Estimation of vocal fold physiology from voice acoustics using an artificial neural network

Zhaoyan Zhang*

Department of Head and Neck Surgery, UCLA School of Medicine, Los Angeles, CA, USA

Keywords: voice inversion; machine learning; voice simulation

Introduction

Many speech applications require estimating vocal fold properties from the produced acoustics. While there have been some previous research solving the inverse problem in voice production, they are often based on lumped-element models of phonation, whose model parameters are difficult to relate to realistic vocal fold properties. This study explores the feasibility of using machine learning methods to infer physiologically realistic vocal fold properties, including vocal fold length, thickness, depth, transverse and longitudinal vocal fold stiffness, vocal fold approximation, and the subglottal pressure, from the produced acoustics.

Methods

The data used in this study are from voice simulations in previous studies using a three-dimensional body-cover vocal fold model [1, 2]. It includes a total of 162,000 voice conditions, each simulates a half-second long voice production under different vocal fold geometric and mechanical properties and the subglottal pressure. For each condition, voice features are extracted from the glottal flow waveform and the output acoustics. Sixteen features are selected for this study, including fundamental frequency, sound pressure level, spectral measures of the voice spectra, and measures of the glottal flow waveform. The nine physiological measures to be estimated include the resting glottal angle, vocal fold vertical thickness, length, depths of the body and cover layers, transverse vocal fold stiffness, longitudinal stiffness of vocal fold cover and body, and subglottal pressure.

The data are randomly divided into three sets, each for training, validation, and testing, respectively. The training set was used to train a feedforward neural network, with different hidden layer configurations.

Results and Discussion

The results show that the subglottal pressure, vocal fold vertical thickness, and vocal fold length can be estimated with very low mean absolute errors. On the other hand, vocal fold stiffness, particularly stiffness in the body layer, consistently have large estimation errors. The performance for estimation of the initial glottal angle is somewhere in between.

This estimation performance is consistent with the findings in our previous parametric studies [1, 2]. These previous studies showed that the subglottal pressure, vocal fold vertical thickness, and length, which have better estimation performance in the present study, have consistent global effect on voice production. In contrast, the effect of vocal fold stiffness, particular the longitudinal stiffness, on voice production is less consistent, which may explain the poor estimation performance for these parameters in the present study.

Acknowledgements

This study was supported by NIH/NIDCD.

References

[1] Zhang, Z., J Acoust Soc Am, 139: 1493-1507, 2016.
[2] Zhang, Z., J Acoust Soc Am, 142: 2311-2321, 2017.

*corresponding author e-mail: zyzhang@ucla.edu