JASA ARTICLE

Principal dimensions of voice production and their role in vocal expression

Zhaoyan Zhang[a)]
Department of Head and Neck Surgery, University of California, Los Angeles, 31-24 Rehab Center, 1000 Veteran Avenue, Los Angeles, California 90095-1794, USA

ABSTRACT:
How we produce and perceive voice is constrained by laryngeal physiology and biomechanics. Such constraints may present themselves as principal dimensions in the voice outcome space that are shared among speakers. This study attempts to identify such principal dimensions in the voice outcome space and the underlying laryngeal control mechanisms in a three-dimensional computational model of voice production. A large-scale voice simulation was performed with parametric variations in vocal fold geometry and stiffness, glottal gap, vocal tract shape, and subglottal pressure. Principal component analysis was applied to data combining both the physiological control parameters and voice outcome measures. The results showed three dominant dimensions accounting for at least 50% of the total variance. The first two dimensions describe respiratory-laryngeal coordination in controlling the energy balance between low- and high-frequency harmonics in the produced voice, and the third dimension describes control of the fundamental frequency. The dominance of these three dimensions suggests that voice changes along these principal dimensions are likely to be more consistently produced and perceived by most speakers than other voice changes, and thus are more likely to have emerged during evolution and be used to convey important personal information, such as emotion and larynx size. © *2024 Acoustical Society of America*. https://doi.org/10.1121/10.0027913

(Received 8 May 2024; revised 20 June 2024; accepted 24 June 2024; published online 9 July 2024)
[Editor: James F. Lynch] Pages: 278–283

I. INTRODUCTION

In addition to linguistic meaning, human voice also conveys personal information about the speaker such as emotion, size, and identity. Many factors can influence the expression of such personal information in the voice. One such factor is the physiology of voice production in the larynx, which may constrain the voice acoustics space to a few principal dimensions, despite the inherent variability in voice both within and between speakers. The existence of such principal dimensions would explain why voice changes along some dimensions (e.g., from breathy to pressed voice) can be produced and perceived consistently by most speakers (Kreiman, 2024). Such principal dimensions would also allow physiological changes aligned with these dimensions (e.g., emotion) to be consistently and reliably encoded in and perceived from the voice.

The goal of this study is to investigate whether human voice and its control are indeed constrained to a few principal dimensions. One approach to answer this question would be collecting voice samples from a large number of speakers. However, this approach does not allow us to isolate constraints due to biological factors (e.g., anatomy and physiology) from other sociocultural factors (e.g., language and culture background), nor does it provide insights into the physiological mechanisms controlling individual principal dimensions. In this study, a computational simulation approach was used. Voice samples were generated from parametric voice simulations using a previously developed three-dimensional continuum model of voice production (Zhang, 2015, 2016a, 2023a), in which vocal fold properties (geometry, stiffness, and position), vocal tract shape, and the subglottal pressure were systematically varied in a large range. The ranges of variation encompassed those reported in humans, thus sampling voice variations both within and between speakers. It will be shown below that a few principal dimensions indeed emerge naturally in this large set of voice samples, and may have played an important role in the evolution of how we encode and decode physiological information in the voice.

II. METHOD

The three-dimensional vocal fold model has been developed in our previous studies (Zhang, 2015, 2016a, 2023a). The model consists of a respiratory system, a three-dimensional vocal fold model, and a vocal tract. The vocal folds are modeled as a transversely isotropic, nearly incompressible, linear material with the plane of isotropy perpendicular to the anterior-posterior direction. A reduced-order formulation is used by projecting the governing equations of the vocal folds into the space spanned by the *in vacuo* eigenmodes of the vocal folds (Zhang, 2015), which significantly improves the computational efficiency and makes it possible to synthesize the large number of voice samples used in this study. The glottal flow is modeled as a

a)Email: zyzhang@ucla.edu

https://doi.org/10.1121/10.0027913

one-dimensional quasi-steady glottal flow model taking into consideration viscous loss up to the point of flow separation from the vocal fold surface, with the flow separation point predicted by an *ad hoc* geometric model. Vocal fold contact is modeled by applying a penalty pressure perpendicular to the vocal fold surface when the two vocal folds are in contact. Despite simplifications made to improve computational efficiency, our model has been shown to qualitatively and quantitatively reproduce observations from experiments and fully resolved simulations [see discussion in Zhang (2023b)].

