Interspeech 2026
27 September - 1 October 2026, Sydney, Australia

Improved modeling of vocal fold contacting and de-contacting in a geometric vocal fold model

Tianyi Zhang[1], *Zihao Huang*[1], *Peter Birkholz*[1]

[1] Institute of Acoustics and Speech Communication, TU Dresden, Germany

tianyi.zhang2@tu-dresden.de; zihao.huang@mailbox.tu-dresden.de;
peter.birkholz@tu-dresden.de

Abstract

This paper presents a geometric model of the vocal folds for more natural-sounding articulatory speech synthesis. In contrast to similar models, we propose a biomechanically-motivated more gradual opening and closing of the glottis during the phases of partial contact of the vocal folds. This was combined with a mechanism to skew the glottal area waveform to the left or right. The model was implemented in the articulatory speech synthesizer and compared against a similar model without the improved contacting and de-contacting mechanisms in a perception experiment with synthetic sentences for a range of pre-phonatory model settings. In an A/B comparison, the synthetic sentences with the new model were preferred over the same sentences with the previous model in 58.1 % of the comparisons. In addition, the mean opinion scores for the quality on a five-point scale were 3.02 and 2.68 for the new and previous models, respectively.

Index Terms: articulatory speech synthesis, phonation

1. Introduction

Neural speech synthesizers are able to synthesize speech that is often indistinguishable from human speech [1]. However, while they work well for high-resource languages and allow high-level control of speaker characteristics and emotions [2, 3], they are less effective for low-resource languages or when precise control of low-level articulatory and acoustic features is needed (e.g. in basic speech research [4, 5]). Articulatory speech synthesis, which is based on simulation models of the vocal tract [6, 7], the vocal folds [8, 9], acoustic wave propagation [10, 11], and articulatory control [12, 13], does allow the individual control of all relevant parameters of the speech production process and is a potential candidate for TTS systems for low-resource languages. However, speech generated with articulatory synthesizers does not yet sound fully natural [14]. A particularly important factor for their naturalness is the voice source model [15].

This paper presents a new voice source model that was designed for high-quality articulatory synthesis. It is derived from the geometric vocal fold model by Titze [16] and is shown in Figure 1. In Titze's model, only the right vocal fold is modeled, while the left vocal fold moves in a mirror image. The upper and lower edges of the vocal fold are assumed to oscillate sinusoidally like a vibrating string (fixed at the anterior and posterior ends) in its fundamental mode. In the coronal plane, the glottis shape is assumed to be trapezoidal with straight lines connecting the upper and lower vocal fold edges. For glottal area computation, all negative values for lateral displacement, which indicate contact between the left and right vocal folds, are set to zero (see Sec. 2).

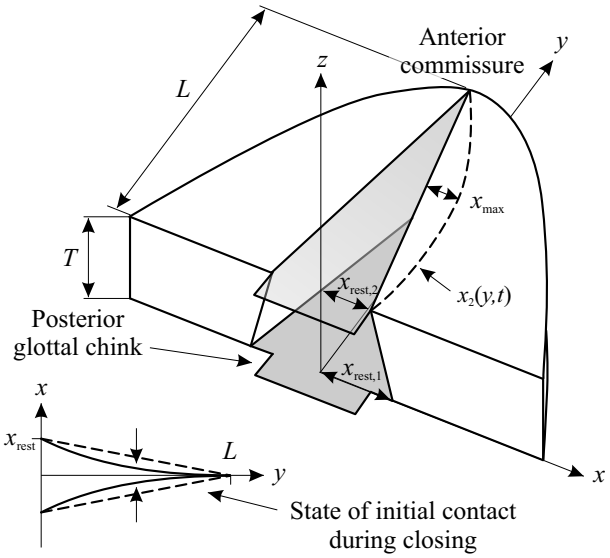

Figure 1: *Schematic of the geometric vocal fold model.*

This basic model was later extended by a posterior chink to model "glottal leakage" [17, 18], by introducing rounded (instead of flat) medial surfaces of the vocal folds [8], and by allowing the generation of skewed glottal area pulses [19], which are commonly observed in human phonation [20, 21, 22].

