Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds

Peter Birkholz, Bernd J. Kröger, and Christiane Neuschaefer-Rube
Clinic for Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen and RWTH Aachen University
Pauwelsstrasse 30, 52074 Aachen, Germany
pbirkholz@ukaachen.de, bkroeger@ukaachen.de, cneuschaefer@ukaachen.de

Abstract

A modified two-mass model of the vocal folds is introduced and applied to the articulatory synthesis of words in six voice qualities. The modified two-mass model uses mass elements that are inclined, instead of parallel, with respect to the dorsoventral axis as a function of the degree of abduction. This allows to produce the continuum of voice qualities from pressed over modal to breathy voices. Furthermore, the model is extended by a variable posterior chink to represent the space between the arytenoid cartilages, like in whispery phonation. Five words were each synthesized with different glottal settings to simulate modal voice, pressed voice, breathy voice, whispery voice, vocal fry, and falsetto. The stimuli were judged by a group of listeners in a forced-choice experiment with respect to the perceived voice qualities. Apart from whispery voice, which was more often judged as breathy than whispery, all voice types were identified as intended with probabilities between 50% (modal voice) and 94% (falsetto), which are well above the chance level of 16.1%.

Keywords: Vocal fold model, two-mass model, triangular glottis, voice quality, articulatory speech synthesis.

1. Introduction

People can speak and sing with a variety of voice qualities. Modal voice is the most common voice quality in speaking. However, other qualities like pressed, creaky, breathy, or whispery phonation are often used in a controlled way to signal paralinguistic information or, in some languages, phonological contrast [1]. In this article, we use the term voice quality interchangeably with phonation type to refer to a state of the glottis. In a broader sense, voice quality may include supraglottal articulatory settings like nasality, which are not considered in this study [2].

Voice qualities actually form a continuum rather than discrete categories. The continuum from pressed over modal to breathy phonation is, for example, associated with an increasing degree of glottal abduction. In conversational speech, the voice quality is varied consistently along this continuum in much the same way, but independently of, fundamental frequency [3]. It was shown to be correlated with a variety of different factors. Campbell and Mokhtari [3] found significant correlations with the interlocutor, speaking style, and speaker intention. For example, talking to children exhibits a higher degree of breathiness than talking to friends, and *giving* information happens with a more tense voice than *requesting* information. Likewise, the phonation type varies with focus, i.e., more prominent syllables are produced with a less tense voice [4]. Voice quality is also an important aspect of emotional speech (e.g. [5, 6]). Gobl et al. [6] demonstrated that voice quality changes alone can impart very different emotional overtones to a message, for example an angry overtone with a tense voice.

The above examples illustrate the importance of voice quality for speech production and perception. However, with regard to speech synthesis technology, voice quality is difficult to control independently in current corpus-based synthesis systems. This seriously complicates the task to generate truly natural conversational speech. Formant synthesis [7] is more flexible in this respect, because it allows to manipulate the glottal flow signal independently from the vocal tract filter function. However, the source models of formant synthesizers specify only acoustic properties of the voice source and are not related to the underlying physiological mechanism. This makes them difficult to control. Furthermore, they do not account for source-filter interaction and still sound very unnatural.

Physiological models of the vocal folds are in principle best qualified for the comprehensive synthesis of voice. A good compromise between realism and complexity are the low-dimensional lumped-mass approximations of the vocal folds, namely the one-mass models (e.g. [8, 9]), two-mass models (e.g. [10, 11]), three-mass models (e.g. [12]), and lumped body-cover models [13, 14]. However, a common restriction of all these models is their approximation as two-dimensional structures in the coronal plane. The third dimension is only taken into account by giving the vocal folds a certain length, which makes the glottal slit rectangular.

p3s 2011, March 14-15, 2011, Vancouver, BC, CA.

Thereby, the opening and closing of the vocal folds always happen simultaneously along the entire length. This gives the synthetic voices usually a pressed to modal voice quality. A gradual closing and a possibly incomplete closure of the glottis, as for breathy voices, is intricate with these models. The representation of a vocal fold by multiple masses in the dorso-ventral direction can solve this issue (e.g. [15, 16]), but renders the model more complicated and computationally more expensive.