Two sets of voice data were generated in this study. The first set includes voice data generated from simulation without a vocal tract, which allows identifying principal dimensions originating from laryngeal mechanisms alone. Voice simulations were performed with parametric variation in model control parameters. These include vocal fold geometry (anterior-posterior length L, medial surface vertical thickness T, and medial-lateral depths of the body and cover layers D_b and D_c), vocal fold stiffness (transverse stiffness in the coronal plane E_t, and longitudinal stiffness in the body and cover layers G_{apb} and G_{apc}), initial (prephonatory) glottal angle α, a measure of vocal fold approximation in the horizontal plane, and the subglottal pressure P_s. The ranges of variations are listed in Table I. For each condition, a half-second long sustained phonation was simulated, for a total of 221 400 conditions. Note that not all combinations of vocal fold length and depth were simulated. Specifically, considering that shorter vocal folds often have smaller depths, conditions combining a vocal fold length of 6 or 10 mm with the largest value of the body and cover layer depths were not simulated. This led to a positive relation between vocal fold length and depths in the PCA analysis, as discussed further below.

In order to investigate possible effects of vocal tract adjustments and source-filter interaction on the principal dimensions, a second set of data were generated from voice simulations with a vocal tract (Fig. 1). To reduce the number of simulation conditions, a small subset of the vocal fold conditions listed in Table I was simulated, with the number of parametric values reduced for vocal fold depths and vocal fold AP shear moduli. These parameters have been shown to have relatively small impact on the voice source in previous studies [e.g., Zhang (2021), (2023b)]. For each vocal fold condition, simulations were performed first with a 17.5-cm long uniform vocal tract, then with vocal tract shapes with constrictions varying in degree introduced at the levels of the false vocal folds, aryepiglottic folds, pharynx, oral cavity, and the lips, respectively, similar to Zhang (2023a), for a total of 144 000 conditions.

TABLE I. Simulation conditions.

Transverse Young's modulus	$E_t = [1, 2, 4]$ kPa
Cover AP shear modulus	$G_{apc} = [1, 10, 20, 30, 40]$ kPa
Body AP shear modulus	$G_{apb} = [1, 10, 20, 30, 40]$ kPa
Vertical thickness	$T = [1, 2, 3, 4.5]$ mm
Cover layer depth	$D_c = [1, 1.5]$ mm
Body layer depth	$D_b = [4, 6, 8]$ mm
Vocal fold length	$L = [6, 10, 17]$ mm
Initial glottal angle	$\alpha = [-1.6°, 0°, 1.6°, 4°, 8°]$
Subglottal pressure	$P_s = 50$–2400 Pa (18 steps)
Vocal tract minimum constriction	[0.2, 0.4, 1, 2] cm^2

For each simulation condition, selected voice outcome measures were extracted. These include voice source measures of perceptual importance, as identified in the psychoacoustic model of voice quality proposed by Kreiman *et al.* (2021), as well as other measures often included in studies of vocal expression of emotion [e.g., Patel *et al.* (2011) and Sundberg *et al.* (2024)]. For voice acoustics, these include the fundamental frequency (f0), sound pressure level (SPL), cepstral peak prominence (CPP), harmonic to noise ratio (HNR), and subharmonic to harmonic ratio (SHR). From the glottal flow, the differences between the first harmonic and the second harmonic (H1-H2), the fourth harmonic (H1-H4), the harmonic nearest 2 kHz (H1-H2k), and the harmonic nearest 5 kHz (H1-H5k) in the spectrum of the time derivative of the glottal flow waveform were extracted. The mean (Q0) and peak-to-peak amplitude (Qamp) of the glottal flow waveform, closed quotient (CQ) of the glottal flow waveform, maximum flow declination rate (MFDR), and normalized MFDRN (MFDR normalized by $\pi \cdot$f0$\cdot$Qamp) were calculated. The mean (Ag0) and peak-to-peak amplitude (Agamp) of the glottal area waveform were also extracted. The glottal resistance (GR) was calculated as the ratio of the subglottal pressure and the mean glottal flow. The peak vocal fold contact pressure (P_c) over vocal fold surface during vocal fold collision was also extracted as a measure of risk of vocal fold injury (Zhang, 2023a).

