*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis

*Peter Birkholz*[1], *Susanne Drechsel*[2], *Simon Stone*[1]

[1]Institute of Acoustics and Speech Communication, TU Dresden, Germany
[2]Department of Speech Science and Phonetics, Martin Luther University of Halle-Wittenberg
peter.birkholz@tu-dresden.de, SusanneDrechsel@gmx.de, simon.stone@tu-dresden.de

## Abstract

We present a geometric vocal fold model that describes the glottal area between the lower and upper vocal fold edges as a function of time. It is based on a glottis model by Titze [J. Acoust. Soc. Am., 75(2), 570–580 (1984)] and has been enhanced to allow the generation of skewed (asymmetric) glottal area waveforms and diplophonic double pulsing. Embedded in the articulatory speech synthesizer VocalTractLab, the model was used for the synthesis of German words with a range of settings for the vocal fold model parameters to generate different male and female voices. A perception experiment was conducted to determine the parameter settings that generate the most natural-sounding voices. The most natural-sounding male voice was generated with a slightly divergent prephonatory glottal shape, with a phase delay of 70° between the lower and upper vocal fold edges, symmetric glottal area pulses, and a little shimmer (double pulsing). The most natural-sounding female voice was generated with a straight prephonatory glottal channel, with a phase delay of 50° between the vocal fold edges, slightly asymmetric glottal area pulses, and a little shimmer.
**Index Terms**: perceptual optimization, articulatory synthesis

## 1. Introduction

The naturalness of synthetic speech generated with an articulatory speech synthesizer largely depends on the voice source model. Voice source models can be roughly divided into three classes: models of the glottal flow (e.g., [1, 2]), geometric models of the vocal folds (i.e., models of the glottal area like [3, 4]), and self-oscillating biomechanical models of the vocal folds [5, 6]. In contrast to glottal flow models, geometric and biomechanical vocal fold models naturally account for acoustic source-filter interaction during the synthesis of speech, which is considered important for the naturalness of the voice signal [7]. Biomechanical models are generally still more realistic than geometric models, because they account for the interaction between flow and structure, but their biomechanical properties are also harder to determine [8, 9], and they are harder to control during connected speech synthesis.

The purpose of the present study was to develop a new voice source model for the articulatory speech synthesizer VocalTractLab [10, 11], which is easier to control (i.e., has a more predictable behaviour) than the currently implemented self-oscillating bar-mass model [12], but still accounts for acoustic source-filter interaction. The basis of our model is the glottis model by Titze [3], which essentially specifies the glottal geometry as a function of time and a set of control parameters. To enable a wider range of oscillation patterns and hence voices for the synthesis, we enhanced this model by two mechanisms to generate *asymmetric* glottal area waveforms as well as *double pulsing* [2]. Asymmetric glottal area waveforms (not to be confused with glottal flow pulse asymmetry, which is caused by acoustic inertance) are frequently observed in human phonation (e.g., [13]) and may have an important impact on the characteristic sound of voices. The other mechanism, double pulsing, is often seen in low-pitched voices or during glottalization, and might make synthetic voices more natural-sounding. Besides these extensions, we evaluated the naturalness of the generated voices for a wide range of control parameter settings in a perception experiment with synthesized words.

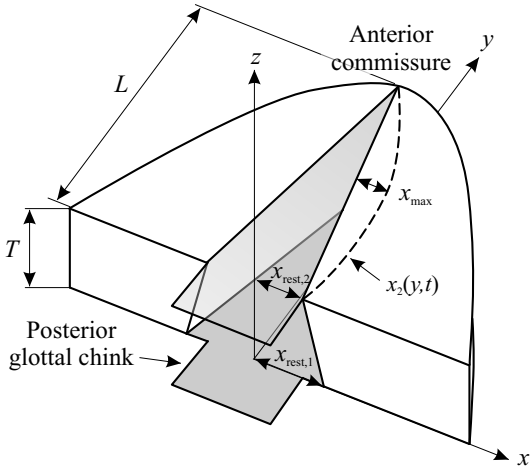


Figure 1: *Parameterization of the glottal geometry. The two gray areas are the lower glottal area $A_1(t)$ and the upper glottal area $A_2(t)$.*