A common simplification of all these model variants is that the edges of the vocal folds perform a sinusoidal oscillation, similar to a freely vibrating string, which is clipped at the mid-sagittal plane. This neglects what happens during the short but important contacting and de-contacting phases during the oscillation cycle [23]. In the contacting phase, the tissue deforms due to the (partial) vocal fold collision and leads to a restoring force that decelerates their further medial movement. In addition, when the vocal folds separate from each other during glottal opening, the driving force has to overcome the adhesive force between them, which should lead to an initially reduced velocity of the lateral movement.

The vocal fold model proposed here includes these effects during the contacting and de-contacting phases, and allows skewed glottal area pulses to be generated. In a perception experiment with synthetic sentences, this model is then compared against a similar previous model [19], which does not include the "soft" contacting and de-contacting, and which has a different mechanism to generate skewed glottal area pulses. In the following, the reference model [19] is called the "2019 model".

10.21437/Interspeech.2026-753

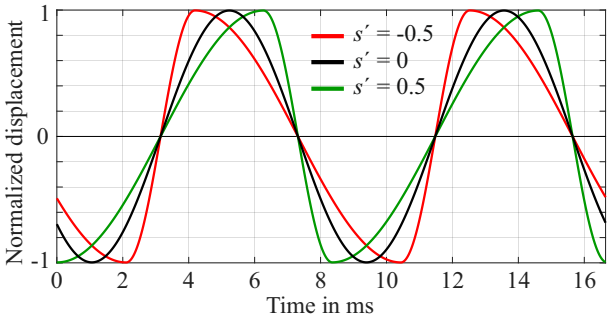

Figure 2: *Oscillation functions $f_{\text{osc}}(\cdot)$ for two periods of the 2019 vocal fold model for three different settings of the pulse shape parameter s'.*

2. Model Design

The purpose of the proposed model, shown in Figure 1, is to generate the time functions of the glottal areas $A_1(t)$ and $A_2(t)$ between the lower and the upper edges of the vocal folds, respectively, which are needed as input for the articulatory synthesizer (see Sec. 3.1). The length L and thickness T of the vocal folds vary with the fundamental frequency f_0 according to

$$L = L_0\sqrt{f_0/f_{\text{nat}}} \quad \text{and} \quad T = T_0 L_0/L, \tag{1}$$

where L_0, T_0, and f_{nat} are speaker-specific reference values for the vocal fold length, thickness, and f_0. The lateral displacements $x_i(y,t)$ of the lower ($i=1$) and upper ($i=2$) edges of the right vocal fold are the sum of a static part $x_{\text{stat},i}(y)$ and a dynamic part $x_{\text{dyn},i}(y,t)$:

$$x_i(y,t) = x_{\text{stat},i}(y) + x_{\text{dyn},i}(y,t). \tag{2}$$

The left vocal fold is a mirror image of the right vocal fold. The *static* (prephonatory) displacement is a linear function of y that equals the adjustable rest displacement $x_{\text{rest},i}$ at the vocal processes and is zero at the anterior commissure, i.e.,

$$x_{\text{stat},i}(y) = x_{\text{rest},i} - y \cdot x_{\text{rest},i}/L, \quad i = 1,2. \tag{3}$$

The *dynamic* displacement is

$$x_{\text{dyn},i}(y,t) = x_{\max} \cdot \sin(y\pi/L) \cdot f_{\text{osc}}(\beta_i(t)), \quad i = 1,2 \tag{4}$$

where $x_{\max}$ is the amplitude of oscillation, and f_{osc} is the "oscillation function" that models the normalized dynamic displacement of the vocal fold edge i as a function of the instantaneous phase $\beta_i(t)$. The amplitude $x_{\max}$ is proportional to the rest length L_0 [24] and to the square root of the transglottal pressure P_{trans} [25], i.e.