In this study, we present a modified two-mass model of the vocal folds with mass elements that are inclined with respect to the dorso-ventral axis as a function of the degree of abduction. This makes the glottal slit triangular in the rest position with the acute angle at the anterior commissure. The mass elements oscillate in the lateral direction, as in the classical two-mass model (TMM), but keep the angle with the dorso-ventral axis constant. In this way, the opening and closing of the glottis becomes progressively more gradual with increasing abduction and allows to simulate oscillation patterns that are typical for pressed, modal, and breathy voices. A similar idea was presented by Childers [17] to reproduce the zipper-like opening and closing of the vocal folds, but it was neither elaborated with a self-oscillating model nor with respect to voice quality variations. Besides the above modification, a posterior chink was added to the model to represent any non-oscillating glottal space between the arytenoid cartilages. This was first proposed by Kröger [18] in combination with a self-oscillating TMM to simulate breathy phonation. With both modifications, the adductive tension of the vocal folds can be independently adjusted from the medial compression and thus allows to define glottal settings for a variety of voice qualities.

The proposed TMM is described in detail in the following section [1]. In Sec. 3 we define laryngeal settings for the vocal fold model in terms of control parameters for six voice qualities, namely modal voice, pressed voice, breathy voice, whispery voice, vocal fry, and falsetto. Acoustic and kinematic properties for each setting will be presented and compared with real data. Section 4 describes a perceptual identification tests with synthetic words in the different voice qualities and its results. We conclude with a discussion in Sec. 5.

2. Vocal fold 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 mid-sagittal plane. In the pre-phonatory rest position, the displacements of the masses at the posterior end are given by $x_{\mathrm{rest}1}$ and $x_{\mathrm{rest}2}$. When $x_{\mathrm{rest}i} \geq 0$, the displacements decrease linearly towards zero at the anterior commissure, so that the pre-phonatory shape of the glottis becomes triangular. Let x_1 and x_2 denote the time-varying horizontal displacements of the masses and l the length of the vocal folds. 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 $A_{\mathrm{folds}i} = 2\int_{z=0}^{l} w_i(z)\mathrm{d}z$. Figure 1b) and c) illustrate the shape of the glottis for different displacements. When the rest displacement $x_{\mathrm{rest}i} < 0$, i.e. when the vocal folds are strongly adducted, then $A_{\mathrm{folds}i} = \max\{0, 2l(x_{\mathrm{rest}i} + x_i)\}$, as in the classical TMM. The potential posterior chink between the aryteniod cartilages is represented by its area A_{chink}, which is added to $A_{\mathrm{folds}1}$ and $A_{\mathrm{folds}2}$ to obtain the total glottal areas $A_i = A_{\mathrm{folds}i} + A_{\mathrm{chink}}$.

The equations of motion for each of the masses are

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

where α_i are the 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 (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 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 [19, p. 916]. The external forces are

$$\begin{aligned} F_1 &= P_1d_1l_{\mathrm{open}1} + 0.25\cdot(P_{\mathrm{sub}} + P_1)d_{\mathrm{in}}l \qquad (3) \\ F_2 &= P_2d_2l_{\mathrm{open}2} + 0.25\cdot(P_2 + P_{\mathrm{supra}})d_{\mathrm{out}}l, \qquad (4) \end{aligned}$$

where $l_{\mathrm{open}1}$ and $l_{\mathrm{open}2}$ 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. [11].

A control parameter q is used to adjust the fundamental frequency of the model as in [10] and scale the length and thickness of the vocal folds as in [20, p. 195]. Ta-

[1] A previous version of the model without the posterior chink was presented in Birkholz, Kröger, Neuschaefer-Rube: "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis", under review for the Journal of the Acoustical Society of America.