## 2. Geometric vocal fold model

The vocal fold model proposed here is based on the model by Titze [3, 4] and shown in Figure 1. The purpose of this model 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, because they are needed for the aeroacoustic simulation in the framework of the articulatory synthesizer (see below). Similar to [4], the length $L$ and thickness $T$ of the glottis are determined by the empirical relations

$$L = L_0\sqrt{f_0/f_{\mathrm{nat}}} \text{ and} \quad (1)$$

$$T = T_0 L_0/L, \quad (2)$$

where $f_0$ is the fundamental frequency, and $L_0$, $T_0$, and $f_{\mathrm{nat}}$ are speaker-specific reference values for the vocal fold length, thickness, and fundamental frequency. The geometry of the glottis is modeled in terms of the lateral displacement $x_i(y,t)$ of the lower ($i = 1$) and upper ($i = 2$) vocal fold edges, each of



http://dx.doi.org/10.21437/Interspeech.2019-2410

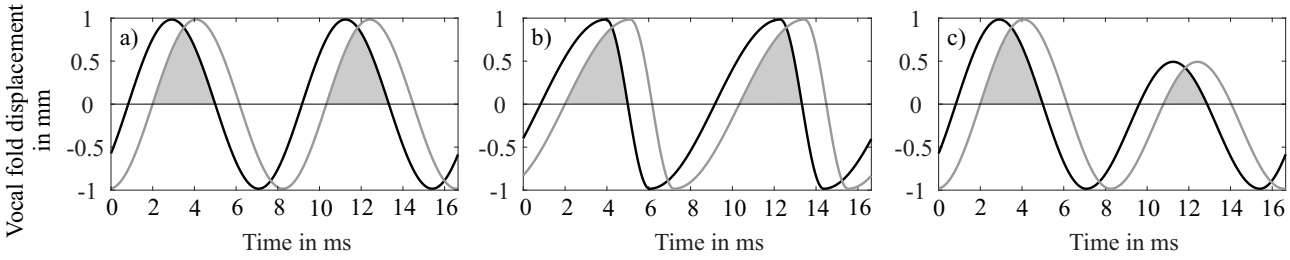


Figure 2: *Examples of simulated vocal fold oscillations. a) Oscillations of the lower (black) and upper (gray) vocal fold edges for the following settings: skewness $s = 0$, double pulsing $d = 0$, phase lag $\phi = 50^\circ$, vocal fold rest displacement $x_{\mathrm{rest},1/2} = 0$ mm. b) Oscillations for the same settings as in a) but with $s = 0.5$. c) Oscillations for the same settings as in a) but with $d = 0.5$. The height of the shaded areas is proportional to the projected glottal area, i.e.,* $\min\{A_1(t), A_2(t)\}$*, when* $A_{\mathrm{chink}} = 0$.

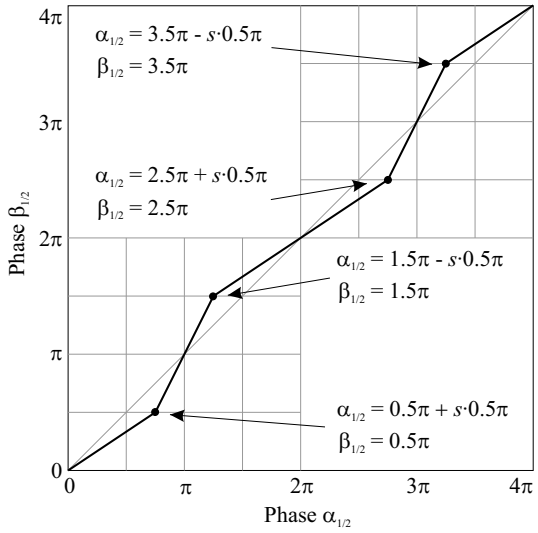


Figure 3: *Piecewise linear warping function to modify the phase of the oscillation such that the time function of the glottal area is skewed. The degree and direction of skewness is controlled by the parameter* $s \in [-1, 1]$. *In this example* $s = 0.5$.

which is the sum of a static part $x_{\mathrm{stat},i}(y)$ and a dynamic part $x_{\mathrm{dyn},i}(y,t)$:

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

Here, $y$ is the position along the anterior-posterior axis. It is assumed that the left and the right vocal folds are symmetric with respect to the midsagittal plane so that only the displacement of the right vocal fold edges is modeled. The *static* (prephonatory) displacement is a linear function of $y$ for both edges, which equals the adjustable rest displacement $x_{\mathrm{rest},i}$ at the level of the vocal processes, and is zero at the anterior commissure, i.e.,

