Title

Voice Feature Selection to Improve Performance of Machine Learning Models for Voice Production Inversion

Permalink
https://escholarship.org/uc/item/32r5k91f

Journal
Journal of Voice, 37(4)

ISSN
0892-1997

Author
Zhang, Zhaoyan

Publication Date
2023-07-01

DOI
10.1016/j.jvoice.2021.03.004

Peer reviewed

Voice feature selection to improve performance of machine learning models for voice production inversion

Zhaoyan Zhang[a]

Department of Head and Neck Surgery, University of California, Los Angeles, 31-24 Rehabilitation Center, 1000 Veteran Ave., Los Angeles, CA 90095-1794

Summary:

Objective.—Estimation of physiological control parameters of the vocal system from the produced voice outcome has important applications in clinical management of voice disorders. Previously we developed a simulation-based neural network for estimation of vocal fold geometry, mechanical properties, and subglottal pressure from voice outcome features that characterize the acoustics of the produced voice. The goals of this study are to 1) explore the possibility of improving the estimation accuracy of physiological control parameters by including voice outcome features characterizing vocal fold vibration; and 2) identify voice feature sets that optimize both estimation accuracy and robustness to measurement noise.

Methods.—Feedforward neural networks are trained to solve the inversion problem of estimating the physiological control parameters of a three-dimensional body-cover vocal fold model from different sets of voice outcome features that characterize the simulated voice acoustics, glottal flow, and vocal fold vibration. A sensitivity analysis is then performed to evaluate the contribution of individual voice features to the overall performance of the neural networks in estimating the physiologic control parameters.

Results and conclusions.—While including voice outcome features characterizing vocal fold vibration increases estimation accuracy, it also reduces the network’s robustness to measurement noise, due to high sensitivity of network performance to voice outcome features measuring the absolute amplitudes of the glottal flow and area waveforms, which are also difficult to measure accurately in practical applications. By excluding such glottal flow-based features and replacing glottal area-based features by their normalized counterparts, we are able to significantly improve both estimation accuracy and robustness to noise. We further show that similar estimation accuracy and robustness can be achieved with an even smaller set of voice outcome features by excluding features of small sensitivity.

Keywords

Voice inversion; vocal fold geometry; vocal fold stiffness; machine learning

[a] zyzhang@ucla.edu .

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Voice inversion in this study refers to the problem of inferring physiological control parameters of the vocal system (e.g., vocal fold geometry and stiffness and subglottal pressure) from voice production outcomes (e.g., acoustics, aerodynamics, and/or vocal fold vibration). A physics-based voice inversion system has important clinical applications. Currently diagnosis of voice disorders in the clinic relies heavily on the physician's experience in perceptual evaluation of the voice as well as other relevant medical information. Although acoustic and aerodynamic tests and endoscopic imaging of vocal fold vibration are routinely performed, interpretation of these test results in the clinic is largely qualitative. Using these test results, a voice inversion system would provide more quantitative information on the underlying physiological control parameters of the vocal system that clinicians can use, together with other relevant information of the patient, to better diagnose voice disorders. Such a voice inversion system may also provide a near-real time feedback of the speaker's voice production strategy, and thus may find applications in monitoring the progress of voice therapy and voice training.

Unlike speech inversion research in which both articulatory and acoustic data can be directly measured in live humans [1], it is very difficult to directly measure vocal fold properties and the subglottal pressure in live humans. As a result, current voice inversion research often has to rely on computational models to establish the relation between vocal fold properties and voice production output [2–9]. In our recent study [9], we used data generated from a three-dimensional continuum model of voice production to train neural networks to predict vocal fold properties (output of neural networks) from voice outcome features characterizing the voice acoustics and glottal flow (input to neural networks). Compared with lumped element models (e.g., [10, 11]), this three-dimensional model [12, 13] is parameterized by realistic geometric and mechanical properties that are often manipulated in the clinic, thus a step closer toward clinical applications. The trained neural network was shown to be able to predict physiological control parameters with reasonable accuracy, particularly for the subglottal pressure, vocal fold length, and vocal fold vertical thickness. Preliminary results showed that the neural network was able to qualitatively predict changes in the subglottal pressure in an excised larynx experiment, indicating the translational potential of simulation-based neural networks toward clinical applications.

Toward clinical applications, it is important that the performance of the neural network is robust to input noise, which includes measurement noise of the input voice features as well as modeling inaccuracy in describing the physics and physiology of human voice production. The results in our previous study [9] showed that the estimation accuracy of the neural network decreased with increasing noise in the inputs to the neural network, more so for some physiological controls than others. The goal of this study is to understand why the estimation of some physiological controls is more robust to noise than others in order to further improve the overall robustness of the network performance to noise. In addition, in [9], we intentionally chose voice features that can be estimated from the voice acoustics and glottal flow, toward applications outside the clinic. However, it is possible that the estimation accuracy of the neural network can be further improved by including voice features characterizing the vocal fold vibratory pattern, e.g., features extracted from endoscopic recordings of vocal fold vibration from clinic examinations.