Principal component analysis (PCA) was then performed to identify potential low-dimensional patterns in the voice outcome space. The analysis was then repeated using the voice outcome data combined with the corresponding model control parameters, in order to identify physiological control mechanisms of the potential principal dimensions.

III. RESULTS

The PCA analysis revealed similar patterns, at least for the first three dominant PCA modes, whether it was applied to the voice outcome data alone or the data combining voice outcome measures and control parameters. Thus, the following focuses on results from the PCA analysis applied to the combined data of voice outcome measures and model controls, for simulations without a vocal tract. This produced a total of 27 PCA modes. The result is shown in Fig. 2. The left panel of Fig. 2 shows the percentage variances explained by individual PCA modes. Two dominant modes can be observed, with the first and second PCA modes explaining about 28% and 15% of the total variance, respectively. The third PCA mode also captured a significant percentage (about 9%) of the total variance. The percentage of variance explained decreased slowly for the remaining 24 modes. The next three panels in Fig. 2 show the contribution of individual model controls and voice outcome measures

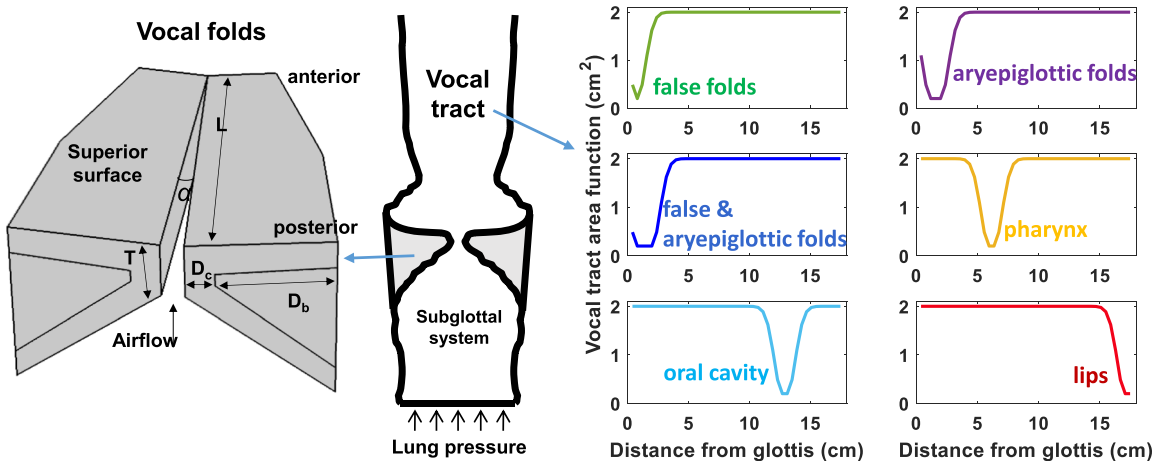

FIG. 1. (Color online) Computational model of human voice production used in this study with vocal fold controls and different vocal tract shapes.

to the first three PCA modes. In the following, we focus on outcome measures with a load equal to or larger than 0.2.

