Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis

Peter Birkholz, Bernd J. Kröger, Christiane Neuschaefer-Rube

Clinic of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
pbirkholz@ukaachen.de, bkroeger@ukaachen.de, cneuschaefer@ukaachen.de

Abstract

Two-mass models of the vocal folds and their variants are valuable tools for voice synthesis and analysis, but are not able to produce breathy voice qualities. The produced voice qualities usually lie between normal and pressed. The reason for this property is that the mass elements are aligned parallel to the dorso-ventral axis. Thereby, the glottis always closes simultaneously along the entire length of the vocal folds. For breathy phonation, however, the closure happens rather gradual. This article introduces a modified two-mass model with mass elements that are inclined with respect to the dorso-ventral axis as a function of the degree of abduction. In this way, the closing phase of the glottis becomes progressively more gradual when the degree of abduction is increased. This model is able to produce the continuum of voice qualities from pressed over normal to breathy voices.

Index Terms: Vocal fold model, triangular glottis, voice quality

1. Introduction

Low-dimensional lumped-mass models of the vocal folds (e.g. [1, 2, 3]) are able reproduce many properties of phonation, like self-sustained oscillations over a wide frequency range, different voice registers, and the phase differences between the upper and lower margins of the vocal folds. However, the simulation of breathy voice qualities was recognized as problematic with this class of models [4]. During breathy phonation, the vocal folds open and close more gradually than during modal phonation, and the glottis does often not close entirely during a vibration cycle [5]. Previous low-dimensional lumped-mass models cannot account for these properties, because their mass elements are aligned parallel to the dorso-ventral axis, so that the opening and closing of the vocal folds always happens simultaneously along the entire length. This gives the synthetic voice usually a pressed or normal voice quality. Vocal fold models with multiple masses along the dorso-ventral dimension (e.g. [6]) can in principle account for gradual and incomplete closure, but at the expense of considerably increased complexity. Another class of models capable to simulate gradual closures and breathy voices are geometric models [7, 8]. However, they are not self-oscillating and therefore less realistic from a physiological point of view.

In this study, we present a new two-mass model (TMM) with mass elements that are inclined with respect to the dorso-ventral axis as a function of the degree of abduction. This makes the opening and closing of the vocal folds progressively more gradual with increasing abduction and results in incomplete closure at high degrees of abduction. This allows to synthesize different degrees of breathiness besides pressed and normal voice qualities. A previous vocal fold model with inclined mass elements somewhat similar to ours was presented by Childers [9], but it was more simplified and not designed for the simulation of voice qualities. The proposed model is introduced in Sec. 2. Section 3 describes the synthesis of vowels for different degrees of abduction with both the classical TMM [1] and the new TMM. Section 4 compares the performance of both models with respect to the voice quality of synthesized vowels on the kinematic, acoustic, and perceptual level.

2. Proposed two-mass model

2.1. Mechanics

Each vocal fold is represented by two mass elements that are connected to a fixed reference frame with springs k_i and dampers r_i ($i = 1, 2$ for the lower and upper mass, respectively) and coupled to each other with an additional spring k_c (Fig. 1). We assume symmetry with respect to the midsagittal plane. In the pre-phonatory rest position, the displacements of the masses at the posterior end (at $z = 0$) are given by $x_{\mathrm{rest}1}(0)$ and $x_{\mathrm{rest}2}(0)$. When $x_{\mathrm{rest}i}(0) \geq 0$, the displacements decrease linearly towards zero at the anterior commissure, so that the pre-phonatory shape of the glottis becomes triangular, i.e. $x_{\mathrm{rest}i}(z) = x_{\mathrm{rest}i}(0)(1 - z/l)$ for $z > 0$, where l is the length of the vocal folds. In the following, we use the shorthand notation $x_{\mathrm{rest}i} \equiv x_{\mathrm{rest}i}(0)$. Let x_1 and x_2 denote the time-varying horizontal displacements of the masses. Then, the half-width of the glottis along the dorso-ventral z-axis is given by $w_i(z) = \max\{0, x_{\mathrm{rest}i}(1 - z/l) + x_i\}$ and the glottal areas between the lower and upper mass pairs are $A_i = 2\int_{z=0}^{l} w_i(z)\mathrm{d}z$. Figure 1b) and c) illustrate the shape of the glottis for different time-varying displacements but the same pre-phonatory rest displacements. When the rest displacement $x_{\mathrm{rest}i} < 0$, i.e. when the vocal folds are strongly adducted, then $A_i = \max\{0, 2l(x_{\mathrm{rest}i} + x_i)\}$, as in the classical TMM.