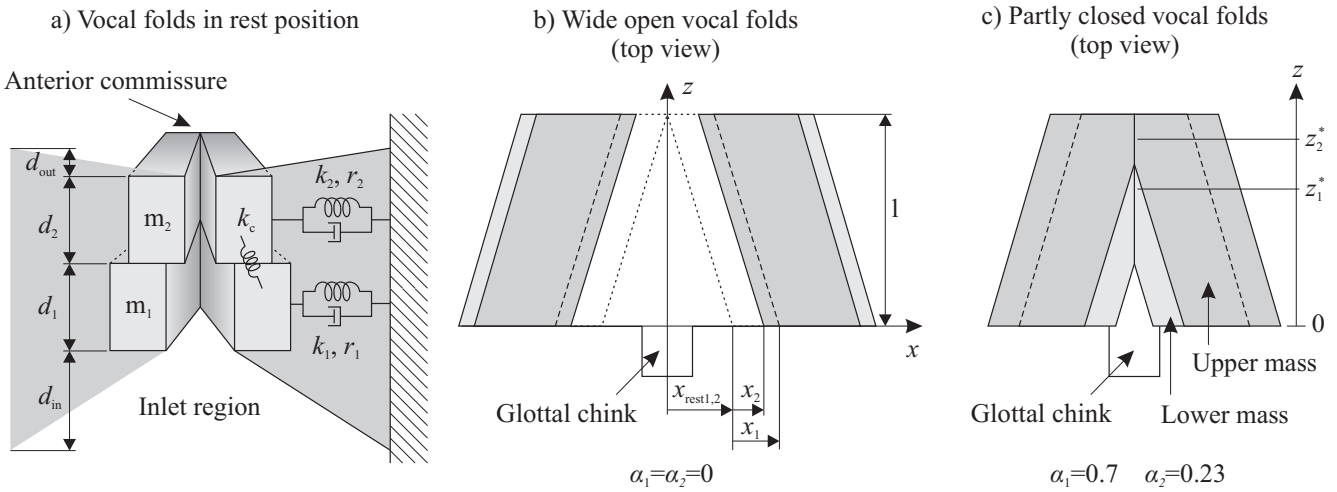

Figure 1. Pseudo-3D view of the model (a) and top view of the model for wide open (b) and partly closed (c) vocal folds. The posterior chink is not drawn in (a) for a better visibility of the other parts.

Table 1. Physiological parameters of the two-mass model.

Parameter	Symbol	Value	Unit
Vocal fold length	l	$13 \cdot \sqrt{q}$	mm
Lower mass thickness	d_1	$2.25/\sqrt{q}$	mm
Upper mass thickness	d_2	$0.75/\sqrt{q}$	mm
Lower mass	m_1	$0.1125/q$	g
Upper mass	m_2	$0.0375/q$	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	k_{col1}	$360 \cdot q$	N/m
Upper collision spring	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

ble 1 summarizes the parameters of the model that represent a male voice. The numerical constants in the table are considered as fixed *speaker-specific parameters*, independent of the phonation type. Apart from d_1, d_2, m_1, m_2, and k_{col1}, they correspond to the standard values of the classical TMM [10]. The new values for d_1, d_2, m_1, m_2, and k_{col1} are motivated in Sec. 3.6. The *control parameters* of the model are the rest displacements at the vocal processes $x_{\mathrm{rest}i}$, the tension parameter q, the posterior chink area A_{chink}, and the subglottal pressure P_{sub}. They define the momentary rest state of the vocal folds and are specified for the different voice qualities in the next section.

For the digital simulations, Eqs. (1) and (2) were approximated by a finite difference scheme analog to [10] to obtain x_1 and x_2 and hence A_1 and A_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 [21, 22]. 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. A noise pressure source with an amplitude proporional to the squared Reynolds number of the glottal flow is added right above the glottis to simulate aspiration noise.

3. Simulation of voice qualities

3.1. Method

The voice qualities considered in this study are modal voice, pressed voice, breathy voice, whispery voice, vocal fry, and falsetto. Unfortunately, there is neither a general agreement on how to name and describe different voice qualities nor a complete understanding of the associated laryngeal settings [23, 2]. Therefore, a short characterization of the voice qualities will be given in the following subsections and laryngeal settings in terms of control parameters for the vocal fold model will be defined. The subglottal pressure was set to 1 kPa for all voice qualities. Although there are indications that the subglottal pressure is typically higher or lower for some voice qualities, we refrained from modifying this parameter due to a lack of quantitative comparative data. Furthermore, the upper and lower mass elements were always displaced equally in the rest positions

($x_{\mathrm{rest1}} = x_{\mathrm{rest2}}$).

For an objective evaluation of the voice quality synthesis, the glottal flow waveform was analyzed during the synthesis of the vowel /a/ for each glottal setting defined below. The following parameters were obtained (with reference to Fig. 3b): fundamental frequency $F_0 = 1/T_0$, open quotient $OQ = (t_3 - t_1)/T_0$, the shape quotient $SQ = (t_2 - t_1)/(t_3 - t_2)$, the closing quotient $CQ = (t_3 - t_2)/T_0$, the minimum flow $u_{\min} = min\{u_{\mathrm{g}}(t)\}$, and the mean flow $u_{\mathrm{mean}} = \int_{t_1}^{t_4} u_{\mathrm{g}}(t)\mathrm{d}t/T_0$, where $T_0 = t_4 - t_1$ and $u_{\mathrm{g}}(t)$ is the glottal volume velocity. The times $t_1 \ldots t_4$ were obtained from a smoothed glottal flow waveform to avoid distortions from ripples in the flow as in Fig 3b) and f). A moving average filter with a length of 1.36 ms was used.