The first mode describes coordination between the respiratory and laryngeal sub-systems in a way that facilitates air passage through the larynx, and thus is likely to play an important role in breathing in order to get as much air in and out of the lungs as possible. This is achieved by laryngeal adjustments to reduce the glottal flow resistance when increasing the subglottal pressure P_s. Such laryngeal adjustments include reducing vocal fold adduction (increasing glottal gap α and reducing vocal fold vertical thickness T) and/or increasing vocal fold length L. When applied to phonation, these adjustments increase vocal intensity SPL at the expense of increased airflow consumption Q_0 and reduced glottal closure (a reduced closed quotient CQ). As a

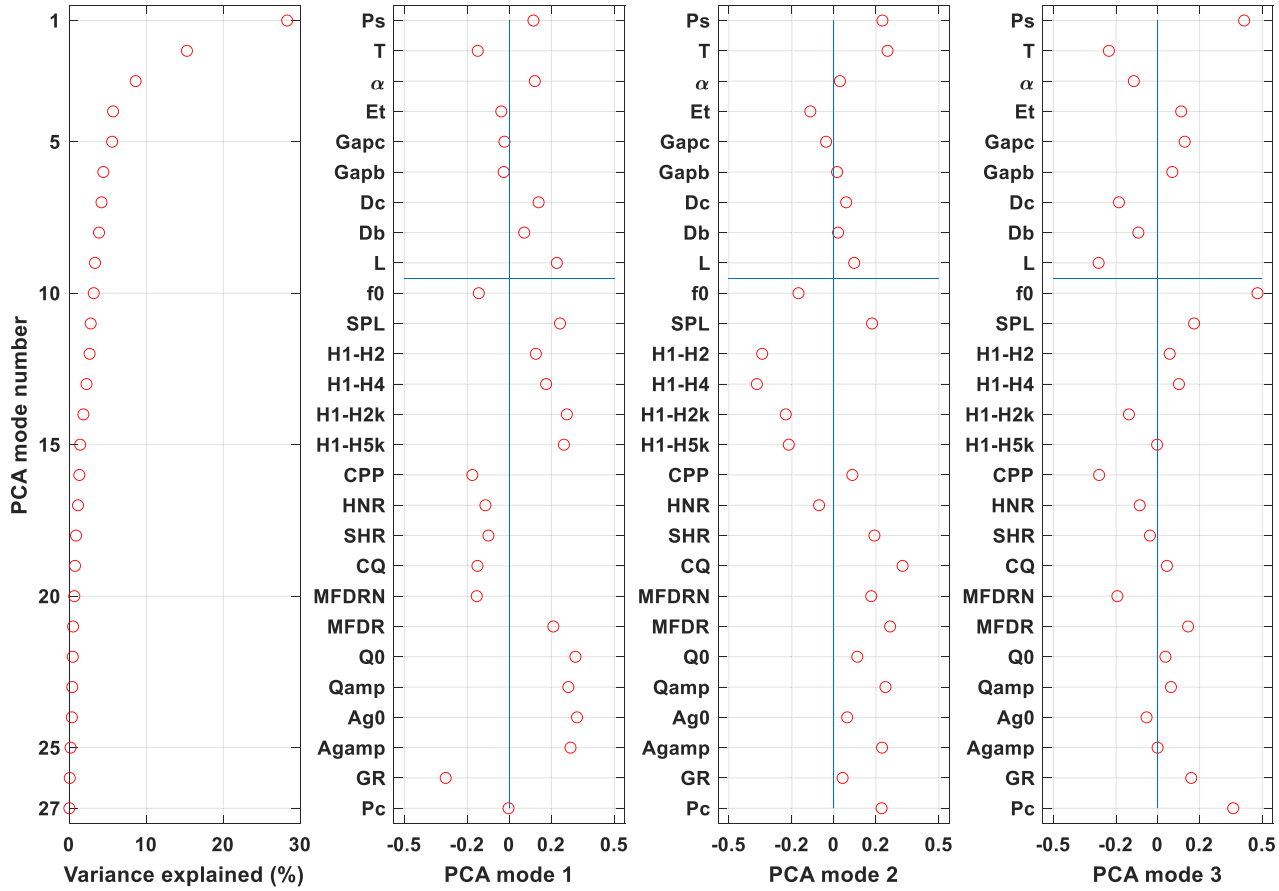

FIG. 2. (Color online) The percentage of variance explained by each PCA mode (left) and the contribution of individual control parameters and voice outcome measures to the first three PCA modes (right). See text for the definition of individual controls and voice outcome measures.

result, the increase in vocal intensity is mostly associated with an increase in energy at the fundamental frequency, and comes at the cost of reduced harmonic energy at high frequencies, as reflected in the increase in the spectral slope measures H1-H2k and H1-H5k. This mode of respiratory-laryngeal adjustments appears to have little effect on the peak vocal fold contact pressure, thus allowing the use of high subglottal pressure without significantly increasing the risk of vocal fold injury.