$$x_{\max} = c \cdot x_{\text{rel}} \cdot L_0 \cdot \sqrt{P_{\text{trans}}}, \tag{5}$$

where $x_{\text{rel}} \in [0,1]$ is an adjustable amplitude factor that is 1 for vowels and can assume smaller values for certain consonants. The constant c was empirically adjusted to obtain $x_{\max} = 0.63$ mm for $P_{\text{trans}} = 800$ Pa and $L_0 = 1.6$ cm. While a slightly different equation was used in [19], the above equation was used for both the 2019 model and the proposed model in the present study. The glottal areas $A_1(t)$ and $A_2(t)$ are obtained as

$$A_i(t) = \left[2\int_0^L \max\{x_i(y,t),0\}dy\right] + A_{\text{chink}}, \quad i = 1,2 \tag{6}$$

where A_{chink} is the area of a posterior glottal chink [17].

The main difference between the proposed model and the 2019 model lies in the type of the oscillation function. The 2019 model used $f_{\text{osc}}(\beta_i(t)) = \cos(\text{warp}(\alpha_i(t)))$, where $\text{warp}(\cdot)$ is a warping function (see [19] for details) and $\alpha_i(t)$ is the (unwarped) instantaneous phase, namely,

$$\alpha_1(t) = 2\pi\int_0^t f_0(\tau)d\tau \tag{7}$$

$$\alpha_2(t) = \alpha_1(t) - \phi. \tag{8}$$

Here, ϕ is the (adjustable) phase delay between the upper and lower vocal fold edges. The warping function modifies the durations of the rising and falling flanks of the cosine function based on a shape parameter $s' \in [-0.5, 0.5]$ as shown in Figure 2. This allows asymmetrical glottal area waveforms to be generated, either with faster glottal opening than closing for $s' < 0$ or vice versa for $s' > 0$. However, informal listening tests indicated that speech signals generated with these asymmetrical waveforms did not always sound natural. This motivated the development of the following new oscillation function.

Based on the ideas discussed in Sec. 1, the proposed function differs from the 2019 model as follows. First, the *medial* movement of the vocal folds starts to decelerate as soon as the left and right vocal folds start to make contact along the glottal midline ("soft contacting"). Second, the *lateral* movement of the vocal folds at the beginning of glottal opening (when the upper or lower vocal fold edges are still in partial contact) happens more gradually ("soft de-contacting"). Finally, the soft contacting and de-contacting phases can be extended in time to generate asymmetrical glottal area waveforms.

These ideas are implemented as follows:

$$f_{\text{osc}}(\alpha) = \begin{cases} \max\{\cos(\alpha), f_1(\alpha), f_2(\alpha)\} & \text{for } \alpha \in [\gamma_1, \gamma_2] \\ \cos(\alpha) & \text{otherwise} \end{cases} \tag{9}$$

Here, α applies to both the (wrapped) instantaneous phases $\alpha_1(t)$ and $\alpha_2(t)$. The functions $f_1(\alpha)$ and $f_2(\alpha)$ model the soft contacting and de-contacting, respectively, by means of natural exponential functions, namely

$$f_1(\alpha) = d_{\min} + (\cos(\gamma_1) - d_{\min})e^{-(\alpha-\gamma_1)/\tau_1}, \tag{10}$$

$$f_2(\alpha) = d_{\min} + (\cos(\gamma_2) - d_{\min})e^{-(\gamma_2-\alpha)/\tau_2}. \tag{11}$$

