lips) A1 lowers F_3 when constricted, reflecting the typical lip-rounding effect (Ladefoged and Maddieson, 1996; Song, 2025). Segments corresponding to the back oral cavity (e.g., A6-A8) have a sustained positive effect on F_3, with A7 being most influential. Segments corresponding to the mid-pharyngeal segments A13-A14 also lower F_3 when constricted. The posteriormost segment (corresponding to glottal region) A16 shows a strong negative effect: when expanded, F_3 decreases.

4. Discussion

4.1. Limitations and future work

We have left unexplored interaction effects on formant values between segments. Future modeling endeavors may seek to quantify such nonlinear interactions, for example in the context of verifying or complementing prior theoretical arguments positing distinctively "stable" and "unstable" regions of the vocal tract (Mrayati et al., 1988; Stevens, 1989; Carré et al., 2017). Such efforts should also be coupled with model selection evaluation (e.g., AIC Akaike, 2003). Given the strive for low computational cost and already high R^2 values, another clear path forward is to include a hyperparameter search for improving the multi-level perceptron model (Bergstra and Bengio, 2012).

4.2. Conclusions

The DVT pipeline allows for leveraging tens of thousands of datapoints to answer foundational questions of theoretical phonetic sciences. Our analysis was focused on examining and describing the contributions of individual tube model segments. The software package necessary to replicate and build upon the use cases described here, or to perform any additional experiments, including illustrations, is freely available online.

5. Acknowledgements

AE was supported by the Swedish Research Council (2025-00209_VR). The results of this work will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2023–00161_VR).