3.2. Modal voice

According to Laver [2], modal voice is characterized by a regular and periodic vibration pattern, where both the ligamental and cartilaginous part of the vocal folds vibrate as a single unit. The adductive tension, medial compression, and longitudinal tension are all moderate. There is no audible friction. In terms of control parameters, we set the posterior chink area to zero (no constant leak) and $x_{\mathrm{rest1,2}}$ to 0.12 mm. The tension parameter q was set to 0.78 for a typical F_0 of about 100 Hz. The corresponding glottal flow waveform is shown in Fig 3a).

3.3. Pressed voice

Pressed voice is mainly characterized by a higher medial compression than modal voice [24] and in consequence by a higher degree of vocal fold adduction. Therefore, $x_{\mathrm{rest1,2}}$ were set to -0.15 mm, which is only slightly more than the limit of -0.2 mm, where the model ceases to oscillate. The other parameters were set as for modal voice. The corresponding glottal flow waveform is shown in Fig 3b).

3.4. Breathy voice

The laryngeal characteristics of breathy voice are minimal adductive tension and weak medial compression [2]. Therefore, the degree of abduction is higher than for model voice. The vocal folds are vibrating, but without ever closing. To model the increased degree of abduction, $x_{\mathrm{rest1,2}}$ were set to 0.35 mm. Furthermore, a posterior chink of 3 mm^2 was added. Figure 3c) shows that in this way the mass elements come close to the glottal midline during vibration, but never touch at the level of the vocal processes. This incomplete closure in addition to the glottal chink gives the projected glottal area an offset of about 3.4 mm^2. Besides other effects on the pulse shape, this results in a glottal flow offset of 144 cm^3/s.

3.5. Whispery voice

Whispery voice is characterized by a low adductive tension and a high to moderate medial compression [2]. The consequence of the low adductive tension is a triangular opening between the arytenoid cartilages, comprising about 1/3 of the full length of the glottis. For the simulation, the triangular opening is represented by a chink area of 9 mm^2. The medial compression is assumed to be high and so $x_{\mathrm{rest1,2}}$ were set to -0.15 mm, like for pressed voice. Hence, with respect to control parameters, pressed and whispery voice only differ in the size of the posterior chink. The corresponding glottal flow waveform is shown in Fig 3d). Like for breathy voice, both the projected glottal area and the glottal flow have a considerable offset. However, the waveforms clearly differ between breathy and whispery simulations, especially with respect to the open quotient.

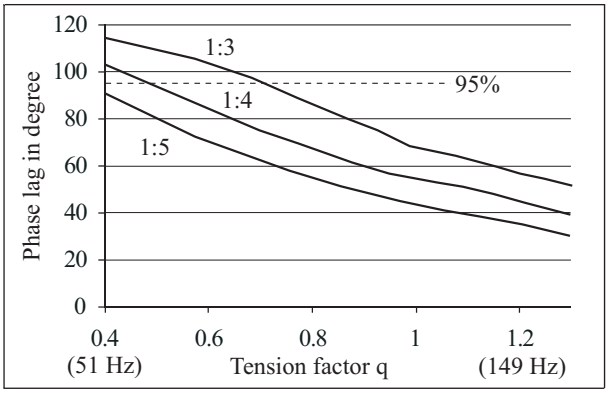

Figure 2. Phase lag between the lower and upper mass element as a function of the tension factor q for three mass and thickness ratios between the upper and lower mass elements. For phase lags greater than 95°, secondary pulses appear in the glottal area waveform.

3.6. Vocal fry

Vocal fry is characterized by periodic vocal fold vibrations at very low frequencies, ranging from approximately 20 to 70 Hz with a mean of 50 Hz [25]. The open quotient of the glottal flow waveform is lower than for other voice types. While Childers and Lee [26] measured values around 45% for OQ, values as low as 13% were reported [27]. A common observation in this mode of phonation is that within one glottal period, the vocal folds may separate twice, or even three times, in quick succession followed by a long closed phase [2, 27]. This phenomenon is called double pulsing or dicrotic dysphonia.