Here, γ_1 is the instantaneous phase where the soft contacting begins, and γ_2 is the instantaneous phase where the soft decontacting ends (see examples in Figure 3). They are calculated as

$$\gamma_1 = \arccos(d_{\text{contact}}) + H(-s) \cdot 0.2\pi \cdot s \tag{12}$$

$$\gamma_2 = 2\pi - \arccos(d_{\text{contact}}) + H(s) \cdot 0.2\pi \cdot s, \tag{13}$$

where $H(\cdot)$ is the Heaviside step function and $s \in [-1,1]$ is a shape parameter to generate asymmetrical glottal area pulses (s has no prime sign to distinguish it from the shape parameter s' of the 2019 model). As Figure 3 illustrates, glottal area pulses are symmetrical for $s = 0$, leaning towards the left for $s < 0$, and leaning towards the right for $s > 0$. The variables d_{contact} and $d_{\min}$ in Eqs. (10) to (13) are the normalized displacement value where the vocal fold edges just start to make contact during closure (as shown at the bottom of Figure 1) and the minimal value of the function, specifically

$$d_{\text{contact}} = -\frac{x_{\text{rest}}}{x_{\max} \cdot \pi} \quad \text{and} \quad d_{\min} = d_{\text{contact}} - 0.2. \tag{14}$$

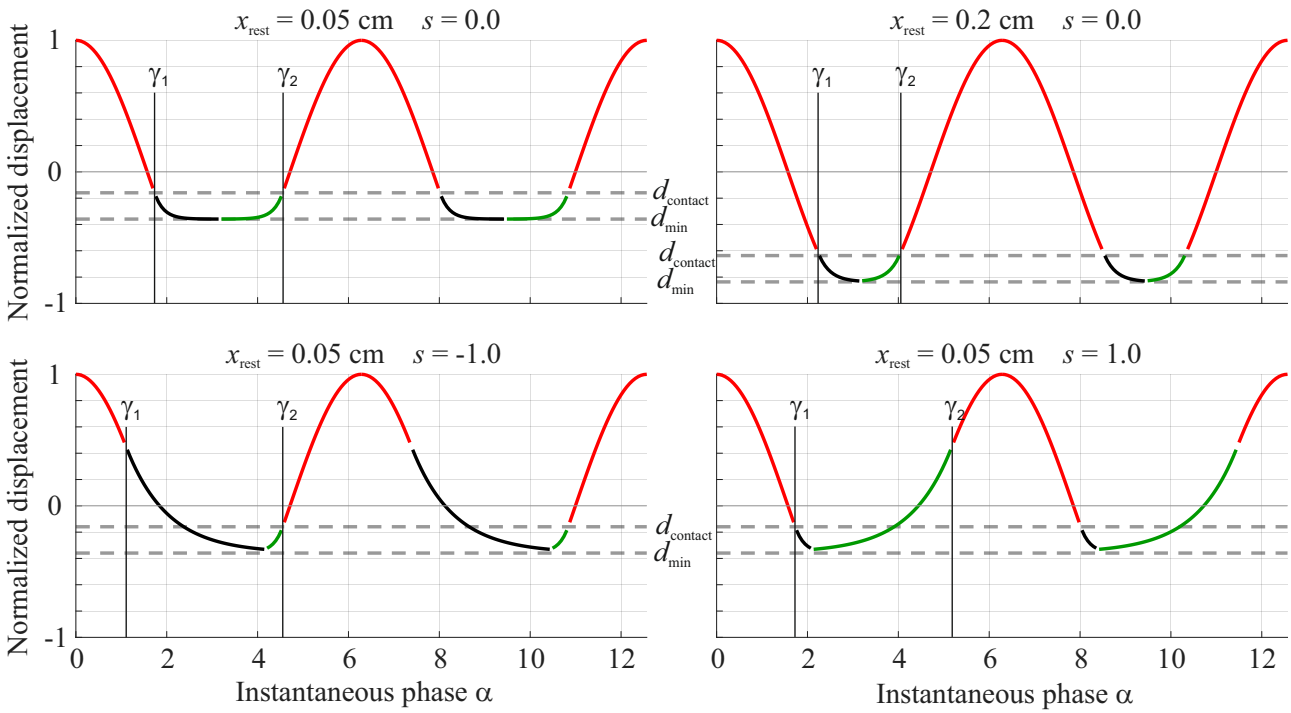

Figure 3: *Oscillation functions $f_{\mathrm{osc}}(\alpha)$ of the proposed model for four combinations of rest displacement x_{rest} and shape parameter s. The curve segments in red, black and green represent segments of the functions $\cos(\alpha)$, $f_1(\alpha)$ and $f_2(\alpha)$, respectively.*