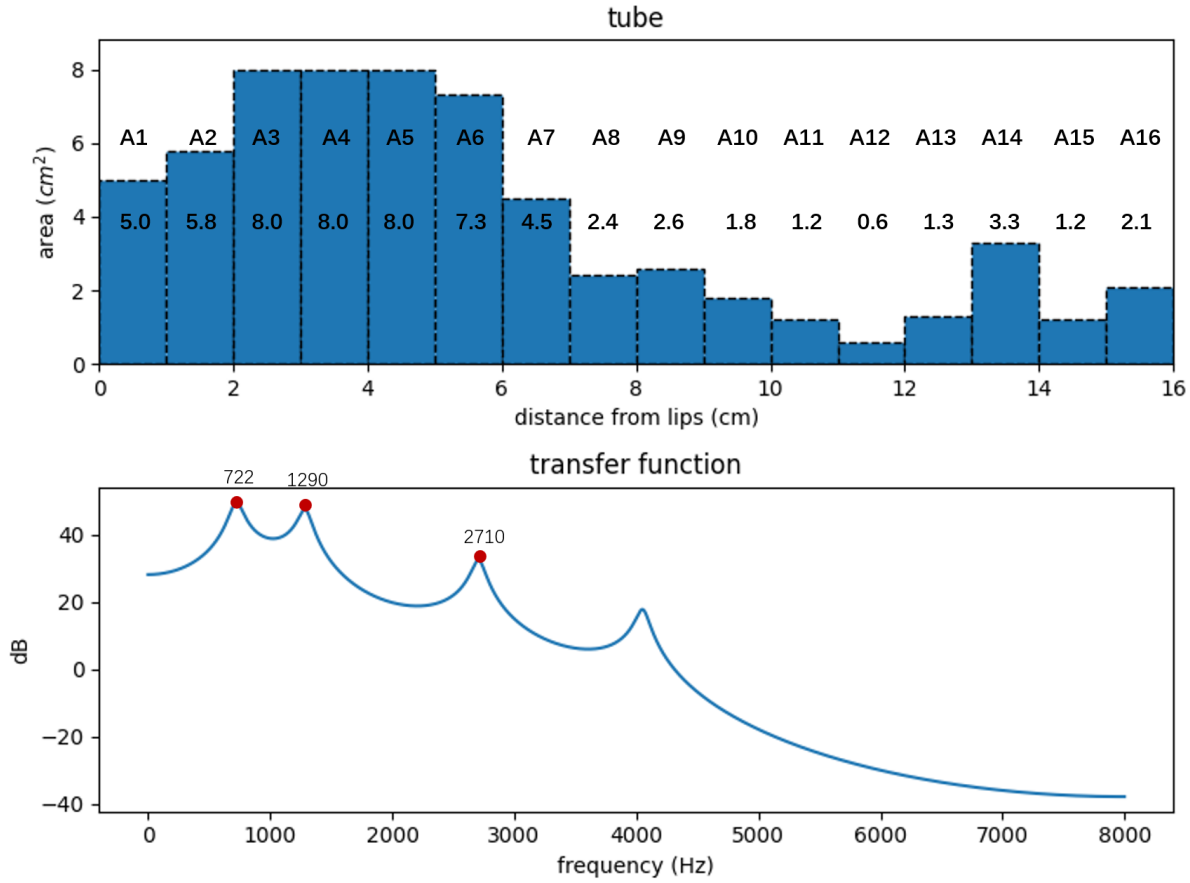

Figure 2: Tube segment parameters and predicted formant frequencies for vowel [a].

6. Bibliographical References

H. Akaike. 2003. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

S. Arlot and A. Celisse. 2009. A survey of cross validation procedures for model selection. *Statistics Surveys*, 4.

J. Ba and R. Caruana. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27.

P. Badin and G. Fant. 1984. Notes on vocal tract computation. *STL QPSR*, 2(3):53–108.

J. Bergstra and Y. Bengio. 2012. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305.

L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone. 2017. *Classification and regression trees*. Routledge.

R. Carré, P. Divenyi, and M. Mrayati. 2017. *Speech: A dynamic process*. Walter de Gruyter GmbH & Co KG.

T. Chen and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

J. Dang, C. H. Shadle, Y. Kawanishi, K. Honda, and H. Suzuki. 1998. An experimental study of the open end correction coefficient for side branches within an acoustic tube. *The Journal of the Acoustical Society of America*, 104(2):1075–1084.

P. Delattre. 1951. The physiological interpretation of sound spectrograms. *Pmla*, 66(5):864–875.

H. K Dunn. 1950. The calculation of vowel resonances, and an electrical vocal tract. *The Journal of the Acoustical Society of America*, 22(6):740–753.

G. Fant. 1971. *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations*. Mouton.

J. H. Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

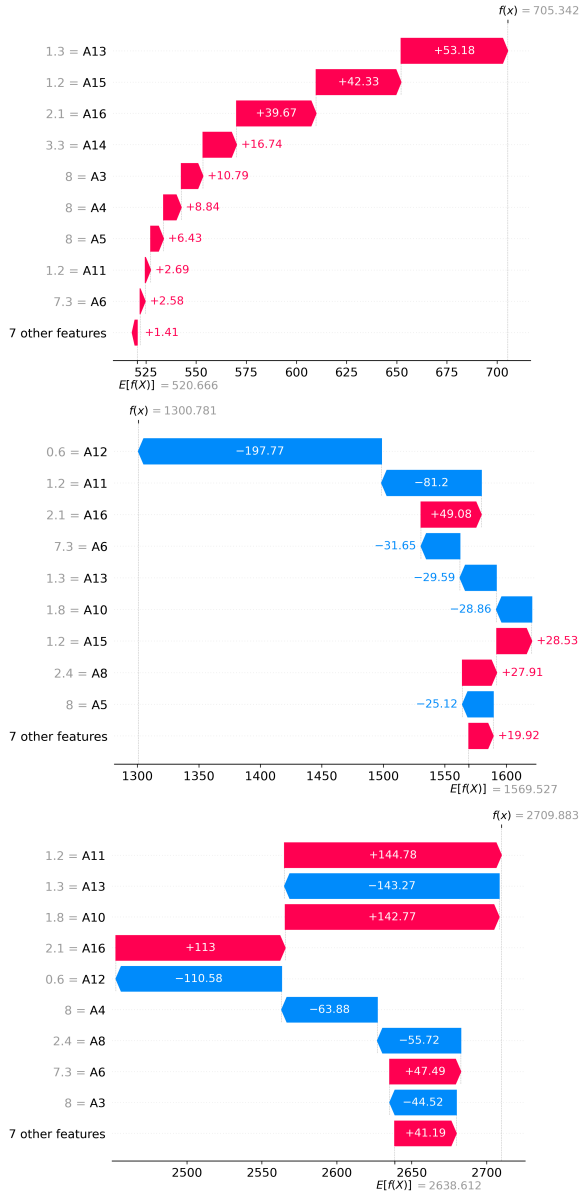

Figure 3: Waterfall plots of SHAP values for predicting F1 (top), F2 (middle), and F3 (bottom) for an example vowel [a], highlighting the contribution of each tube segment in a single example.