With regard to the control parameters of the vocal fold model, the low F_0 is easily to obtain with a low tension parameter. We chose $q = 0.34$ for this simulation for a F_0 of about 43 Hz. However, with the standard values for the speaker-specific model parameters (cf. [10]), we found no way to simulate the double pulsing oscillation pattern. To obtain such a pattern, we modified them in such a way that the phase lag between the lower and upper mass elements was increased, as illustrated by their displacement curves in Fig. 3e). This creates two time intervals within one glottal cycle with a gap between both the lower and upper mass

pairs, as the projected glottal area curve shows – one primary pulse and a smaller secondary pulse. An effective way to increase the phase lag is to change the distribution of the vocal fold mass. In the classical TMM, the total mass and thickness of one vocal fold are $m_1 + m_2 = 0.15$ g and $d_1 + d_2 = 3$ mm. They are partitioned at a ratio of 1:5, i.e., the lower mass and thickness is five times that of the upper mass and thickness. Figure 2 illustrates what happens when this ratio is changed to 1:4 or 1:3 – the phase lag is increased by an approximately constant offset over the whole range of frequencies. For phase lags of more than 95%, we observed the occurence of secondary pulses. Hence, a partitioning of the total mass and thickness of 1:4 or 1:3 creates a double pulse pattern at the very low frequencies of vocal fry, but a single pulse pattern for higher frequencies of the other voice qualites. We adopted a mass and thickness ratio of 1:3 for this study (cf. Tab. 1). Furthermore, we found that a moderate increase of the lower collision spring constant further supports the occurence of double pulsing at low frequencies and it was hence increased by 50% with respect to the standard value. The glottal rest shape for vocal fry was set to a moderate degree of abduction with $x_{\mathrm{rest}1,2} = 0.15$ mm and the posterior chink area was set to zero.

3.7. Falsetto

Falsetto is characterized by regular vocal fold vibrations at noticeably higher frequencies than modal voice. The glottis often remains slightly open resulting in an audible friction noise component [2]. For the simulation, q was set to 2.0 for a frequency of 225 Hz (the upper frequency limit of the proposed model for self-sustained oscillations is 260 Hz). The generation of friction noise was accounted for by a moderate abduction with $x_{\mathrm{rest}1,2} = 0.15$ mm and a small posterior chink of 1 mm.

3.8. Comparison with real glottal flow data

The glottal settings for the voice qualities are summarized in Tab. 2 together with the kinematic and acoustic measurements. For comparison, corresponding data from real voices are given for F_0, OQ, SQ, and CQ. The data for modal, pressed, and breathy voices were adapted from [24] and are the mean values of five male speakers. The data for vocal fry and falsetto were summarized as "typical values" in [26]. The simulated data for CQ of modal, pressed, and breathy phonation come close to corresponding human data. CQ is known to be the most effective time-domain parameter of the glottal flow for the discrimination of breathy, modal, and pressed voices [24]. For OQ and SQ, the absolute deviations are obviously greater than for CQ. However, when the voice qualities are ordered as in Fig. 4, similar *relative* changes from one voice quality to the other can be observed for both OQ and SQ between the human data and the simulated data. With regard to the mean glottal flow rates, our data in Tab. 2 are generally compatible with the numbers reported by Ladefoged [23]. For a subglottal pressure of about 700 Pa, he measured about 120 cm^3/s for modal voice, about 400 cm^3/s for whispery voice, and values greater than 500 cm^3/s for breathy voice. Besides these objective glottal flow inspections, the synthesized vowels were informally judged to be acceptable representatives of the considered voice qualities by the authors of the study. Therefore, the glottal settings were adopted for the synthesis of words for the identification experiment.

Table 2. Characteristics and parameters for the phonation types.

	Modal	Pressed	Breathy	Whispery	Vocal fry	Falsetto
Control parameters of the vocal fold model						
Q	0.78	0.78	0.78	0.78	0.34	2.0
P_{sub} (kPa)	1.0	1.0	1.0	1.0	1.0	1.0
$x_{0,1}$ (mm)	0.12	-0.15	0.35	-0.15	0.15	0.15
A_{chink} (mm^2)	0	0	3	9	0	1
Kinematic and acoustic simulation results						
F_0 (Hz)	103	110	95	81	43	225
Phase lag (deg)	89	88	79	63	116	17
u_{min} (cm^3/s)	0	0	144	347	0	61
u_{mean} (cm^3/s)	217	150	414	448	202	300
OQ (%)	61	54	87	63	48	88
SQ (%)	130	149	77	101	192	134
CQ (%)	26	22	49	31	16	38
Acoustic features of real voices						
F_0	102	105	103	-	$\leq$52	$\geq$275
OQ (%)	84	70	96	-	45	99
SQ (%)	215	218	115	-	350	150
CQ (%)	27	22	45	-	-	-

4. Perceptual identification task

A perceptual identification task based on words was conducted to determine if listeners are able to differentiate the voice qualities defined in the previous section.