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

The *dynamic* displacement is modeled as

$$x_{\mathrm{dyn},i}(y,t) = x_{\max} \sin(y\pi/L) \sin(\beta_i(t)), \quad i = 1, 2 \quad (3)$$

where $\beta_i(t)$ and $x_{\max}$ are the phase and amplitude of the oscillation, respectively. Based on the observations by Titze [14], the amplitude $x_{\max}$ is modeled as

$$x_{\max} = c \cdot x_{\mathrm{rel}} \cdot \sqrt{P_{\mathrm{trans}}} \cdot L_0/L,$$

where $c$ is a constant, $x_{\mathrm{rel}} \in [0, 1]$ is an adjustable amplitude factor, and $P_{\mathrm{trans}}$ is the transglottal pressure. The constant $c$ is adjusted such that $x_{\max} = 1$ mm for the standard settings $P_{\mathrm{trans}} = 800$ Pa, $L = L_0$, and $x_{\mathrm{rel}} = 1$. Values of $x_{\mathrm{rel}} < 1$ allow to simulate active stiffening of the vocal folds. In the basic model by Titze [3], the oscillation phase $\beta_i(t) = 2\pi f_0 t + \phi_i$ for a constant fundamental frequency $f_0$ and some phase offsets $\phi_i$. Hence, the vocal fold edges oscillate sinusoidally in time like ideal strings that are fixed at their two ends. Accordingly, the generated glottal area waveforms of Titze's model were always symmetric. In the present study, we implemented two mechanisms to extent the flexibility of this oscillation:

1. The phase function can be warped to generate different velocities of the opening and closing movements and hence a skewness of the glottal area waveforms.
2. The amplitude of every second oscillation cycle can be reduced by a certain degree to introduce double pulsing in the generated voice.

To implement these effects, we first expressed the *unwarped* phases $\alpha_1$ and $\alpha_2$ of the oscillations of the lower and upper edges, respectively, as

$$\begin{aligned} \alpha_1(t) &= 2\pi \int_0^t f_0(\tau) d\tau \\ \alpha_2(t) &= \alpha_1(t) - \phi, \end{aligned}$$

where $f_0(t)$ is the frequency contour[1] and $\phi$ is the (adjustable) phase delay between the upper and lower vocal fold edges. The intervals of $\alpha_1(t)$ and $\alpha_2(t)$ were restricted to $[0, 4\pi]$, i.e.,

$$\alpha_i(t) := \alpha_i(t) \bmod 4\pi, \quad i = 1, 2.$$

To obtain the *warped* phases $\beta_1$ and $\beta_2$, the warping function shown in Figure 3 was applied to $\alpha_1$ and $\alpha_2$. The parameter $s \in [-1, 1]$ controls the direction and degree of warping. Figures 2a and b show simulated vocal fold displacements at the lower and upper edges for $s = 0$ (no warping) and $s = 0.5$, respectively. Like the displacement curves, also the glottal area waveforms are skewed to the right for $s > 0$, and to the left for $s < 0$.

To implement double pulsing, the equation (3) for the dynamic displacement of the vocal fold edges was extended as follows. For $\beta_i \leq 1.5\pi$ and $\beta_i \geq 3.5\pi$, Eq. (3) was applied, and for $1.5\pi < \beta_i < 3.5\pi$, the displacement was modeled with a reduced amplitude as

$$\begin{aligned} x_{\mathrm{dyn},i}(y,t) = x_{\max} \sin(y\pi/L) \cdot \\ [(1 - d/2)(\sin(\beta_i(t)) + 1) - 1], \end{aligned}$$

[1] To make the voices more natural, small quasi-random fluctuations ("flutter") are added to the intended $f_0$ contour according to [2].

where $d \in [0, 1]$ is the adjustable degree of diplophonic double pulsing. Figure 2c illustrates the effect for $d = 0.5$ on the dynamic displacement. Here, for every 2nd oscillation cycle the maximum lateral (opening) displacement of the vocal fold edges is reduced by 50%. The glottal areas $A_1(t)$ and $A_2(t)$ are determined by setting all negative values of $x_1(y, t)$ and $x_2(y, t)$ to zero and adding an adjustable area $A_{\text{chink}}$ of a potential posterior glottal chink (based on [15]):

$$A_i(t) = \left[2 \int_0^L \max\{x_i(y, t), 0\} dy\right] + A_{\text{chink}}, \quad i = 1, 2. \tag{4}$$

For the synthesis of speech, the vocal fold model was implemented in the articulatory speech synthesizer VTL. At the acoustic level, this synthesizer represents all parts of the vocal system uniformly in terms of a branched tube model, i.e., a concatenation of short cylindrical tube sections [16]. The glottis is represented by two adjacent tube sections, which assume the areas $A_1(t)$ and $A_2(t)$ according to Eq. (4) and the lengths $L/2$ according to Eq. (2). Based on these tube sections, the glottal flow is calculated according to [17].