Note that their values differ between the lower and upper vocal fold edges, as x_{rest} is either $x_{\mathrm{rest},1}$ or $x_{\mathrm{rest},2}$. The constant 0.2 was empirically determined. Finally, the decay parameters τ_1 and τ_2 in Eqs. (10) and (11) are calculated in such a way that the exponentials are tangent to the function $\cos(\alpha)$ at the positions γ_1 and γ_2. An important feature of the proposed model is that the oscillation function approaches that of a pure cosine function (as in the 2019 model) when the rest displacement of the vocal folds increases and d_{contact} decreases accordingly (see top row of Figure 3). This is important, for example, for realistic transitions from vowels to voiceless consonants.

3. Perceptual Evaluation

3.1. Stimulus Generation

Both the proposed vocal fold model (NEW) and the 2019 baseline model (OLD) were implemented in the articulatory speech synthesizer VocalTractLab 2.4 (VTL) (`www.vocaltractlab.de`, [26, 27, 28]). At the acoustic level, the synthesizer represents the vocal tract as a branched tube system composed of concatenated short cylindrical tube sections [29]. The glottis is modeled by two adjacent tube sections with time-varying areas $A_1(t)$ and $A_2(t)$ according to Eq. (6), each having a length of $T/2$, where T denotes the vocal fold thickness (Eq. (1)). Based on these tube sections, glottal flow is computed as in [30], including configuration-dependent flow separation [31]. Speech stimuli were generated by copy synthesis of nine sentences using the default male speaker model "JD2", preserving original phone durations and f_0 contours (as shown in: `https://youtu.be/J_n01Xe4aME`). Speaker-specific vocal fold reference parameters were $L_0 = 1.6\,\mathrm{cm}$, $T_0 = 0.45\,\mathrm{cm}$, $f_{\mathrm{nat}} = 120\,\mathrm{Hz}$ and $A_{\mathrm{chink}} = 0.02\,\mathrm{cm}^2$. For modal phonation, stimuli were synthesized for all combinations of the following four vocal fold model parameters:

- Lower rest displacement $x_{\mathrm{rest},1} \in \{0.1,\ 0.4,\ 0.7\}$ mm
- Upper rest displacement $x_{\mathrm{rest},2} \in \{0.1,\ 0.4,\ 0.7\}$ mm
- Phase lag $\phi \in \{40^\circ,\ 70^\circ,\ 100^\circ\}$
- Pulse shape $s \in \{-1,\ 0,\ +1\}$, $s' \in \{-0.5,\ 0,\ +0.5\}$

This resulted in 81 parameter combinations per sentence and per model. Across nine sentences and two models, a total of 1,458 utterances were generated. Note that s (proposed model) and s' (2019 model) both control pulse skewness but have different corresponding range limits.

For the glottal configuration of consonants, the same phase lag and pulse-shape settings were used, while abduction parameters were adjusted as required. The double-pulsing parameter of the 2019 model was fixed to zero to ensure comparability, as this feature is not available in the proposed model. Flutter (quasi-random f_0 fluctuation) was set to 25% for all stimuli in both models. All stimuli are available in the supplemental material (`https://www.vocaltractlab.de/index.php?page=birkholz-supplements`).

3.2. Experimental Design and Procedure

A total of 27 listeners (8 female, 19 male; age range 18–47 years) participated in the experiment. Twenty-two were native German speakers; the remaining participants reported high proficiency in German. All participants had normal hearing.