Specifically, this study aims to find a voice feature set that optimizes both estimation accuracy and robustness to measurement noise in input voice features through a sensitivity analysis. The sensitivity quantifies changes in the estimation error (differences between estimated and true physiological control parameters) with increasing Gaussian noise added to the input voice features, simulating measurement noise or modeling inaccuracy. A small sensitivity indicates a small contribution to the overall voice inversion accuracy of the corresponding voice feature, and we hypothesize that this voice feature can then be dropped from the neural network without much decrease in voice inversion performance. A very large sensitivity indicates a large contribution to the voice inversion performance by the corresponding voice feature. However, a very large sensitivity also means that the voice inversion accuracy will decrease rapidly with measurement noise, particularly if accurate measurement of the corresponding voice feature is difficult (e.g., features based on the glottal flow through inverse filtering). Ideally, the network should consist of voice features with comparable, moderate sensitivity to balance estimation accuracy and robustness to noise.

In this study, due to the lack of human data, this sensitivity analysis is performed using the dataset from numerical simulations as in [9]. However, it is reasonable to assume the general findings should be able to translate to human data and thus be useful in developing similar machine learning models based on human data when they become available.

Dataset and methods

Dataset and voice features

The same dataset as in [9] is used in this study. The dataset consists of different voice outcome data, including voice acoustics, glottal flow waveform, vocal fold vibration, and the corresponding vocal fold properties that produced the voice. The data were generated from parametric simulations using the three-dimensional body-cover vocal fold model developed in [12, 13]. This model has been shown to be able to produce different voice types (regular, subharmonics, chaotic) and voice qualities (breathy, modal, pressed, vocal fry, or strained voices) observed in human voice [12, 13, 14], and has been qualitatively validated against experiment [15–17]. Simulations were performed with parametric variations in nine physiological control parameters (Figure 1), including vocal fold length L, vocal fold medial surface vertical thickness T, vocal fold depths in the medial-lateral direction of the body and cover layers D_b and D_c, vocal fold transverse stiffness E_t, vocal fold longitudinal shear moduli in the body and cover layers G_{apb} and G_{apc}, initial glottal angle or degree of vocal fold approximation α, and subglottal pressure P_s. A detailed list of the parametric conditions can be found in [9]. In total the dataset includes 95,028 phonating voice conditions.

For each voice condition in the dataset, voice features are extracted from the voice acoustics, glottal flow waveform, and vocal fold vibration. The voice features are grouped in four sets, based on 1) the specific voice outcome they are designed to characterize (e.g., acoustics, flow, or vocal fold vibratory pattern) and 2) the requirement for special equipment or calibration procedures in data collection (Table 1). The first set VFa includes features that can be extracted from the voice acoustics, either directly or through inverse-filtering, and do not require proper amplitude calibration so that they can be measured without special equipment other than a microphone. As we have no prior knowledge as to which voice features might improve the estimation performance, we include acoustics-related features that are known to be perceptually importance. These include the fundamental frequency F0, the amplitude differences between the first harmonic and the second harmonic (H1–H2), the fourth harmonic (H1–H4), the harmonic nearest 2 kHz (H1–H2 k), and the harmonic nearest 5 kHz (H1–H5 k) in the spectrum of the time derivative of the glottal flow waveform [18], cepstral peak prominence (CPP), harmonic-to-noise ratio (HNR), and subharmonic to harmonic ratio (SHR; [19]).

The second set VFf consists of voice features derived mostly from the glottal flow waveform as well as acoustics-related voice features that require proper microphone calibration, including the closed quotient (CQ), sound pressure level (SPL), perturbations of the peak amplitude (AmpPert) and period (PeriodPert) of the glottal flow waveform, maximum flow declination rate (MFDR), maximum flow acceleration rate (MFAR), mean glottal flow rate (Qmean), and peak-to-peak amplitude of the glottal flow waveform (Qamp).

The third feature set VFv consists of features derived from vocal fold vibration, including the mean glottal area (Ag0), peak-to-peak amplitude of the glottal area waveform (Agtamp), minimum glottal area (Agmin), vertical phase difference in vocal fold vibration between the upper and lower margins of the vocal fold medial surface (VPD), longitudinal phase difference in vocal fold vibration between the anterior quarter and mid-membranous locations (LPD). Finally, the last feature set VFvn is similar to VFv, except that the three glottal area measures are normalized by the vocal fold length squared. For details of the voice feature extraction process, the reader is referred to our previous studies [9, 12, 13].

The first two voice feature sets (VFa and VFf) are the voice features used in our previous study [9], whereas VFv and VFvn are newly added in this study. The two vibration phase measures, VPD and LPD, are added in order to improve estimation accuracy of vocal fold stiffness, which was relatively low in [9].