The equations of motion for each of the masses are

$$\begin{aligned} F_1 &= m_1\ddot{x}_1 + r_1\dot{x}_1 + k_1x_1 + k_{\mathrm{col}1}\alpha_1(x_1 + x^*_{\mathrm{rest}1}) \\ &\quad + k_\mathrm{c}(x_1 - x_2) \qquad (1) \\ F_2 &= m_2\ddot{x}_2 + r_2\dot{x}_2 + k_2x_2 + k_{\mathrm{col}2}\alpha_2(x_2 + x^*_{\mathrm{rest}2}) \\ &\quad + k_c(x_2 - x_1), \qquad (2) \end{aligned}$$

where α_i are the time-varying relative portions of the length l, where the left and right masses are in contact ($0 \leq \alpha_i \leq 1$, cf. Fig. 1b and c), and $x^*_{\mathrm{rest}i}$ are the rest displacements in the middle of these portions along the z-axis (at z^*_1 and z^*_2 in Fig. 1c). $k_{\mathrm{col}1}$ and $k_{\mathrm{col}2}$ are the spring constants of the additional springs that repel the left and right vocal folds during

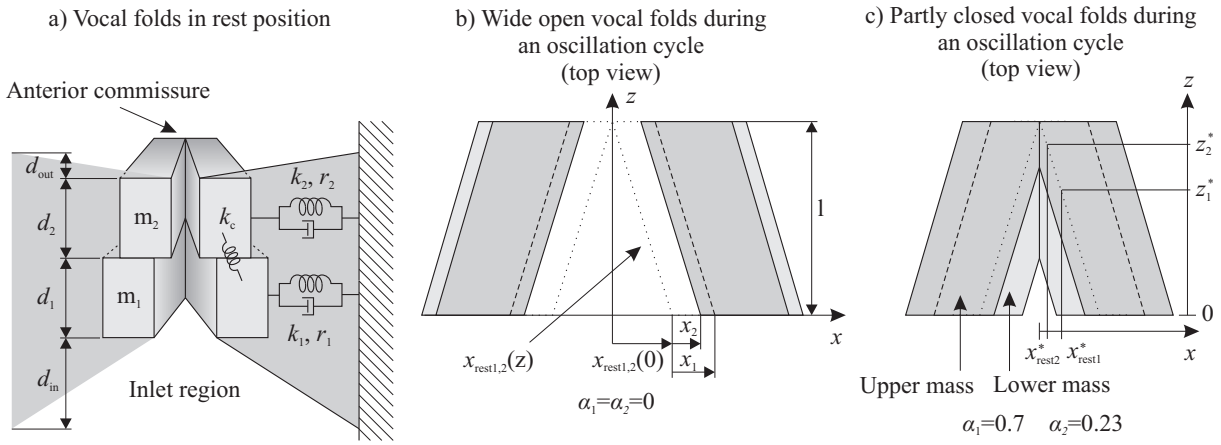

Figure 1: (a) Pseudo-3D view of the model. (b,c) Top view of the model for a wide open and a partly closed glottis during an oscillation cycle with the same pre-phonatory rest displacements. The dotted lines show the vocal fold margins in the rest position, i.e. $x_{\mathrm{rest}1,2}(z)$.