In contrast, the second PCA mode describes a vocal control strategy in which increasing the subglottal pressure P_s is accompanied by simultaneously increasing vocal fold vertical thickness T and reducing vocal fold transverse stiffness E_t, which can be achieved by activating the thyroarytenoid muscles (Zhang, 2016b). This adjustment allows simultaneous increase in vocal intensity SPL and high-frequency harmonic energy (reduced H1-H2k and H1-H5k). While the increase in vocal intensity increases the peak-to-peak amplitudes (Agamp and Qamp) of the glottal area and glottal flow waveforms, the increase in the mean flow rate Q_0 is much smaller in comparison, thus conserving air consumption while increasing vocal intensity. The duration of glottal closure during phonation is significantly increased, as reflected in a significantly increased closed quotient CQ. However, this adjustment of vocal fold thickening while increasing subglottal pressure does significantly increase the peak vocal fold contact pressure P_c, thus potentially increasing the risk of vocal fold injury.

The third PCA mode mainly functions to control the fundamental frequency f_0. This is achieved by a combination of adjustments including increasing subglottal pressure P_s, increasing vocal fold approximation (decreasing α), and shortening (reducing L), stiffening (increasing E_t, G_{apc}, and G_{apb}), and thinning (reducing T) the vocal folds. Such adjustments tend to increase the peak vocal fold contact pressure P_c, which is mostly due to the significantly increased subglottal pressure.

Note that the first and third PCA modes have a moderate contribution from the vocal fold cover layer depth, despite the small effects of vocal fold depths on the voice source identified in previous studies [e.g., Zhang (2021)]. This is likely due to a positive relationship between the vocal fold length and cover layer depth in the simulation design, as mentioned earlier in the method section, rather than a direct effect of cover layer depth on the voice outcome measures.

Similar observations can be made when the PCA was applied to the voice outcome data alone. The first three PCA modes were almost identical to the PCA modes in Fig. 2, and each accounted for 37%, 20%, and 9% of the total variance, respectively.

The same three PCA modes were observed when the PCA analysis was applied to the second set of data, or voice data simulated with a vocal tract. When applied to the combined data of outcome measures and control parameters, the first three PCA modes captured 23%, 17%, and 11% of the total variance, respectively. This suggests that these principal dimensions are robust enough to remain the same across different vowels.

IV. DISCUSSION AND CONCLUSION

Our results showed that three principal dimensions in the voice outcome space emerge naturally from the physiology of voice production. These dimensions accounted for about 66% of the variance in the voice outcome space and 52% in the combined voice outcome-control space. These principal dimensions remain almost unchanged with or without a vocal tract, and across different vocal tract shapes.

The dominance of these three dimensions suggests that voice changes along these principal dimensions are likely to be more consistently produced and perceived by most speakers than other voice changes, and thus are more likely to have emerged during evolution and be used to convey important information. For example, the first two dimensions reflect energy balance between low- and high-frequency harmonics. Variations in this energy balance lead to voice changes along the continuum from a breathy to pressed voice quality, which are reliably produced and perceived by most speakers (Kreiman, 2024) and have been used to convey meaning in many languages [e.g., Keating *et al.* (2023)].