Neural network training

The voice features and the corresponding nine physiological control parameters are first z-score normalized and then used in the training of feedforward neural networks. During training, the dataset is randomly divided into three sets, each for training (70%, 66,520 conditions), validation (15%, 14,254 conditions), and testing (15%, 14,254 conditions). The neural network consists of an input layer (voice features), an output layer (estimated physiological control parameters), and a number of hidden layers of interconnected neurons in between (Figure 2). Each neuron receives inputs from the preceding layer, transforms them using an activation function, and passes them as inputs to the next layer. The goal of the training process is to find parameters of the activation functions that minimize the difference between the network prediction and target output (truth) in the training data. In this study, the neural network is trained using the scaled conjugate gradient method using the Matlab Deep Learning Toolbox. We have explored networks of different number of hidden layers, and it is found that networks with four hidden layers with 200 neurons in each layer provide reasonable performance accuracy. It is possible that the estimation performance can be further improved with more hidden layers or neurons in each hidden layer, which will be explored in future studies. The results reported below are obtained using this network configuration.

Similar to [9], the performance of trained neural networks is evaluated on the testing dataset by calculating the mean absolute error (MAE) between the truth and the estimates from the network. The trends are qualitatively similar when evaluated using root mean squared errors. To evaluate the robustness of the estimation performance to measurement noise, Gaussian noise with a standard deviation equivalent to 2% and 5% of the standard deviation of the corresponding voice feature in the entire dataset is added to the testing data, and the resulting MAEs are calculated.

Sensitivity analysis

A sensitivity analysis is performed to better understand the contribution of individual voice features to the overall network performance [20]. Using the same noise-adding procedure as described above in the last section but instead of adding noise to all voice features, we add noise to one voice feature at a time while keeping other voice features the same, and evaluate its effect on the MAEs for the nine physiological control parameters. Specifically, for each voice feature-physiological control pair, the increase in MAE will be calculated as,

$$\Delta MAE = MAE \quad (5\%\ noise) - MAE(\ no\ noise) \tag{1}.$$

A larger ΔMAE value indicates that estimation accuracy of the corresponding physiological control is more sensitive to noise/errors in the corresponding voice feature.

Results and Discussion

Figure 3 compares the MAEs of neural networks trained using different voice feature sets. The top panels show the MAEs for individual physiological controls when noise is added to all voice features, whereas the bottom panels show the MAEs when noise is added to individual voice features one at a time in the sensitivity analysis. Note that the MAEs are z-score normalized. In general, the MAE decreases as more voice features are included in the network. However, the network also becomes less robust to noise with increasing number of voice features. The combined voice feature set VFa+VFf (middle column), which was the voice feature set used in [9], provides a reasonable balance between estimation accuracy and robustness to noise. In contrast, while the combined use of all three voice feature sets (VFa+VFf+VFv) gives the lowest MAEs in the absence of noise, the MAEs increase significantly with additive noise (right column, figure 3; note the different vertical scales across different columns). The sensitivity data in the bottom panels of figure 3 show that this increase in MAEs is largely due to high sensitivity in MAEs to the voice features quantifying the absolute amplitudes of the glottal flow waveform (Qmean and Qamp) and glottal area waveform (Ag0 and Agtamp).

Note that the MAE results in the middle column of figure 3 are similar to those in [9], despite the differences in the neural network configuration and training algorithm. This indicates that the general trends of MAEs in figure 3 are related to the dataset rather than the specific neural network configuration or training method.

Such high sensitivity to the absolute amplitudes of the glottal flow and area waveforms is undesirable for practical applications in which accurate measurement of either glottal flow or area waveform is difficult. Measurement of the glottal flow is often achieved through inverse filtering (e.g., [21]), which is likely to introduce errors in the absolute amplitude of the glottal flow waveform. While the glottal area waveform can be extracted from recordings of vocal fold vibration, conversion of the glottal area from pixels to real units is difficult due to lack of proper calibration as well as errors associated with varying imaging angle and limited spatiotemporal resolutions [3, 22–23].

To improve robustness to noise, we explore the possibility of excluding Qmean and Qamp and/or using normalized glottal area measures (VFvn instead of VFv) in the neural network. The results are shown in Figure 4, for different combinations of voice features. Both the MAEs and robustness to noise are improved in all three designs shown. Although the ΔMAEs associated with the normalized glottal-area voice features still dominate those of other voice features, they are much smaller (less than 0.02) than those of the absolute glottal area measures or the two flow measures Qmean and Qamp (about 0.2–0.3) in Figure 3. The inclusion of normalized glottal area-based voice features (VFvn) also reduces the sensitivity of MAEs to Qmean and Qamp (middle column, Figure 4).

Figure 4 also shows that in addition to the normalized glottal area-based voice features, the MAE is also sensitive to changes in F0, SPL, MFDR, and LPD. In particular, the estimated vocal fold stiffness shows large sensitivity to LPD, which is partially responsible for the improved estimation accuracy of vocal fold stiffness compared with that in our previous study [9]. On the other hand, ΔMAE is much smaller for most voice features in VFa (except F0), which includes the four spectral shape measures, CPP, and HNR. This small sensitivity indicates that these voice features can be excluded without much degradation in estimation accuracy of the neural network. This is confirmed in Figure 5, which shows comparable performance of the neural network with these voice features excluded.