collision. For simplicity, we use linear springs in our model, because the nonlinear spring characteristics of the classical model have a relatively little effect on the oscillations according to [10, p. 916]. The external forces are

$$F_1 = P_1 d_1 l_{\mathrm{open1}} + 0.25 \cdot (P_{\mathrm{sub}} + P_1) d_{\mathrm{in}} l \quad (3)$$

$$F_2 = P_2 d_2 l_{\mathrm{open2}} + 0.25 \cdot (P_2 + P_{\mathrm{supra}}) d_{\mathrm{out}} l, \quad (4)$$

where l_{open1} and l_{open2} are the lengths of the open partitions between the upper and lower mass pairs ($0 \leq l_{\mathrm{open}i} \leq l$), i.e. the partitions where the masses are *not* in contact. d_1, d_2, d_{in}, and d_{out} are explained in Tab. 1. P_{sub}, P_1, P_2, and P_{supra} denote the subglottal pressure, the pressures between the lower and upper masses, and the supraglottal pressure, respectively. The second terms on the right-hand side of Eqs. 3 and 4 are the hinge moments on the lower and upper masses due to the mean pressures in the inlet and outlet regions. The classical TMM neglects these forces, but we consider it as more realistic to include them like e.g. [2].

Table 1: Mechanical parameters of the two-mass model. Refer to the main text for q, α_1, and α_2.

Parameter	Symbol	Value	Unit
Vocal fold length	l	$1.3 \cdot \sqrt{q}$	cm
Lower mass thickness	d_1	$0.25/\sqrt{q}$	cm
Upper mass thickness	d_2	$0.05/\sqrt{q}$	cm
Lower mass	m_1	$0.125/q$	g
Upper mass	m_2	$0.025/q$	g
Lower spring constant	k_1	$80 \cdot q$	N/m
Upper spring constant	k_2	$8 \cdot q$	N/m
Coupling spring constant	k_c	$25 \cdot q^2$	N/m
Lower collision spring cons.	k_{col1}	$240 \cdot q$	N/m
Upper collision spring cons.	k_{col2}	$24 \cdot q$	N/m
Lower damping ratio	ζ_1	$0.1 + \alpha_1$	-
Upper damping ratio	ζ_2	$0.6 + \alpha_2$	-
Inlet region length	d_{in}	4.0	mm
Outlet region length	d_{out}	1.0	mm

A control parameter q is used to adjust the fundamental frequency of the model as in [1] and scale the length and thickness of the vocal folds as in [7, p. 195]. Table 1 summarizes the parameters of the model. Their values were adopted from [1]. For the digital simulations, Eqs. 1 and 2 were approximated by a finite difference scheme analog to [1] to obtain x_1 and x_2 at a rate of 44100 Hz.

2.2. Aerodynamic-acoustic model

The model of the vocal folds was implemented in the framework of the articulatory speech synthesizer VocalTractLab (`www.vocaltractlab.de`). The synthesizer approximates the trachea, the glottis, and the vocal tract as a series of abutting cylindrical tube sections with variable lengths. Two tube sections with the time-varying lengths d_1 and d_2 and areas A_1 and A_2 represent the glottis. The aerodynamic-acoustic simulation is based on a transmission-line representation of the tube system [11, 12]. The simulation assumes a Bernoulli-type flow from the subglottal region to the glottis section with the minimum diameter and flow detachement without dynamic pressure recovery at the exit of this section. This differs from the original assumptions by Ishizaka and Flanagan [1] and conforms with more recent findings about the pressure distribution in the glottis [13]. A dipole noise source injects white noise with an amplitude proportional to the squared Reynolds number of the glottal flow right above the glottis to simulate aspiration noise.

3. Simulation experiments

At the physiological level, pressed, normal, and breathy voice qualities mainly differ in terms of the degree of glottal abduction (and hence glottal rest area), which is greatest for breathy voice, least for pressed voice, and somewhere in between for normal voice. We examined for both the classical TMM and the new TMM to what extend these models can reproduce this relationship between voice qualities and degrees of abduction by synthesizing the vowel /a/. The classical TMM was implemented along with the new model in VocalTractLab. With respect to the aerodynamic-acoustic part, it was simulated analogous to the new model.