4.1. Material

Each of the five German words "Ananas" ([ʔananas]), "Banane" ([banaːnə]), "Mandarine" ([mandariːnə]), "Melone" ([meːloːnə]), and "Orange" ([oːraŋʒə]) was synthesized with modal voice, pressed voice, breathy voice, whispery voice, vocal fry, and falsetto using the articulatory synthesizer VocalTractLab (www.vocaltractlab.de). First, gestural scores were manually created for the detailed re-synthesis of natural recordings of the five words spoken in a neutral style. In these scores, the phonation gestures that specify the glottal shape in terms of glottal abduction and posterior chink area were then replaced by the settings for the voice qualities defined in Tab. 2. For modal, pressed, breathy, and whispery voice, the pitch contour was reproduced from the natural recordings. The average F_0 was 108 Hz. For vocal fry stimuli, the original pitch contour was lowered by one

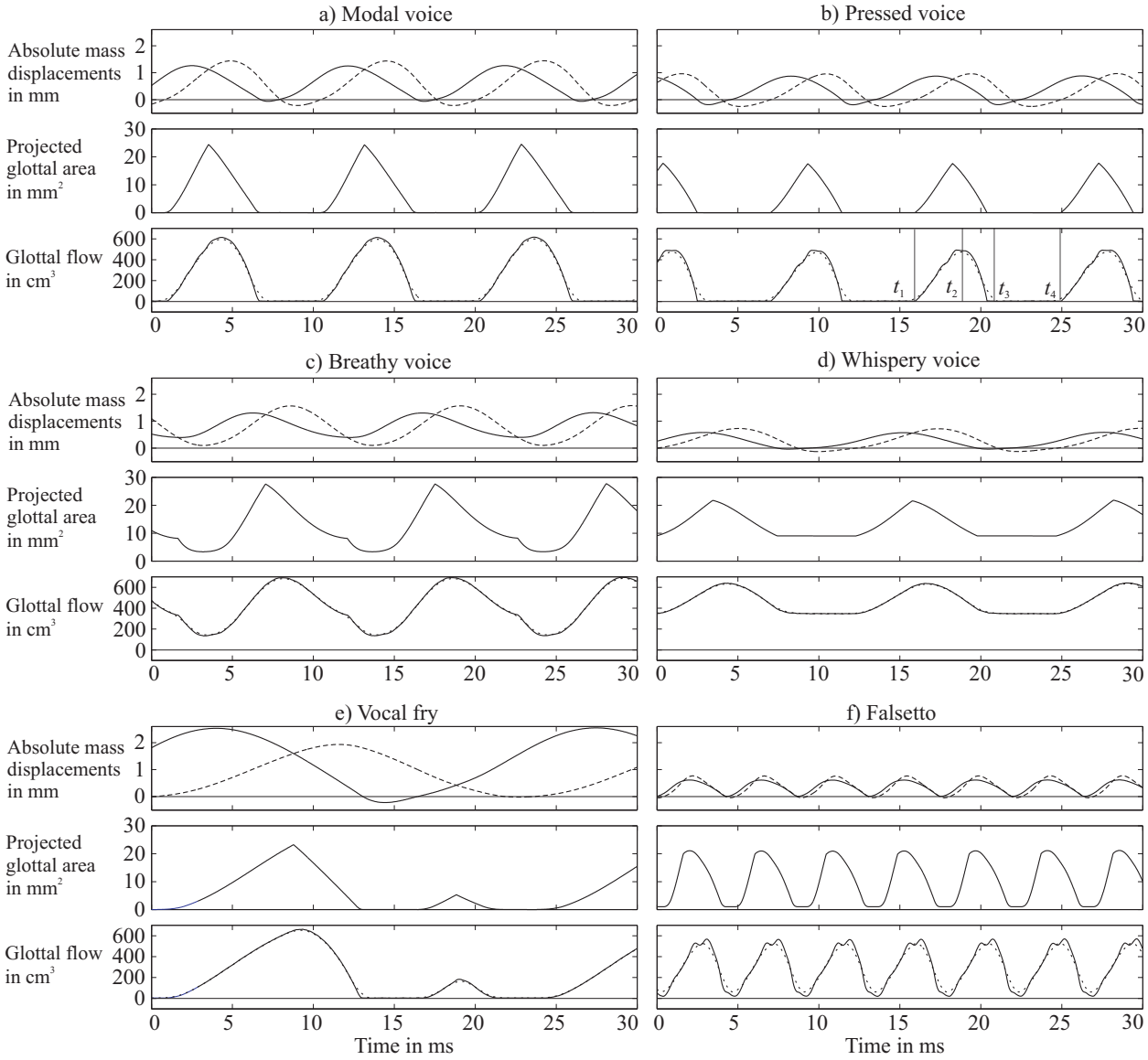

Figure 3. Absolute displacements of the lower (solid line) and upper (dashed line) mass elements, the projected glottal area, and the glottal flow for simulations of the six voice types. The vocal tract was shaped for the vowel /a/. The dashed lines in the glottal flow panels are the smoothed waveforms used for the determination of the open quotient, the shape quotient, and the closing quotient.