Table 2 shows the MAEs in real units with and without additive noise, obtained using VFa+VSf+VFvn and excluding Qmean and Qamp (Fig. 4, right column). Compared with the results in [9], the estimation accuracy improves by more than 25%, except for the body-layer longitudinal stiffness and vocal fold depth, which are improved by about 7–10%. The improvement at conditions with 5% additive noise is even higher.

Conclusions

In this study we show that although inclusion of voice features characterizing vocal fold vibration improves the neural network’s accuracy in estimating physiological control parameters, it also makes it susceptible to measurement noise, with MAEs increasing significantly with additive noise. Sensitivity analysis shows that this large increase in MAEs is mostly due to sensitivity of the neural network to voice features quantifying the absolute amplitudes of the glottal flow and glottal area waveforms (Qmean, Qamp, Ag0, Agtamp, Agmin), particularly when all five features are included in the neural network. By excluding Qmean and Qamp and replacing the glottal area features with their normalized counterparts (Ag0N, AgtampN, AgminN), the neural network is able to improve both estimation accuracy and robustness to noise.

We further show that similar estimation accuracy and robustness can be achieved with an even smaller set of voice features, by excluding voice features with negligible sensitivity which include the spectral shape measures, CPP, and HNR. It is unclear why these voice features have a small contribution to the estimation accuracy of the neural network, despite their perceptual importance reported in the literature. It is possible that voice features with large sensitivity and thus large contribution to network estimation accuracy (e.g., glottal area, MFDR) have a strong and global relationship with the physiological controls to be estimated [12, 13] so that this relationship is easily captured in the training process, whereas the relationship between voice features of small sensitivity (e.g., spectral shape) and vocal fold properties is more complex, localized, and nonlinear and presumably difficult to be learned during training.

The results of this study indicate that machine learning models of voice inversion and the dataset used should be carefully designed to allow efficient learning of both the global, simple relationships and the more subtle, localized, yet perceptually important relationships in the dataset. The selected voice features should have comparable, moderate sensitivity over the dataset to balance estimation accuracy and robustness to noise. Our study shows that this can be achieved by excluding amplitude-related voice outcome features that have a global and simple relationship with the physiological controls [12, 13] or using their normalized counterparts. It is possible that increasing the complexity of the neural network (more layers and neurons, different activation functions, etc.) may allow the neural network to better learn the localized and nonlinear relationships in voice production, which may further improve the estimation performance of the neural network.

Overall our study presents a range of network options with varying accuracy and robustness to noise that can be selected depending on specific applications. When only acoustic data are available, the use of VFa and VFf provides an acceptable option, but suffers from moderate robustness. In the clinic, when endoscopic recordings of vocal fold vibration are available, the use of VFa+VFf+VFvn and excluding (Qmean, Qamp) provides better accuracy and improved robustness to noise.

An important step toward clinical applications is to validate the findings of this study in humans. Our preliminary study showed a reasonable estimation accuracy of the subglottal pressure when compared to excised larynx experiments. While measurement of vocal fold stiffness is difficult in humans, the subglottal pressure and vocal fold geometry (length, thickness, and glottal width) may be measured with reasonable accuracy in humans. Future work will focus on systematic validation of the neural network against human data across a large range of voice conditions, and its usefulness as a clinical tool in monitoring the trends of changes of voice production overtime or during clinical intervention.

Acknowledgements

The work was supported by NIH grant R01DC009229.