## 3. Perceptual parameter optimization

To find out which control parameter settings of the vocal fold model generate the most natural-sounding male and female voices, each of five German words was synthesized with a range of parameter value combinations. These synthetic words were used as stimuli in a perception experiment where listeners were asked to evaluate the naturalness of the synthetic voices. The five German words were "Büro" [byː.ʁoː], "Hände" [hɛn.də], "Musik" [muː.ziːk],"Thema" [teː.maː], and "Töne" [tøː.nə]. They contain mostly voiced sounds, have two syllables each, and were chosen to cover all tense German vowel qualities.

Each of the five words was synthesized in 405 variants of a male voice, and in 405 variants of a female voice. For the male voice, the "anatomic" vocal fold parameters were set to $T_0 = 4.5$ mm, $L_0 = 16$ mm, and $f_{\text{nat}} = 120$ Hz, and for the female voice, they were set to $T_0 = 4$ mm, $L_0 = 12$ mm, and $f_{\text{nat}} = 200$ Hz. The 405 variants per word and gender were generated for all combinations of

- three values for the rest displacement $x_{\text{rest},1}$ of the lower vocal fold edge (0.0, 0.3, and 0.6 mm), which also determined the posterior glottal chink area $A_{\text{chink}} = 2 \cdot x_{\text{rest},1} \cdot 4$ mm,
- three values for the rest displacement $x_{\text{rest},2}$ of the upper vocal fold edge (0.0, 0.3, and 0.6 mm),
- three values for the phase lag $\phi$ between the upper and lower edges (30°, 70°, and 90°),
- five values for the skewness $s$ of the glottal area pulses (-0.4, -0.2, 0.0, 0.2, and 0.4), and
- three values for the degree of double pulsing $d$ (0.0, 0.05, and 0.1).

Hence, there were 4050 stimuli in total for the perception experiment (5 words × 2 genders × 405 variants). All stimuli were synthesized in terms of gestural scores in the software VocalTractLab, similar to the procedure described in [18]. For each word and gender, all 405 stimulus variants were synthesized with the same phone durations and $f_0$ contour, which were reproduced from natural utterances of the respective words. The vocal tract anatomy and the articulatory phone targets of the male and female stimuli were based on a male model speaker [10] and a female model speaker [19], respectively. For all stimuli, the time constants of the articulatory gestures, which define the speed of target approximation of the model articulators, were set to 15 ms for all gestures, and the subglottal pressure was set to 800 Pa. All stimuli were saved as mono WAV files with a sampling rate of 22050 Hz and 16 bit quantization.

For the perceptual evaluation, 30 normal-hearing native German participants were recruited (15 female, 24-56 years). The task of each participant was to listen to a sequence of 405 stimuli (out of the 4050) and rate the naturalness of the voice of each stimulus. Each listener was individually seated in a quiet room in front of a computer screen. The stimuli were presented over closed headphones (K240 by AKG) via an external sound card (Aureon XFire 8.0 HD by Terratec) attached to a computer. After a stimulus was presented, the listener had to click one of four buttons with the labels "unnatural", "rather unnatural", "rather natural", "natural", corresponding to the scores 0, 1, 2, and 3, respectively. After a selection, the next simulus was automatically played. The listener had the possibility to repeat the playback of each stimulus once, if desired.

The subsets of 405 stimuli presented to the individual listeners were formed in such a way that they contained either only male or only female stimuli, and that each of the 405 combinations of vocal fold model parameters was represented in terms of one randomly selected word. Therefore, fifteen participants listened to only female stimuli, and the other 15 participants listened to only male stimuli. The stimuli were distributed in such a way within the two groups of listeners that each of the five words with each the male and the female voice was rated exactly three times for each parameter combination. Hence there were 15 responses for each parameter combination and gender. The order in which the stimuli (and so the parameter combinations) were presented was individually randomized for each listener.