To prevent listener fatigue while ensuring full coverage of the parameter space, each participant was assigned only one subset of stimuli. For each glottis model, the complete set of 729 stimuli (9 sentences × 81 parameter combinations) was evenly partitioned into nine subsets according to the parameter combinations, with each subset containing 81 stimuli and representing all nine sentences equally (9 stimuli per sentence).

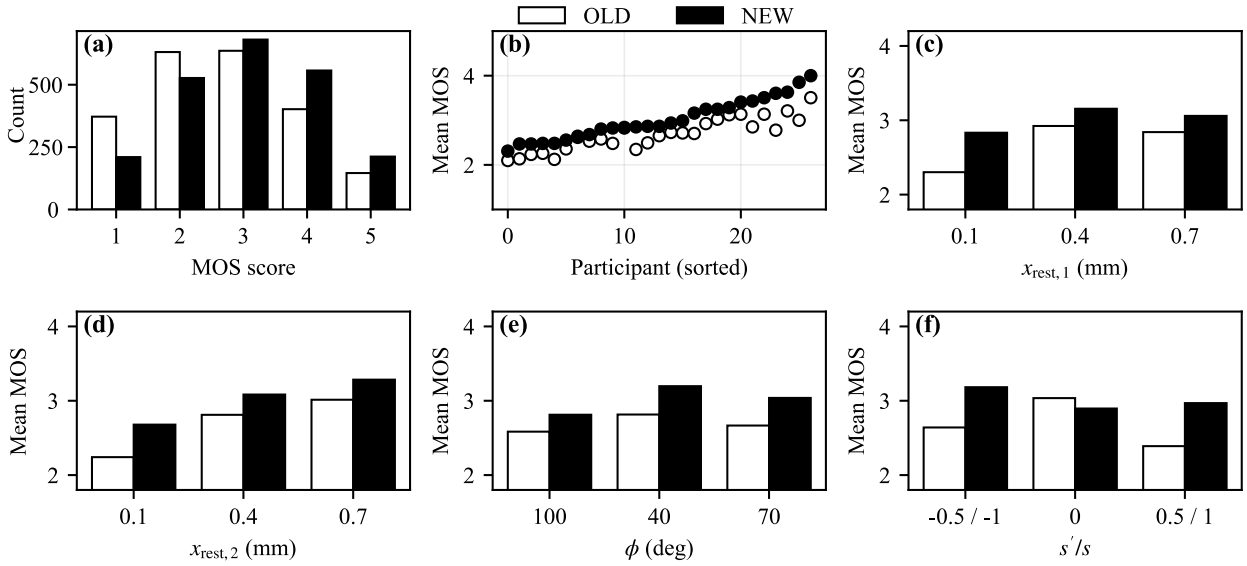

Figure 4: *(a) MOS score distribution for both glottis models. (b) Participant-wise mean MOS score. (c–f) Mean MOS across parameter levels* $x_{\text{rest},1}$*,* $x_{\text{rest},2}$*,* φ*, and* s*. The parameter* s' $[-0.5,\ 0.5]$ *of the OLD corresponds to a linear rescaling of* s $[-1,\ 1]$*, with* $s' = \frac{1}{2}s$*.*

For evaluation, the corresponding subsets of the NEW and OLD models were combined, yielding 162 stimuli per participant. The nine subsets together constituted one complete round. With 27 participants in total, three complete rounds were obtained.

Each participant performed two perceptual tasks: an AB preference test comparing the OLD (2019 model) and NEW models, and a MOS test for the evaluation of naturalness. In the AB task, each trial presented a pair of stimuli synthesized with identical sentence and parameter settings but generated using either the NEW or the OLD vocal fold model. Presentation order within each pair was randomized, and participants performed a forced-choice judgment indicating which stimulus they preferred.

In the MOS task, stimuli were presented individually in fully randomized order. Listeners rated each sample on a five-point scale ranging from 1 (very unnatural) to 5 (very natural). Participants were instructed to evaluate each stimulus independently and to avoid direct comparisons between samples. Both tasks were conducted in a quiet environment using headphones with calibrated playback levels.