References

1. Mitra V, Nam H, Espy-Wilson CY, Saltzman E, Goldstein L. Retrieving tract variables from acoustics: a comparison of different machine learning strategies. IEEE journal of selected topics in signal processing, 2010;4:1027–1045. [PubMed: 23326297]
2. Dollinger M, Hoppe U, Hettlich F, Lohscheller J, Schuberth S, Eysholdt U. Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE Transactions on Biomedical Engineering, 2002;49:773–781. [PubMed: 12148815]
3. Tao C, Zhang Y, Jiang J. Extracting physiologically relevant parameters of vocal folds from high-speed video image series. IEEE Transactions on Biomedical Engineering, 2007;54:794–801. [PubMed: 17518275]
4. Qin X, Wang S, Wan M. Improving reliability and accuracy of vibration parameters of vocal folds based on high-speed video and electroglottography. IEEE Transactions on Biomedical Engineering, 2009;56:1744–1754. [PubMed: 19272979]
5. Hadwin P, Galindo G, Daun K, Zanartu M, Erath B, Cataldo E, Peterson S. Non-stationary Bayesian estimation of parameters from a body cover model of the vocal folds. The Journal of the Acoustical Society of America, 2016;139:2683–2696. [PubMed: 27250162]
6. Gomez P, Schützenberger A, Kniesburges S, Bohr C, Dollinger M. Physical parameter estimation from porcine ex vivo vocal fold dynamics in an inverse problem framework. Biomechanics and Modeling in Mechanobiology, 2018;17:777–792. [PubMed: 29230589]
7. Hadwin P, Motie-Shirazi M, Erath B, Peterson S. Bayesian inference of vocal fold material properties from glottal area waveforms using a 2D finite element model. Applied Science, 2019;9:2735.
8. Gomez P, Schutzenberger A, Semmler M, Dollinger M. Laryngeal pressure estimation with a recurrent neural network. IEEE Journal of Translational Engineering in Health and Medicine, 2019;7:2000111.
9. Zhang Z. Estimation of vocal fold physiology from voice acoustics using machine learning. The Journal of the Acoustical Society of America, 2020;147:EL264–EL270. [PubMed: 32237804]
10. Ishizaka K, Flanagan J. Synthesis of voiced sounds from a twomass model of the vocal cords. Bell Labs Technical Journal, 1972;51:1233–1268.
11. Story BH, Titze IR. Voice simulation with a body-cover model of the vocal folds. The Journal of the Acoustical Society of America, 1995;97:1249–1260. [PubMed: 7876446]
12. Zhang Z. Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. The Journal of the Acoustical Society of America, 2016;139:1493–1507. [PubMed: 27106298]
13. Zhang Z. Effect of vocal fold stiffness on voice production in a three-dimensional body-cover phonation model. The Journal of the Acoustical Society of America, 2017;142:2311–2321. [PubMed: 29092586]
14. Zhang Z. Vocal instabilities in a three-dimensional body-cover phonation model. The Journal of the Acoustical Society of America, 2018;144:1216–1230. [PubMed: 30424612]
15. Zhang Z, Mongeau L, Frankel S. Experimental verification of the quasi-steady approximation for aerodynamic sound generation by pulsating jets in tubes. The Journal of the Acoustical Society of America, 2002;112:1652–1663. [PubMed: 12398470]
16. Zhang Z, Luu T. Asymmetric vibration in a two-layer vocal fold model with left-right stiffness asymmetry: Experiment and simulation. The Journal of the Acoustical Society of America, 2012;132:1626–1635. [PubMed: 22978891]
17. Farahani M, Zhang Z. Experimental validation of a three-dimensional reduced-order continuum model of phonation. The Journal of the Acoustical Society of America, 2016;140:EL172–EL177. [PubMed: 27586776]

18. Kreiman J, Gerratt B, Garellek M, Samlan R, Zhang Z. Toward a unified theory of voice production and perception. Loquens, 2014;1:e009. [PubMed: 27135054]
19. Sun X. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, 2002;I-333–I-336.
20. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 2003;160:249–264.
21. Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech communication, 1992;11:109–118.
22. Deng J, Hadwin P, Peterson S. The effect of high-speed videoendoscopy configuration on reduced-order model parameter estimates by Bayesian inference. The Journal of the Acoustical Society of America, 2019;146:1492–1502. [PubMed: 31472542]
23. Schlegel P, Kunduk M, Stingl M, Semmler M, Döllinger M, Bohr C, Schützenberger A. Influence of spatial camera resolution in high-speed videoendoscopy on laryngeal parameters. PLOS ONE, 2019;14:e0215168.

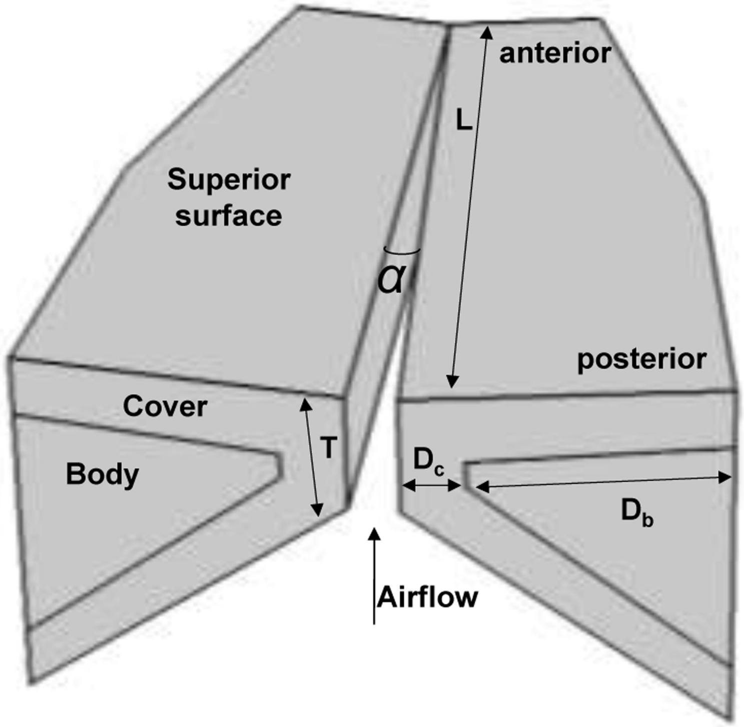

Figure 1:
The body-cover vocal fold model used to generate the dataset of this study.