## 4. Results and discussion

The experimental results are shown in Figure 4 for the male voice and in Figure 5 for the female voice. Here, each matrix shows the naturalness scores for all value combinations of two parameters of the vocal fold model, summed across all responses. This allows to assess how suitable different regions in the parameter space are for the synthesis of natural-sounding voices. For example, both male and female voices were preferred with $x_{\text{rest},1} \geq 0.3$ mm and $x_{\text{rest},2} \geq 0.3$ mm, i.e., with slightly abducted vocal folds. The phase lag $\phi$ seems to have a minor effect on the naturalness, although smaller values tended to be slightly preferred. With regard to the skewness of the glottal area pulses, moderate values of $-0.2 \leq s \leq 0.2$ were generally preferred over more extreme values. For double pulsing the values of 0.0 and 0.05 were preferred over 0.1.

The highest naturalness scores were given to the setting $\{x_{\text{rest},1} = 0.3 \text{ mm}, x_{\text{rest},2} = 0.6 \text{ mm}, \phi = 70°, s = 0.0, d = 0.05\}$ for the male voice (score 2.2 on the $0 \ldots 3$ scale), and to $\{x_{\text{rest},1} = 0.6 \text{ mm}, x_{\text{rest},2} = 0.6 \text{ mm}, \phi = 50°, s = 0.2, d = 0.05\}$ for the female voice (score 2.1). These positions in the parameter space are indicated by the white frames in the matrixes in Figures 4 and 5. Accordingly, the most natural male voice was generated with a slightly divergent prephonatory glottis ($x_{\text{rest},2} > x_{\text{rest},1}$), with symmetric glottal area pulses, and with a little double pulsing. The most natural female voice was generated with parallel prephonatory vocal fold sides ($x_{\text{rest},2} = x_{\text{rest},1}$), with slightly skewed

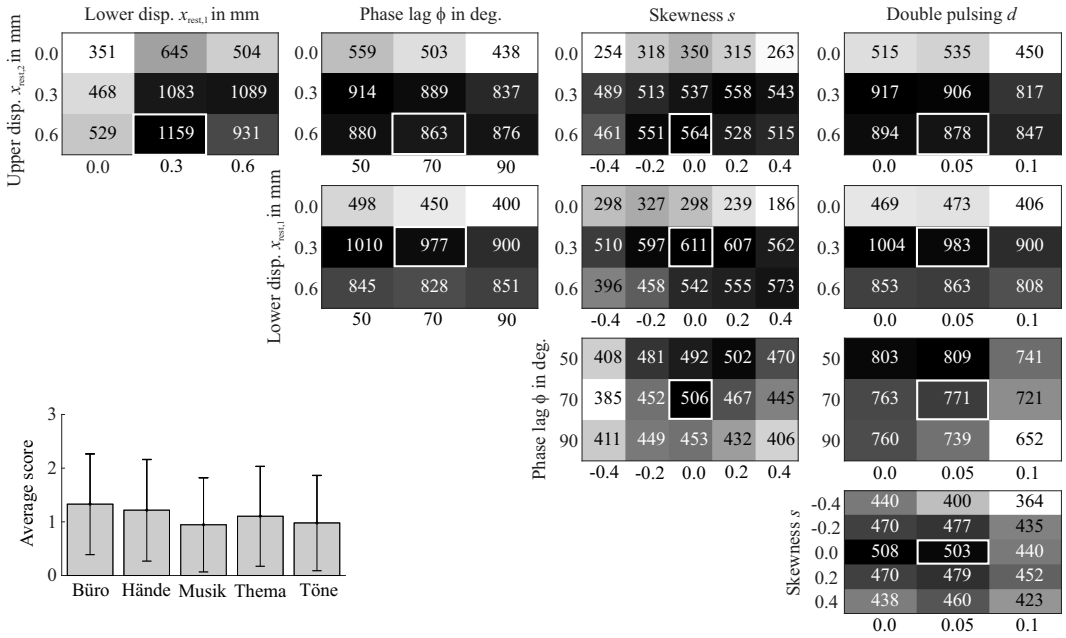


Figure 4: *Naturalness scores for the male speaker. Each matrix is a 2D projection of the five-dimensional space of the examined vocal fold model parameters and shows the total naturalness score (pooled across all other dimensions, raters, and words) for each combination of parameter values. The white frames mark the parameter combination that was perceived as most natural. The bar graph in the bottom left corner shows the average naturalness scores for the individual words (pooled across all responses).*