Since laryngeal anatomy predates the emergence of language, it is likely that voice changes along these principal dimensions may have evolved to convey primarily physiological differences or changes in the speaker' physiological state, rather than linguistic contrasts, in both humans and animals [see, e.g., Anikin *et al.* (2023)]. The existence of these principal dimensions suggests that physiological changes aligned with these dimensions can be more reliably encoded in the voice and perceived from the voice than other physiological changes. One such example is emotion, which strongly impacts both respiratory and laryngeal activities. Much previous research on vocal expression of emotion in humans has shown that different emotions can be differentiated along a small number of principal dimensions in the voice acoustic space (Laukka *et al.*, 2005; Patel *et al.*, 2011; Scherer *et al.*, 2017; Sundberg *et al.*, 2024). Interestingly, the three principal components identified from human voice data in Patel *et al.* (2011) are qualitatively similar to the first three PCA modes identified in our study using data generated from computational simulations: our three modes qualitatively correspond to their second, first, and third components, respectively. Voice changes along a continuum aligned largely with the first two principal dimensions of this study also signal arousal across many species [see, e.g., Congdon *et al.* (2019) and Schwartz *et al.* (2022)]. In addition to the principal dimensions of the voice outcome space, in this study we were also able to identify the physiological control mechanisms underlying the first three dimensions. A better understanding of how emotion impacts physiology underlying the principal dimensions may allow different emotions to be better differentiated and monitored

at the physiological level than at the acoustic level, which is worth further investigation in the future.

Larynx size differences related to sex and age are another example of physiological properties directly impacting the three dimensions, and can be reliably perceived from the voice. In general, male vocal folds are longer and thicker than female vocal folds, which are again longer and thicker than vocal folds in children. These differences in length and thickness align perfectly with the third dimension, which impacts mostly the fundamental frequency of the voice, and thus can be reliably encoded in the voice primarily through changes in the fundamental frequency. Size differences can also be encoded separately along the first two dimensions: the length difference can be encoded along the first dimension and the thickness difference along the second dimension. Thus, changes in voice quality, which are described in the first two dimensions, may also influence voice gender perception, depending on how a specific voice is represented along the first two principal dimensions. This potentially differential encoding of size differences along the three principal dimensions may partially explain why gender perception from the voice is dominated by the fundamental frequency, but may also be manipulated by changes in voice quality [e.g., Skuk and Schweinberger (2014) and Zhang *et al.* (2022)].

Both the second and third PCA modes of this study have a large impact on the peak vocal fold contact pressure, an important contributing factor to vocal fold injury. This large impact is mostly due to the use of high subglottal pressure and to a lesser degree the increased vocal fold thickness. Both dimensions are often considered an indicator of arousal [e.g., Patel *et al.* (2011)]. This appears to support the general impression that angry voices, produced with high arousal and power, tend to be harmful to the vocal folds, due to the potentially high vocal fold contact pressure. On the other hand, the first PCA mode suggests that the negative impact of high subglottal pressure on the peak vocal fold contact pressure can be mitigated by reducing vocal fold adduction, although at the cost of reduced harmonic production at high frequencies. A balance between high-frequency harmonic production and peak vocal fold contact pressure can be achieved at some combinations of the first two PCA modes, which may correspond to a flow phonation configuration described in Gauffin and Sundberg (1989) or those targeted in resonant voice therapy (Verdolini-Marston *et al.*, 1995).

In this study, the principal dimensions were identified from simulation data generated from parametric variations in the independent model control parameters. In humans, the model controls are not necessarily independent of each other. For example, vocal fold geometry, stiffness, and glottal gap are controlled by the same set of laryngeal muscles and thus often co-vary [see a review in Zhang (2023b)]. Tongue movement may affect the vertical position of the larynx, which again impacts both vocal fold configuration and vocal tract length (Vilkman *et al.*, 1996). In addition to these anatomical and physiological factors, how we produce voice is also influenced by socio-cultural factors. The fact that the three principal dimensions in this study only account for about 50% of the variance suggests that there is plenty of room for other factors to come in and further shape the way speakers produce and perceive voice. Understanding how these additional factors further constrain vocal control and the voice outcome space would provide important insight into how emotion and other physiological information of the speaker are encoded in the voice and how to reliably extract such information from the voice. This will be addressed in future studies.