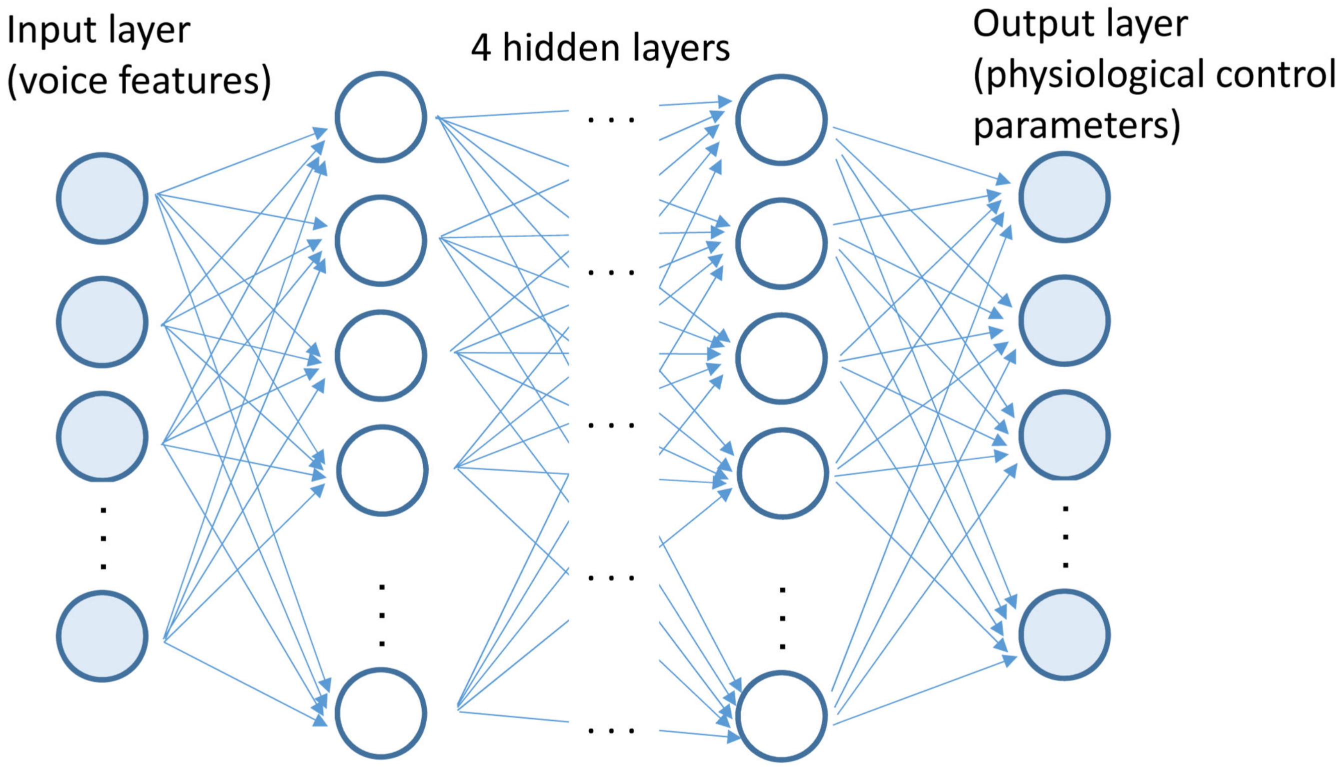

Figure 2:

Schematic of the feedforward neural network. The neural network consists of an input layer of voice features, an output layer of vocal fold properties (geometry and mechanical properties) and subglottal pressure that produce the voice features, and four hidden layers that are trained to capture the input-output relationship.