Firstly, we determined for each model the range of rest displacements, for which a self-sustained oscillation was possible at a subglottal pressure of 1 kPa and $F_0 = 120$ Hz. These ranges were then sub-divided in 10 or 11 equally spaced values. For the classical TMM, the displacement $x_{\mathrm{rest}1,2}$ was varied from -0.15 to 0.35 mm in steps of 0.05 mm. For the new

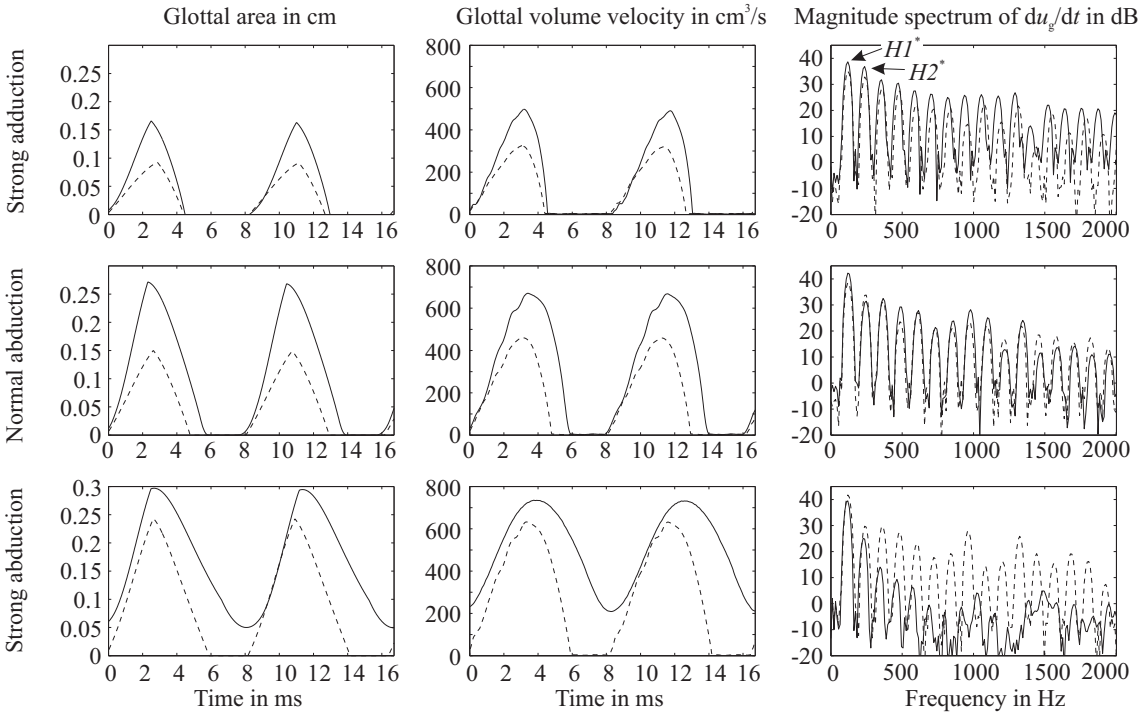

Figure 2: Simulated glottal area waveform (left), glottal flow waveform (middle column), and magnitude spectrum of the first derivative of the glottal flow (right) for strong adduction (top), normal abduction (middle row), and strong abduction (bottom row). Waveforms of the classical TMM are shown as dashed lines and those of the new TMM as solid lines. Strong adduction, normal abduction, and strong abduction correspond to $x_{\mathrm{rest1,2}} = -0.15$ (-0.2) mm, $x_{\mathrm{rest1,2}} = 0.05$ (0.2) mm, and $x_{\mathrm{rest1,2}} = 0.35$ (0.7) mm for the classical (new) model, respectively.

model, $x_{\mathrm{rest1,2}}$ was varied between -0.2 and 0.7 mm in steps of 0.1 mm. For all these degrees of abduction (x_{rest1} and x_{rest2} were set equal in all cases) we synthesized the vowel /a/ using a subglottal pressure of 1 kPa. The tension factor q was adjusted for $F_0 = 120$ Hz.