ACKNOWLEDGMENTS

The author thanks Jody Kreiman for helpful discussions and comments on an earlier version of this paper. This study was supported by research Grant No. R01 DC020240 from the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health.

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the author upon reasonable request.

Anikin, A., Canessa-Pollard, V., Pisanski, K., Massenet, M., and Reby, D. (**2023**). "Beyond speech: Exploring diversity in the human voice," Iscience **26**(11), 108204.

Congdon, J. V., Hahn, A. H., Filippi, P., Campbell, K. A., Hoang, J., Scully, E. N., Bowling, D. L., Reber, S. A., and Sturdy, C. B. (**2019**). "Hear them roar: A comparison of black-capped chickadee (*Poecile atricapillus*) and human (*Homo sapiens*) perception of arousal in vocalizations across all classes of terrestrial vertebrates," J. Comp. Psychol. **133**, 520–541.

Gauffin, J., and Sundberg, J. (**1989**). "Spectral correlates of glottal voice source waveform characteristics," J. Speech. Lang. Hear. Res. **32**, 556–650.

Keating, P., Kuang, J., Garellek, M., and Esposito, C. M. (**2023**). "A cross-language acoustic space for vocalic phonation distinctions," Language **99**(2), 351–389.

Kreiman, J. (**2024**). "Information conveyed by voice quality," J. Acoust. Soc. Am. **155**(2), 1264–1271.

Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (**2021**). "Validating a psychoacoustic model of voice quality," J. Acoust. Soc. Am. **149**, 457–465.

Laukka, P., Juslin, P., and Bresin, R. (**2005**). "A dimensional approach to vocal expression of emotion," Cognit. Emotion **19**(5), 633–653.

Patel, S., Scherer, K. R., Björkner, E., and Sundberg, J. (**2011**). "Mapping emotions into acoustic space: The role of voice production," Biol. Psychol. **87**(1), 93–98.

Scherer, K. R., Sundberg, J., Fantini, B., and Eyben, F. (**2017**). "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," J. Acoust. Soc. Am. **142**, 1805–1815.

Schwartz, J. W., Sanchez, M. M., and Gouzoules, H. (**2022**). "Vocal expression of emotional arousal across two call types in young rhesus macaques," Anim. Behav. **190**, 125–138.

Skuk, V. G., and Schweinberger, S. R. (**2014**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech Lang. Hear. Res. **57**(1), 285–296.

Sundberg, J., Salomão, G. L., and Scherer, K. R. (**2024**). "Emotional expressivity in singing. Assessing physiological and acoustic indicators of two opera singers' voice characteristics," J. Acoust. Soc. Am. **155**(1), 18–28.

Verdolini-Marston, K., Burke, M. D., Lessac, A., Glaze, L., and Caldwell, E. (**1995**). "A preliminary study on two methods of treatment for laryngeal nodules," J. Voice **9**, 74–85.
Vilkman, E., Sonninen, A., Hurme, P., and Korkko, P. (**1996**). "External laryngeal frame function in voice production revisited: A review," J. Voice **10**, 78–92.
Zhang, Z. (**2015**). "Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control," J. Acoust. Soc. Am. **137**(2), 898–910.
Zhang, Z. (**2016a**). "Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model," J. Acoust. Soc. Am. **139**(4), 1493–1507.
Zhang, Z. (**2016b**). "Mechanics of human voice production and control," J. Acoust. Soc. Am. **140**(4), 2614–2635.
Zhang, Z. (**2021**). "Contribution of laryngeal size to differences between male and female voice production," J. Acoust. Soc. Am. **150**(6), 4511–4521.
Zhang, Z. (**2023a**). "The influence of source-filter interaction on the voice source in a three-dimensional computational model of voice production," J. Acoust. Soc. Am. **154**(4), 2462–2475.
Zhang, Z. (**2023b**). "Vocal fold vertical thickness in human voice production and control: A review," J. Voice (published online).
Zhang, Z., Zhang, J., and Kreiman, J. (**2022**). "Effects of laryngeal manipulations on voice gender perception," in *Proceedings of Interspeech 2022*, pp. 1856–1860.