T. Fushiki. 2011. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21:137–146.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep learning*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

M. Halle. 1983. On distinctive features and their articulatory implementation. *Natural language and linguistic theory*, 1(1):91–105.

S. Harper, L. Goldstein, and Shrikanth Narayanan. 2020. Variability in individual constriction contributions to third formant values in american english/r/. *The Journal of the Acoustical Society of America*, 147(6):3905–3916.

W. L. Henke. 1966. *Dynamic articulatory model of speech production using computer simulation*. Ph.D. thesis, Massachusetts Institute of Technology.

U. Ingard. 1953. On the theory and design of acoustic resonators. *The Journal of the Acoustical Society of America*, 25(6):1037–1061.

IPA International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

R. Jakobson, C. G. Fant, and M. Halle. 1951. *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press.

G. James, D. Witten, T. Hastie, and R. Tibshirani. 2021. *An introduction to statistical learning : with applications in R*, second edition edition. Springer texts in statistics. Springer, New York.

K. Johnson. 2012. *Acoustic and auditory phonetics*, 3. ed. edition. Wiley-Blackwell, Malden, MA.

D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

P. Ladefoged and I. Maddieson. 1996. *The sounds of the world's languages*. Phonological theory. Blackwell, Oxford.

J. Lee, S. Shaiman, and G. Weismer. 2016. Relationship between tongue positions and formant frequencies in female speakers. *The Journal of the Acoustical Society of America*, 139(1):426–440.

J. Liljencrants and G. Fant. 1975. Computer program for vt-resonance frequency calculations. *STL-QPSR*, 16:15–21.

B. Lindblom, J. Sundberg, P. Branderud, and H. Djamshidpey. 2007. On the acoustics of spread lips. *Proceedings of Fonetik*, 50(1):13–16.