3.3. Statistical Analysis

In addition to a descriptive statistical analysis, MOS ratings were analyzed using a linear mixed-effects model (LMM) to estimate fixed effects while accounting for random variability. The model was specified as follows:

$$\begin{aligned}\text{MOS}_{ij} =& \beta_0 + \beta_1 \text{NEW}_{ij} + \boldsymbol{\lambda}^\top \mathbf{x}_{ij} \\ &+ \boldsymbol{\gamma}^\top (\text{NEW}_{ij}\mathbf{x}_{ij}) + u_i + v_j + \varepsilon_{ij},\end{aligned} \tag{15}$$

where $\mathbf{x} = (x_{\text{rest},1}, x_{\text{rest},2}, \phi, s)^\top$; β_0 denotes the baseline MOS corresponding to the reference condition (OLD model with $x_{\text{rest},1} = 0.1$, $x_{\text{rest},2} = 0.1$, $\phi = 40°$, $s = -1$); β_1 denotes the overall advantage of the NEW model relative to OLD; $\boldsymbol{\lambda}$ represents the main effects of the parameters; $\boldsymbol{\gamma}$ captures the corresponding interaction effects between model identity and each parameter. Random intercepts were included for listener (u_i) and sentence (v_j). For the AB preference, a binomial test was conducted.

Table 1: *Selected fixed-effect estimates of the LMM.*

Symbol	Effect	Level	$\hat{\beta}$	p
β_0	Intercept	baseline	1.933	$< .001$
β_1	NEW	(vs OLD)	0.911	$< .001$
$\gamma_{6,0.4}$	NEW $\times x_{\text{rest},1}$	0.4	-0.299	$< .001$
$\gamma_{6,0.7}$	NEW $\times x_{\text{rest},1}$	0.7	-0.313	$< .001$
$\gamma_{7,0.4}$	NEW $\times x_{\text{rest},2}$	0.4	-0.164	.022
$\gamma_{7,0.7}$	NEW $\times x_{\text{rest},2}$	0.7	-0.167	.021
$\gamma_{8,70}$	NEW $\times \varphi$	70	-0.009	.894
$\gamma_{8,100}$	NEW $\times \varphi$	100	-0.154	.034
$\gamma_{9,0}$	NEW $\times s$	0	-0.682	$< .001$
$\gamma_{9,1}$	NEW $\times s$	1	-0.037	.61

4. Results and Conclusions

From the descriptive statistics (Figure 4), the NEW model demonstrates an overall advantage in MOS ratings. Figure 4(b) illustrates substantial inter-participant variability in rating behavior, confirming the necessity of accounting for random listener effects through the linear mixed-effects model. The LMM revealed a significant main effect of model at the baseline configuration ($\beta_1 = 0.911$, $p < .001$), indicating higher perceived naturalness for the NEW model. Although most interaction coefficients were negative, model-based predictions derived from Eq. (15) showed that the NEW model achieved higher MOS ratings than the OLD model in 75.3% of the 81 parameter combinations. Consistent with the LMM results, the binomial test for the AB preference task indicated that the proportion of preferences for the NEW model (58.1%, 95% CI [56.0%, 60.2%], $p < .001$) was significantly above chance level. Additional plots illustrating MOS ratings across individual parameter combinations are provided in the supplemental material.

Overall, the results indicate that the proposed vocal fold model yields a statistically significant overall improvement in perceived naturalness compared to the 2019 baseline model.

5. Acknowledgements

This research was funded by the German Research Foundation (DFG), grant BI 1639/9-1, and by the "Zentrales Innovationsprogramm Mittelstand (ZIM)" by the German Federal Ministry for Economic Affairs and Energy (BMWK), grant no. KK5049503FG3.

6. Generative AI Use Disclosure

No generative AI tools were used to generate scientific ideas, research methodology, experimental results, analyses, or conclusions. All technical content was created, verified, and approved by the authors, who take full responsibility for the work.

7. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*, 2022, pp. 2709–2720.

[3] D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee, "EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 2365–2380, 2025.

[4] Y. Xu, A. Lee, W.-L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PLOS ONE*, vol. 8, no. 4, p. e62397, 2013.

[5] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger, and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: an articulatory synthesis study," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[6] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533–553, 2002.

[7] K. Iskarous, L. Goldstein, D. H. Whalen, M. Tiede, and P. Rubin, "CASY: The Haskins configurable articulatory synthesizer," in *International Congress of Phonetic Sciences*, 2003, pp. 185–188.

[8] I. R. Titze, "A four-parameter model of the glottis and vocal fold contact area," *Speech Communication*, vol. 8, pp. 191–201, 1989.

[9] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, "A review of lumped-element models of voiced speech," *Speech Communication*, vol. 55, no. 5, pp. 667–690, 2013.

[10] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, pp. 199–229, 1982.

[11] K. van den Doel and U. M. Ascher, "Real-time numerical solution of Webster's equation on a nonuniform grid," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1163–1172, 2008.

[12] E. L. Saltzman and K. G. Munhall, "A dynamic approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382, 1989.

[13] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan, "A modular architecture for articulatory synthesis from gestural specification," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4458–4471, 2019.

[14] P. K. Krug, S. Stone, and P. Birkholz, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, Budapest, Hungary, 2021, pp. 102–107.

[15] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[16] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *The Journal of the Acoustical Society of America*, vol. 75, no. 2, pp. 570–580, 1984.

[17] B. Cranen and J. Schroeter, "Modeling a leaky glottis," *Journal of Phonetics*, vol. 23, pp. 165–177, 1995.

[18] ——, "Physiologically motivated modelling of the voice source in articulatory analysis/synthesis," *Speech Communication*, vol. 19, pp. 1–19, 1996.

[19] P. Birkholz, S. Drechsel, and S. Stone, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis," in *Proc. of the Interspeech*, 2019, pp. 3765–3769.

[20] P. Woo, "Quantification of videostrobolaryngoscopic findings - measurements of the normal glottal cycle," *The Laryngoscope*, vol. 106, no. S79, pp. 1–27, 1996.

[21] H. Pulakka, P. Alku, S. Granqvist, S. Hertegard, H. Larsson, A.-M. Laukkanen, P.-A. Lindestad, and E. Vilkman, "Analysis of the voice source in different phonation types: simultaneous high-speed imaging of the vocal fold vibration and glottal inverse filtering," in *Interspeech 2004*, Jeju Island, Korea, 2004, pp. 1121–1124.

[22] P. Alku, T. Murtola, J. Malinen, A. Geneid, and E. Vilkman, "Skewing of the glottal flow with respect to the glottal area measured in natural production of vowels," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2501–2509, 2019.

[23] C. T. Herbst, "Electroglottography – an update," *Journal of Voice*, vol. 34, no. 4, pp. 503–526, 2020.

[24] Z. Zhang, "Contribution of laryngeal size to differences between male and female voice production," *The Journal of the Acoustical Society of America*, vol. 150, no. 6, pp. 4511–4521, 2021.

[25] I. R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 901–906, 1989.

[26] P. Birkholz, "3D-Artikulatorische Sprachsynthese," Ph.D. dissertation, University of Rostock, 2005.

[27] ——, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLOS ONE*, vol. 8, no. 4, p. e60603, 2013.

[28] S. Stone and P. Birkholz, "Monophthong vocal tract shapes are sufficient for articulatory synthesis of German primary diphthongs," *Speech Communication*, vol. 157, p. 103041, 2024.

[29] P. Birkholz, *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin, 2005.

[30] P. Birkholz, D. Jackèl, and B. J. Kröger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[31] P. Birkholz and D. Pape, "How modeling entrance loss and flow separation in a two-mass model affects the oscillation and synthesis quality," *Speech Communication*, vol. 110, pp. 108–116, 2019.