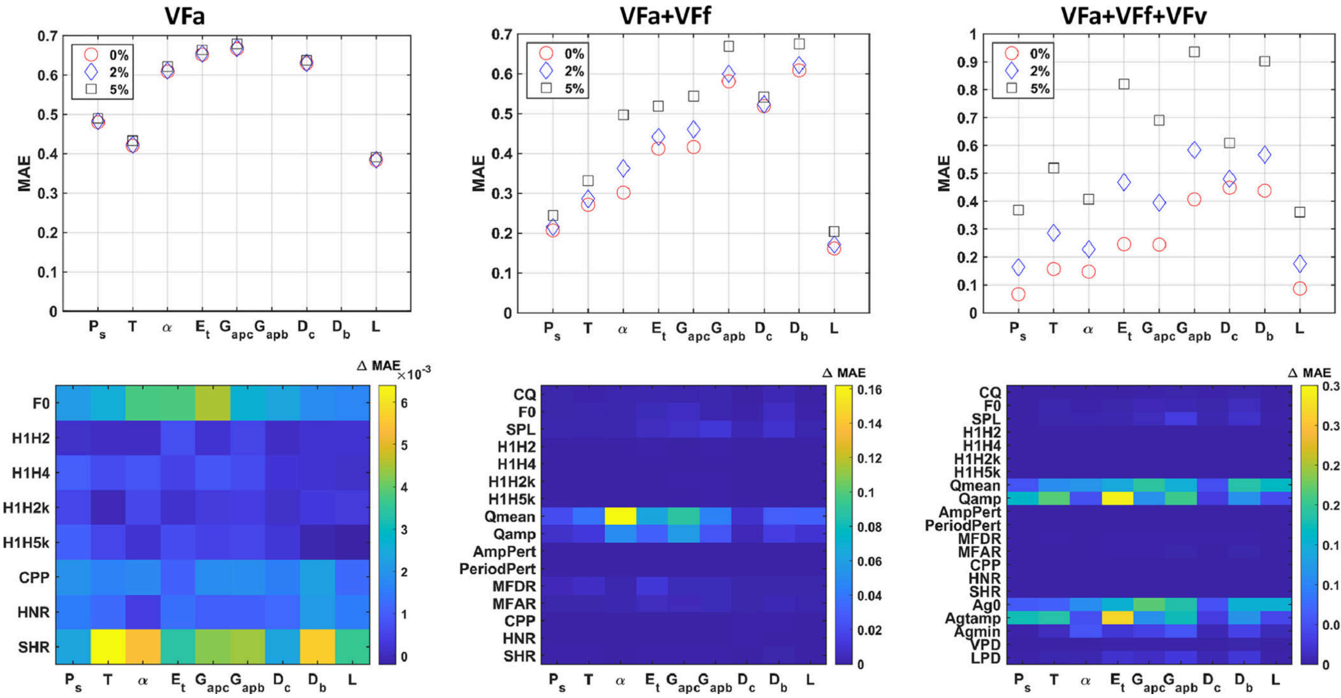

Figure 3:

Performance with different voice features sets. Top: MAEs without noise (circles) and with 2% (diamonds) and 5% (squares) additive noise to all voice features. Bottom: changes in MAEs due to addition of 5% noise to individual voice features.

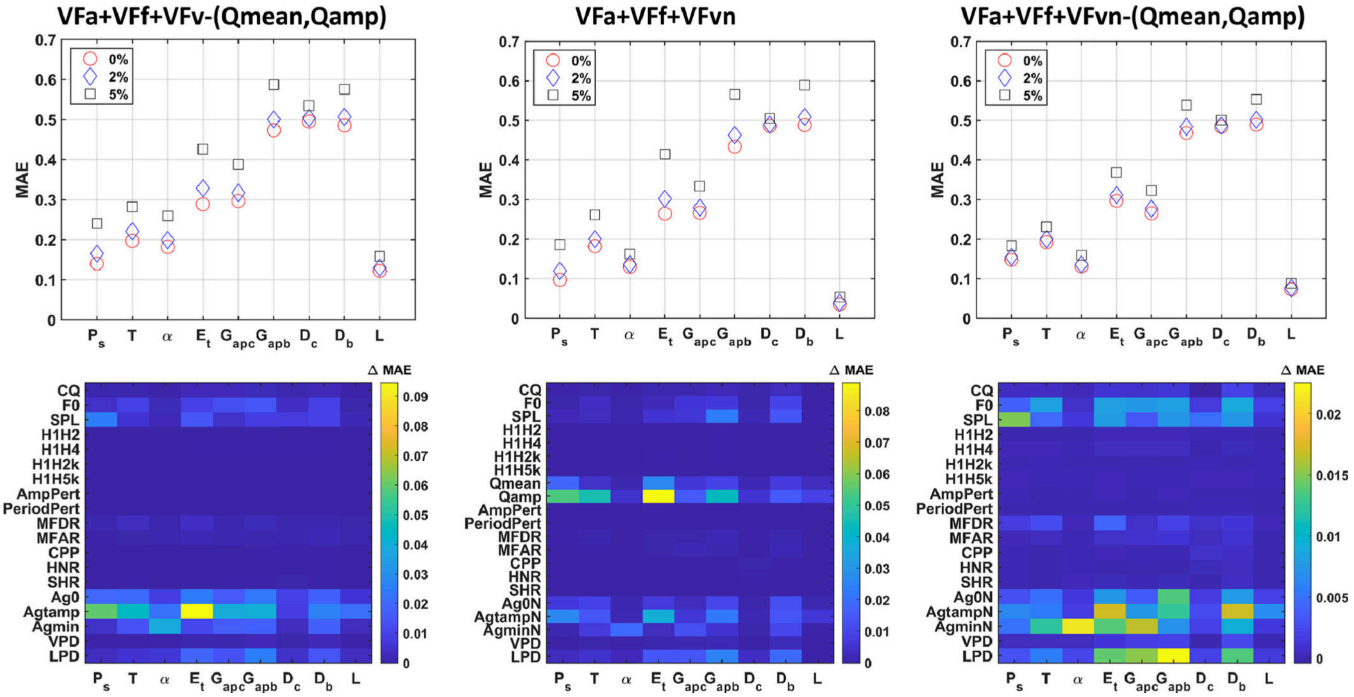

Figure 4:

Excluding amplitude-related voice features improves robustness to noise. Top: MAEs without noise (circles) and with 2% (diamonds) and 5% (squares) additive noise to all voice features. Bottom: changes in MAEs due to addition of 5% noise to individual voice features.