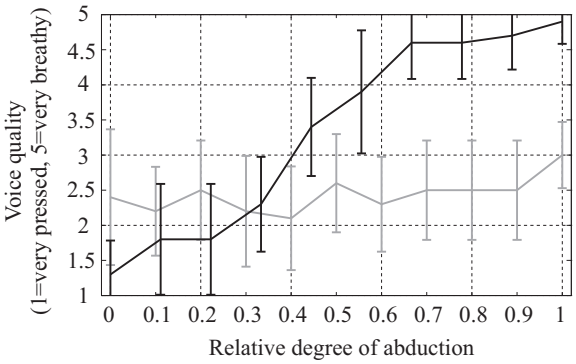

Figure 3: Averaged perceived voice qualities for different degrees of abduction of the classical TMM (gray) and the new TMM (black). The vertical bars indicate the 2σ-ranges.

There are several acoustic measures known to correlate with the degree of abduction and the perceived voice quality [15]. Some of these measures were selected to assess the ability of the two models to simulate the continuum of voice qualities. With regard to the *acoustic* performance of the models we calculated the mean open quotient OQ, speed quotient SQ, and closing quotient CQ of the simulated glottal flow waveforms of five periods in the middle of each item, the harmonic richness factor HRF, and $H_1^* - H_2^*$. OQ is defined as pulse width divided by fundamental period, SQ as rise time divided by fall time, and CQ as fall time divided by fundamental period [16]. HRF was calculated according to [16] and $H_1^* - H_2^*$ according to [5]. Furthermore, the models were evaluated at the *kinematic* level using the OQ, SQ, and CQ of the projected glottal area waveform, i.e. $\min\{A_1(t), A_2(t)\}$. Finally, the models were evaluated *perceptually*. Ten listeners were asked to rate the voice quality of each item on a discrete scale from 1 (very pressed) to 5 (very breathy). All vowel stimuli of both models were presented over earphones to one subject after the other in a different randomized order for each subject. Each stimulus could be repeated once on request. The subjects were not trained before the task but asked to judge the stimuli according to their associations with the according voice qualities.

4. Results and discussion

The results are shown in Figures 2, 3, and 4. For both the classical and the new model, all kinematic and acoustic data change into the expected direction when the degree of abduction is increased. However, the amount of change varies for most parameters between the two models. The arrows at the left and the right side of the upper two panels in Fig. 4 show exemplarily the values measured for male subject 1 in [14] for pressed and breathy phonation, respectively. They indicate that the new model generates glottal area and flow waveforms for breathy phonation that come closer to these real values than the classi-