B. E. F. Lindblom and J. E. F. Sundberg. 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4B):1166–1179.

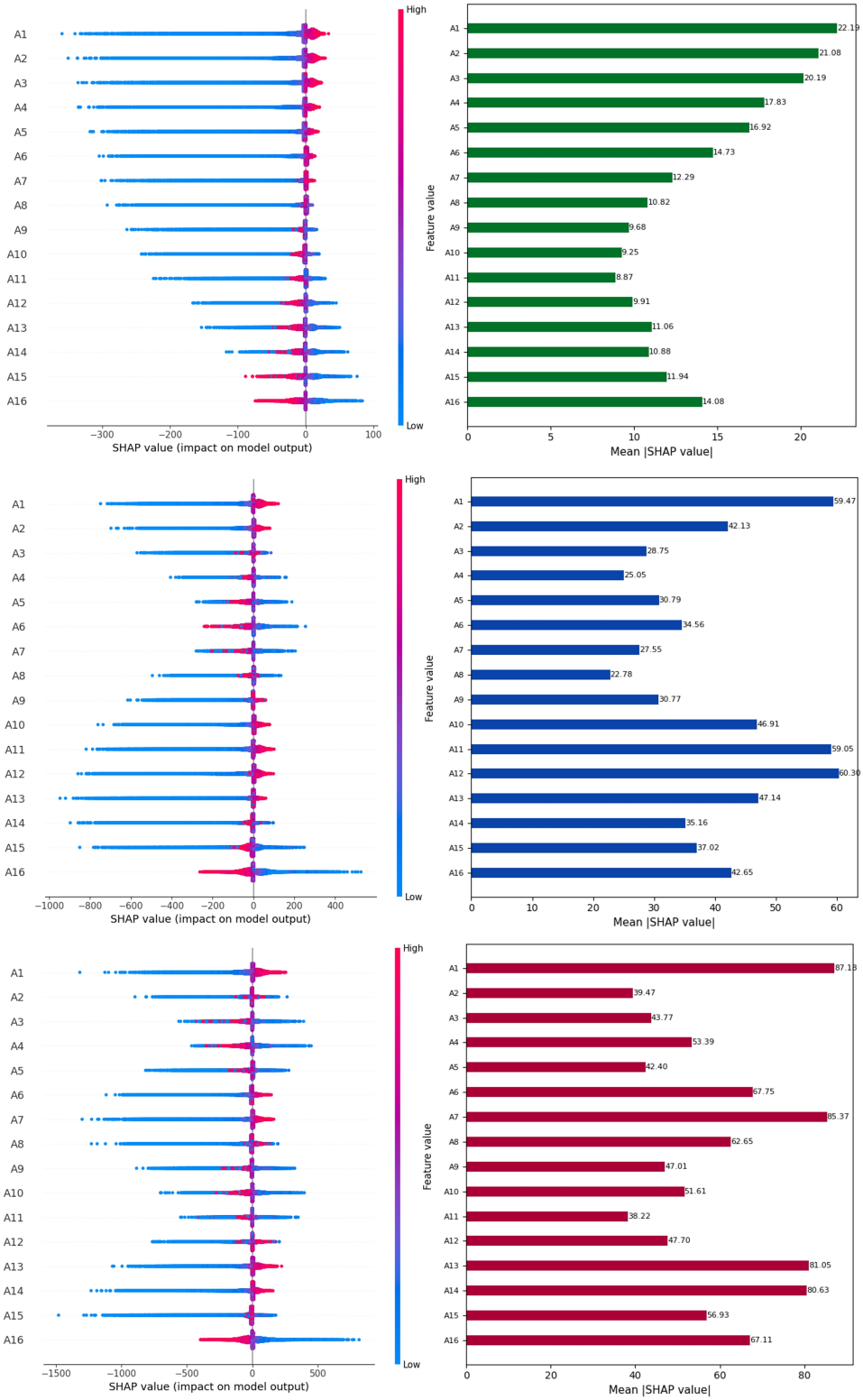

Figure 4: Beeswarm summary plots (left) and bar summary plots (right). The beeswarm summary plots show SHAP values for F1 (top), F2 (middle), and F3 (bottom), illustrating the influence of each tube segment across the entire dataset. Color coding denotes changes to formant frequencies, with blue-colored SHAP values indicating a decrease in segment area, and red-colored SHAP values indicating an increase. Negative SHAP values indicate lower formant frequencies; higher SHAP values indicate higher formant frequencies. The bar summary plots show how large the effect is of each segment by taking the mean of the absolute SHAP values.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

S. M. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

P. Mermelstein. 1973. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082.

C. Molnar. 2020. *Interpretable Machine Learning*. Leanpub.

M. Mrayati, R. Carré, and B. Guérin. 1988. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, 7(3):257–286.

G. E. Peterson and H. L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

G. O. Russell. 1928. *The vowel, some X-ray and photo larynogoperiskopik evidence; with a number of palatograms; giving thus all measurements of the vocal cavities, in three planes, from which the precise computation of each vowel's cavity tone can be made and the buccal position as it was in that subject reproduced ...* Ohio State University Press.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

L. S. Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

R. Song. 2025. Revisiting the third formant: Computational analysis of vocal tract constrictions using tube models and neural networks. Master's thesis, Uppsala University.

K. N. Stevens. 1989. On the quantal nature of speech. *Journal of Phonetics*, 17(1–2):3–45.

K. N. Stevens. 1998. *Acoustic phonetics*. Current studies in linguistics series, 30. MIT Press, Cambridge, Mass.

K. N. Stevens and A. S. House. 1955. Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America*, 27(3):484–493.

K. N. Stevens, S. Kasowski, and C. G. M. Fant. 1953. An electrical analog of the vocal tract. *The Journal of the Acoustical Society of America*, 25(4):734–742.

J. Q. Stewart. 1922. An electrical analogue of the vocal organs. *Nature*, 110(2757):311–312.

B. H. Story. 2005. Synergistic modes of vocal tract articulation for american english vowels. *The Journal of the Acoustical Society of America*, 118(6):3834–3859.

B. H. Story. 2019. History of speech synthesis. In *The Routledge handbook of phonetics*, pages 9–33. Routledge.

E. Thomas. 2017. *Sociophonetics: an introduction*. Bloomsbury Publishing.

M. Tronnier and A. G. Ekström. 2025. Teaching speech acoustics through vocal tract modeling. In *Proceedings from FONETIK*.

K. Zhang, R. Song, R. Tu, J. Edlund, J. Beskow, and A. G. Ekström. 2024. Modeling, synthesis and 3d printing of tube vocal tract models with a codeless graphical user interface. In *Proceedings from FONETIK*, pages 155–160.

X. Zhou, C. Y. Espy-Wilson, M. Tiede, and S. Boyce. 2007. An articulatory and acoustic study of "retroflex" and "bunched" american english rhotic sound based on mri. In *INTERSPEECH*, pages 54–57.