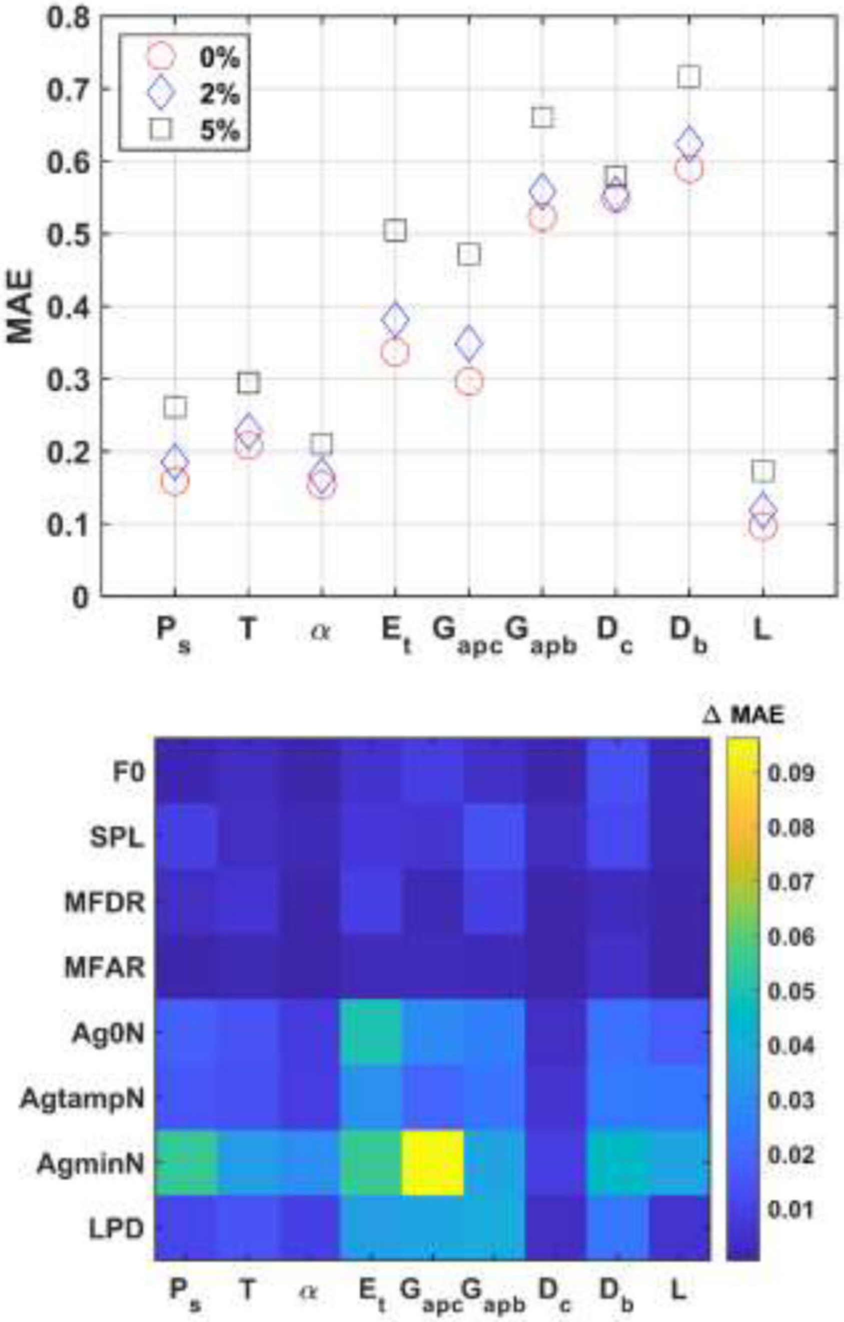

Figure 5:

Performance with voice features of small sensitivity excluded. Top: MAEs without noise (circles) and with 2% (diamonds) and 5% (squares) additive noise to all voice features. Bottom: changes in MAEs due to addition of 5% noise to individual voice features.

Table 1:

Voice features used in network training.

Feature sets	Voice features
VFa (acoustics)	F0, H1-H2, H1-H4, H1-H2K, H1-H5K, CPP, HNR, SHR
VFf (flow)	CQ, SPL, AmpPert, PeriodPert, MFDR, MFAR, Qmean, Qamp
VFv (vibration)	Ag0, Agtamp, Agmin, VPD, LPD
VFvn (vibration normalized)	Ag0N, AgtampN, AgminN, VPD, LPD

Table 2:

MAEs in real unit for neural network trained using VFa+VFf+VFvn-(Qmean, Qamp).

Vocal fold properties	MAEs (0%/5% noise)	Improvement over ref. [9]
P_s	98.6/122.8 Pa	28.2%/23.9%
T	0.23/0.28 mm	27.4%/26.6%
α	0.21/0.25°	58.7%/72.9%
E_t	0.36/0.45 kPa	26.3%/26.5%
G_{apc}	3.52/4.31 kPa	33.9%/40.1%
G_{apb}	6.24/7.21 kPa	13.1%/15.5%
D_c	0.11/0.11 mm	7.1%/9.6%
D_b	0.70/0.80 mm	10.9%/13.8%
L	0.35/0.41 mm	60.0%/63.4%