octave. Likewise, for falsetto stimuli, the pitch contour was raised by 14 semitones. A higher pitch for falsetto would actually be more realistic, but we had to make sure that the highest F_0 value of the contour stayed below the upper frequency limit of the vocal fold model of 260 Hz. The supraglottal articulation and the gestural timing was kept equal in all voice quality variants of a word. The synthetic samples can be downloaded from `www.vocaltractlab.de` under "Supplemental material".

4.2. Subjects and task

A set of 36 stimuli war prepared for the identification task: one stimulus for each word and voice quality plus one randomly selected duplicate word of each voice quality for an intra-judge reliability test. These stimuli were presented to a total of 18 listeners (7 males and 11 females) in groups of one to four people over loudspeaker from a distance of about two meters. For each group, the order of the stimuli was re-randomized to minimize contectual effects in the ratings. Each stimulus was presented three times in succession to the subjects with short pauses inbetween. The subjects had to estimate the voice quality of each stimulus in a forced-choice mode. Because we knew that not all subjects would be familiar with the phonetic terms for the voice qualities, we decided to label them on the question-

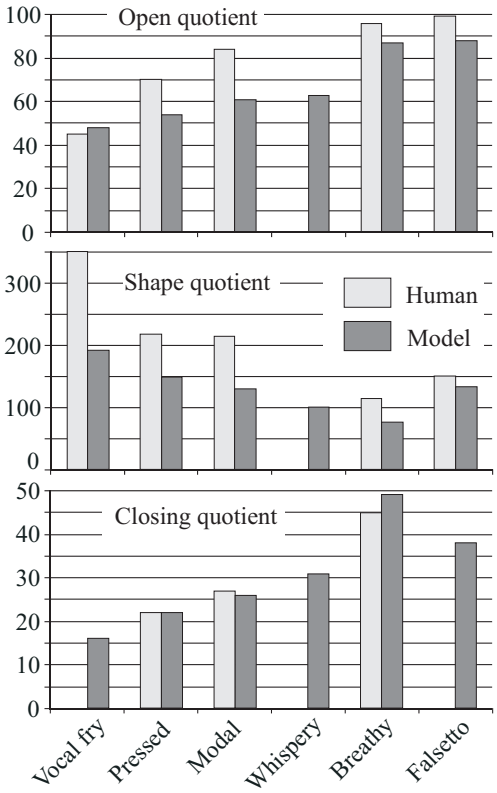

Figure 4. Open quotients, shape quotients, and closing quotients of the glottal flow waveforms of humans and model simulations according to Tab. 2.

naires with the following descriptive terms: "normal" for modal voice, "pressed" for pressed voice, "soft breathy" for breathy voice, "hard breathy" for whispery voice, "low and creaky" for vocal fry, and "high and thin" for falsetto. Before the task, several examples of each voice quality spoken by the authors were presented to the subjects for training.

4.3. Results

Table 3 presents the results of the identification task in terms of a confusion matrix with scores given in percent. Vocal fry and falsetto stimuli were perceived as intended in over 90% of the responses. Pressed stimuli were recognized rather reliably, too, with a score of 68%, and misclassified as modal in 19% of the cases. Modal and breathy stimuli were perceived as intended in 50% and 53% of the cases, respectively. Modal stimuli were most often misclassified as pressed (33%), and breathy stimuli as modal (23%) or whispery (18%). Whispery voice stimuli were most often perceived as breathy (69%), and in 31% of the cases as intended. Interestingly, none of the whispery voice stimuli was misclassified as modal or pressed, like some breathy stimuli were. The intra-judge reliability was determined by comparing the responses for the stimuli that occured twice and turned out to be 68%.

Table 3. Confusion matrix for the perceptual identification of the phonation types. All numbers are in percent (rounded). The highest value in each row is printed in bold letters.