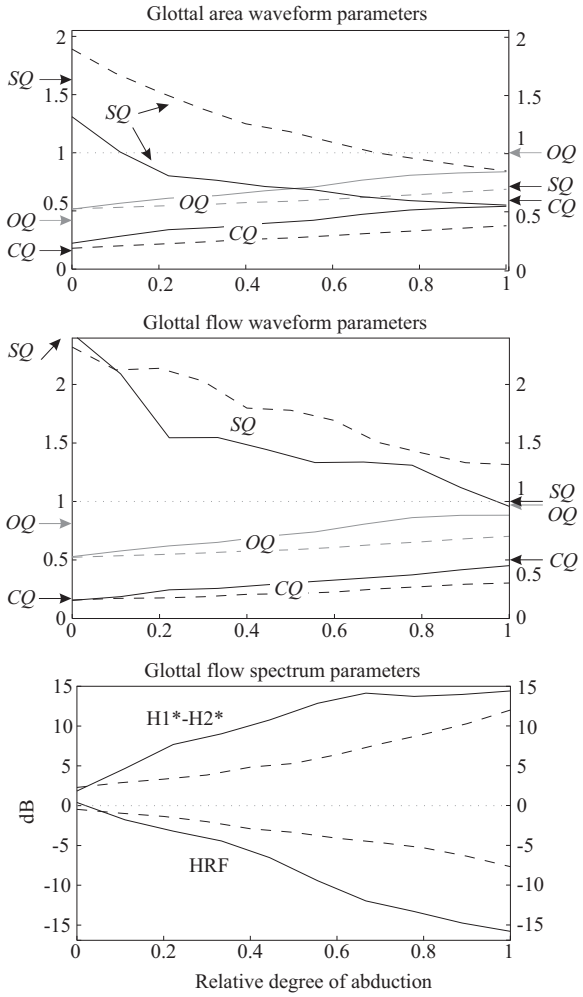

Figure 4: Glottal area and flow waveform time parameters (open quotient OQ, closing quotient CQ, speed quotient SQ) and glottal flow spectrum parameters ($H1^* - H2^*$ and HRF) as a function of the relative degree of abduction for the classical model (dashed lines) and the new model (solid lines). The arrows on the left and right side of the graphs indicate the *measured* values for OQ, SQ, and CQ of subject 1 in [14] for pressed and breathy voice, respectively.

cal model in most cases, especially for CQ, which was shown to be the most effective time-domain parameter to characterize the considered voice qualities [16]. The values for the second male subject measured in [14] (not shown) are very similar to those of the first subjects and support the results. Note furthermore that a glottal leak remains for breathy phonation with the new model, whereas the classical model always closes completely (bottom row of Fig. 2). The glottal leak gives rise to stronger aspiration noise that supports the perception of breathiness. The glottal flow spectrum parameters $H1^* - H2^*$ and HRF were also found to effectively characterize voice quality [15]. The higher range of these parameters over the different degrees of abduction for the new model also supports its ability to better simulate the differences in voice quality. In the perception test, all stimuli synthesized with the classical model were perceived quite undifferentiated as normal to slightly pressed (Fig. 3). In contrast, the stimuli of the new model cover the whole continuum of voice qualities. As in reality, the perceived voice quality correlates with the degree of abduction. In conclusion, the proposed model allows the synthesis of a continuum of voice qualities that was previously not possible with this type of model. This was achieved without a significant increase of model complexity. The proposed modifications can also be applied to more sophisticated low-dimensional lumped-mass models, e.g. body-cover models [3].

5. References

[1] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *The Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[2] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, and A. Hirschberg, "A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prothesis design," *Acta Acustica united with Acustica*, vol. 84, pp. 1135–1150, 1998.

[3] I. T. Tokuda, M. Zemke, M. Kob, and H. Herzel, "Biomechanical modeling of register transitions and the role of vocal tract resonators," *Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1528–1536, 2010.

[4] F. Avanzini, S. Maratea, and C. Drioli, "Physiological control of low-dimensional glottal models with applications to voice source parameter matching," *Acta Acustica united with Acustica*, vol. 92, pp. 731–740, 2006.

[5] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[6] I. R. Titze, "The human vocal cords: a mathematical model," *Phonetica*, vol. 28, pp. 129–170, 1973.

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

[8] B. Cranen and J. Schroeter, "Physiologically motivated modelling of the voice source in articulatory analysis/synthesis," *Speech Communication*, vol. 19, pp. 1–19, 1996.

[9] D. G. Childers, D. M. Hicks, G. P. Moore, and Y. A. Alsaka, "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1309–1320, 1986.

[10] K. Ishizaka and J. L. Flanagan, "Acoustic properties of longitudinal displacement in vocal cord vibration," *The Bell System Technical Journal*, vol. 56, no. 6, pp. 889–918, 1977.

[11] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Interspeech 2004*, Jeju Island, Korea, 2004, pp. 1125–1128.

[12] 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.

[13] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1249–1260, 1995.

[14] 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.

[15] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1410–1413.

[16] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatrica et Logopaedica*, vol. 48, pp. 240–254, 1996.