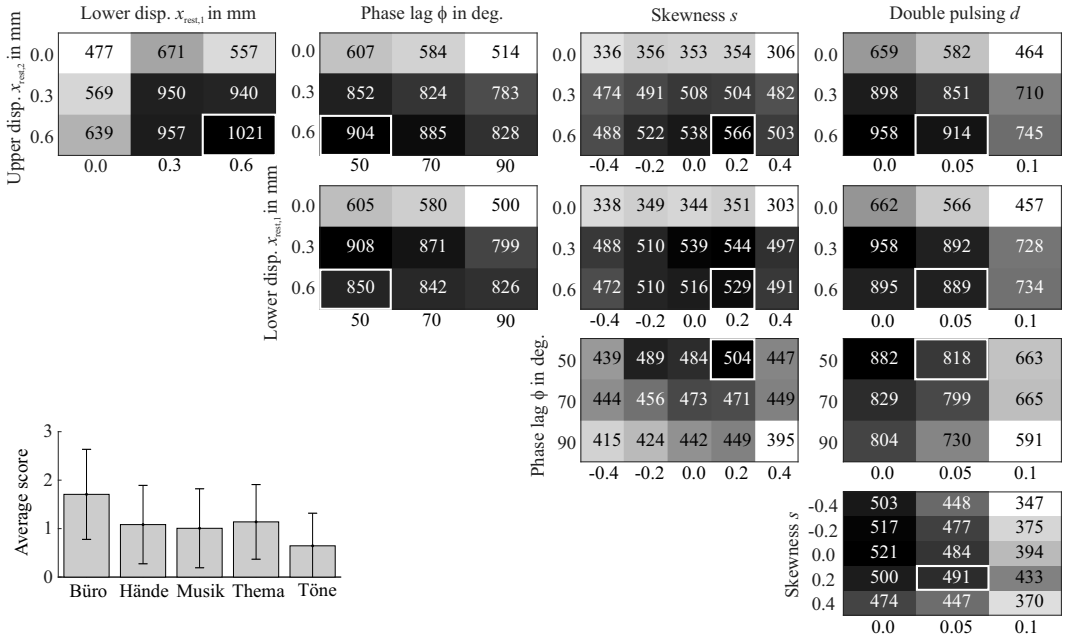


Figure 5: *Naturalness scores for the female speaker, analogous to Figure 4.*

glottal area pulses, and with a little double pulsing. The five words synthesized with these "best" settings can be listened to at `http://www.vocaltractlab.de/index.php?page=birkholz-supplements`.

## 5. Conclusions

The newly introduced parameters "skewness" and "double pulsing" do enhance the range of natural-sounding voices that can be generated with the vocal fold model, and they even contributed to the most preferred male and female voices. Hence, this study is a further step towards truly natural-sounding articulatory speech synthesis.

## 6. Acknowledgements

This study was funded by the German Federal Ministry for Economic Affairs and Energy, grant number ZF4443004BZ8.

# 7. References


[1] G. Fant, J. Liljencrants, and Q. guang Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

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

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

[5] P. Birkholz, "A survey of self-oscillating lumped-element models of the vocal folds," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2011*, B. J. Kröger and P. Birkholz, Eds. TUDPress, Dresden, 2011, pp. 47–58.

[6] 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.

[7] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, 1994.

[8] M. P. de Vries, H. K. Schutte, and G. J. Verkerke, "Determination of parameters for lumped parameter models of the vocal folds using a finite-element method approach," *Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3620–3628, 1999.

[9] N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, and A. Hirschberg, "An in vitro setup to test the relevance and the accuracy of low-order vocal folds models," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 479–490, 2007.

[10] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.

[11] ——, "VocalTractLab [software]," 2017. [Online]. Available: http://www.vocaltractlab.de

[12] P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis," in *Interspeech 2011*, Florence, Italy, 2011, pp. 2681–2684.

[13] 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.

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

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

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

[17] 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.

[18] 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," *Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1503–1512, 2015.

[19] S. Drechsel, Y. Gao, J. Frahm, and P. Birkholz, "Modell einer Frauenstimme für die artikulatorische Sprachsynthese mit VocalTractLab," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, P. Birkholz and S. Stone, Eds. TUDPress, Dresden, 2019, pp. 239–246.