		Perceived					
		Modal	Pressed	Breathy	Whispery	Vocal fry	Falsetto
Intended	Modal	**50**	33	0	8	7	1
	Pressed	19	**68**	1	6	6	1
	Breathy	23	6	**53**	18	1	0
	Whispery	0	0	**69**	31	0	0
	Vocal fry	0	8	0	1	**91**	0
	Falsetto	1	3	0	3	0	**94**

5. Discussion

The synthesis of vowels in Sec. 3 demonstrated that the presented two-mass modal is able to map physiological characteristics of different voice qualities to corresponding acoustic characteristics that generally agree with previous data. Furthermore, the perceptual identification task showed that the considered glottal settings were perceived as different voice qualities. For modal voice, pressed voice, breathy voice, vocal fry, and falsetto, the voice qualities were perceived as intended with probabilies between 50% and 94%, which are well above the chance level of 16.1%. Only whispery voice was misclassified in the majority of cases as breathy voice, which we primarily attribute to the close auditory relationship between both voice types. This close relationship may be the reason why in fact no language of the world contrasts breathy and whispery voice phonologically [23], while other voice qualities are used contrastively. The high identification scores for vocal fry and falsetto are probably due to their special F_0 as a major perceptual cue.

Let us now outline some limitations of the current study. First, the identification task is difficult per se, because people are not used to think in terms of voice quality. They rather associate voice quality with speaking styles, moods, or emotions. Furthermore, each person seems to have somewhat different associations with certain terms for voice qualities. Some terms may as well be associated with a changed supraglottal articulation, like a narrowed pharyngeal region for a "pressed voice". However, in this study, supraglottal articulation was exactly the same for all voice quality variants of a word. We are aware that the short "training" with natural utterances in different voice qualities that we did before the perception task could not completely unify the associations. In this light, it would be interesting to know the performance of people in the discrimination of voice qualities of natural utterances. Such data would be a useful baseline for comparisons, but would pose substantial requirements to the experimental design.

Despite the limitations, our results suggest a few indications how to improve the modeling of the voice qualities. First, the synthetic modal voice was misclassified as pressed in 33% of the cases, but never as breathy or whispery. Therefore, this quality should be synthesized with a somewhat more abducted setting. The synthetic breathy voice might be better distinguishable from whispery voice, if the degree of abduction is even more increased. However, the model oscillates for greater abduction only if the posterior chink area is reduced. Furthermore, the whispery voice quality could be improved by further increasing the posterior chink area at the expense of a reduced oscillation amplitude of the ligamental vocal folds.

Finally, for future studies it would be conceivable to reproduce the voice of one *particular* real person that speaks with different voice qualities instead to compare the generic model voice with averaged measurements of previous publications.

References

[1] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, pp. 383–406, 2001.

[2] J. Laver, *The phonetic description of voice quality*. Cambridge University Press, 1980.

[3] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *The 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 2417–2420.

[4] M. Vainio, M. Airas, J. Järvikivi, and P. Alku, "Laryngeal voice quality in the expression of focus," in *Interspeech 2010*, Makuhari, Japan, 2010, pp. 921–924.

[5] G. Klasmeyer and W. F. Sendlmeier, "The classification of different phonation types in emotional and neutral speech," *Forensic Linguistics*, vol. 4, pp. 104–124, 1997.

[6] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[7] 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.

[8] S. Adachi and J. Yu, "Two-dimensional model of vocal fold vibration for sound synthesis of voice and soprano singing," *Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3213–3224, 2005.

[9] J. Liljencrants, "A translating and rotating mass model of the vocal folds," *STL-QPSR*, vol. 1, pp. 1–18, 1991.

[10] 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.

[11] 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.

[12] I. T. Tokuda, J. Horacek, J. G. Svec, and H. Herzel, "Comparison of biomechanical modeling of register transitions and voice instabilities with excised larynx experiments," *Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 519–531, 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] 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.

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

[16] A. Yang, J. Lohscheller, D. A. Berry, S. Becker, U. Eysholdt, D. Voigt, and M. Döllinger, "Biomechanical modeling of the three-dimensional aspects of human vocal fold dynamics," *Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 1014–1031, 2010.

[17] 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.

[18] B. J. Kröger, "Simulation of vocal fold oscillation behaviour by a self-oscillating glottis model," *Journal de Physique*, vol. 4, pp. 457–460, 1994.

[19] 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.

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

[21] 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.

[22] 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.

[23] P. Ladefoged, "Discussion of phonetics: a note on some terms for phonation types," in *Vocal Physiology: Voice Production*, O. Fujimura, Ed. New York: Raven Press, Ltd., 1988, pp. 373–375.

[24] 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.

[25] M. Blomgren and Y. Chen, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2649–2658, 1998.

[26] D. G. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[27] R. L. Whitehead, D. E. Metz, and B. H. Whitehead, "Vibratory patterns of the vocal folds during pulse register phonation," *Journal of the Acoustical Society of America*, vol. 75, no. 4, pp. 1293–1297, 1984.