Modeling Different Voice Qualities for Female and Male Talkers Using a Geometric-Kinematic Articulatory Voice Source Model: Preliminary Results

Bernd J. Kröger[1]
Peter Birkholz
Jim Kannampuzha
Christiane Neuschaefer-Rube

Abstract: Modeling natural sounding voice qualities – for example the pressed-modal-breathy voice quality continuum which widely occurs during normal speech production – is a crucial point in speech synthesis. A parametric voice source model using prescribed sinusoidal vocal fold vibration patterns (i.e. extended Titze model) is introduced in this paper. This voice source model was adapted for synthesis of a typical male and female voice. A simulation experiment was performed by varying glottal abduction/adduction in order to generate a voice quality continuum from pressed over modal towards breathy. A parameter analysis of the resulting waveshapes of glottal flow and its time derivative was carried out in terms of the LF-model. This analysis indicates that our parametric voice source model is flexible enough to generate the modal to breathy but not the modal to pressed voice quality continuum for the male as well as for the female voice. It can be hypothesized that a self-oscillating voice source

[1] This paper is dedicated to the 60th Birthday of Bernd Pompino-Marschall. I met Bernd Pompino-Marschall in the early 90th of the last century as a young scientist when Bernd already was Professor in Berlin. I thank him especially for supporting me in my hard years between PhD and getting my permanent position at RWTH Aachen University. In these years I always was welcome to visit his Lab at ZAS and at Humboldt University of Berlin and I was employed in Berlin from 1999 to 2000. Great times! I learned a lot! Thank you for all, Bernd!

model is needed in order to generate the whole spectrum of vocal fold vibration patterns occurring during normal speech production.

1 Introduction

The problem of generating natural sounding voices as well as generating changes in voice quality is not yet solved for parametric voice source models used in articulation-based speech synthesis systems. Since natural voice quality is a key for reaching high quality synthetic speech, corpus-based synthesis is currently used if high quality speech is needed (Clark et al., 2007). But changes in voice quality as they occur during the production of utterances (e.g. normal to pressed and normal to breathy, Klatt and Klatt (1990) as well as differences in voice quality due to different speaking styles or due to different emotional states of speakers cannot be realized easily in corpus-based approaches. In addition, if the synthesis of different voices is demanded (e.g. male vs. female or child vs. adult), corpus-based systems need one complete speech corpus for each voice. Furthermore, modeling huge differences in loudness as well as modeling huge differences in pitch is not unproblematic if corpus-based speech synthesis is used. Thus, if a variety of natural sounding voice qualities is demanded, i.e. different speakers, different voice qualities, huge loudness and pitch ranges, it would be advantageous to use other approaches, e.g. articulation-based or comparable parametric speech synthesis approaches. But currently the segmental as well as prosodic quality of these systems is not as high as that of corpus-based approaches.

The quality of articulation-based speech synthesis increased dramatically over the last decades. Huge corpora of kinematic data (EMA, EPG, MRI) are now available for developing realistic models for vocal tract geometries of speech sounds as well as for developing approaches for modeling coarticulation and vocal tract kinematics (Badin et al., 2002; Engwall, 2003; Birkholz and Kröger, 2006; Serrurier and Badin, 2008). In addition, the knowledge concerning vocal tract acoustics and aerodynamics increased (see the overview on vocal tract acoustics given by Stevens, 1998), which now allows the development of high quality articulatory-acoustic models (e.g. Birkholz et al., 2007). Furthermore a lot of knowledge is currently available concerning the voice source and its integration in articulation-based synthesis systems (e.g. Titze, 1989a; Story and Titze, 1995; Titze and Story, 2002).

Breathy voice quality, modal (or normal) voice quality, and pressed voice quality (sometimes used synonymously with the terms creaky or laryngealized) can occur within the realization of a single utterance (Klatt and Klatt, 1990). Thus these qualities can be seen as phonation subtypes occurring during speech, i.e. during the use of the chest register as basic phonation type. The change from modal to breathy voice quality occurs within sentence production, for example by changing from a normal or stressed to an unstressed syllable (Gobl and Chasaide, 1988). A change from normal to pressed voice quality often occurs at the end of an utterance in order to signalize finality (Klatt and Klatt, 1990). In this paper we will not focus on creak or laryngealization but on breathy, modal, and pressed. The voice qualities breathy and pressed can easily be reached from modal voice quality at nearly all pitch levels within the chest register - i.e. without any "register break ", (see Sundberg, 1987, p. 50 for the definition of the term "register break ") - while creaky voice, laryngealized voice, as well as breathy-laryngealized voice only occur at very low pitches (Klatt and Klatt, 1990) and thus should be treated as phonation types occurring within a separate voice register - i.e. vocal fry or pulse register, (see Sundberg, 1987, p. 50) for the definition of these registers - or at least as special types of vocal fold vibration occurring at the lower edge of the chest register.

The pressed-modal-breathy voice quality continuum is realized by varying mainly one physiological voice source control parameter, i.e. *glottal or vocal fold abduction/adduction*. (i) During glottal abduction vocal folds become more and more separated from each other by an outward translation or rotation of the arytenoid cartilages, but this separation is not sufficiently large to cause vocal fold vibration to cease. This abduction process changes voice quality from modal to breathy. (ii) After glottal adduction, vocal folds are (softly) adducted. A further increasing adduction results in medial compression that activates a force which occurs perpendicular to the length of the vocal folds on the level of the vocal processes and modal voice quality now changes towards pressed voice quality.

A lot of evidence exists for the fact that two dimensions of control are sufficient in order to describe phonemic contrasts produced by the voice source as well as the phonetic variability of the voice source during normal (i.e. emotionally neutral) utterance production. (i) Firstly, changes in fundamental frequency (F0) lead to distinctive contrast in many languages (i.e. tone languages). In addition, changes in F0 occur within all languages as a phonetic realization feature for intonation and thus for prosody. Physiologically, changes in F0 mainly result from varying *vocal fold tension* which occurs parallel to the length of the vocal folds. (ii) Secondly, changes in voice quality with respect to breathy vs. modal vs. pressed occur as a phonemic contrast in some languages (Klatt and Klatt, 1990; Gordon and Ladefoged, 2001). In addition, these changes in voice quality occur within many languages as a phonetic feature of stress as well as sentence or utterance finality (see above). Furthermore, vocal fold or glottal abduction is the main control dimension for the linguistically important voiced-voiceless sound contrast while glottal adduction in addition is the main control dimension for the voiced sound to glottal stop contrast: If abduction is stronger than for changing voice quality from normal to breathy, vocal fold vibration ends and voicelessness occurs; if adduction is stronger than for changing voice quality from normal to pressed, vocal vibration ends as well and a glottal stop is produced.

Voice source models should reflect these facts by introducing *two main voice source control parameters*, i.e. (i) vocal fold tension resulting from forces occurring parallel to the length of the vocal folds and (ii) glottal abduction/adduction including medial compression resulting from forces perpendicular to the length of the vocal folds on the level of the vocal processes. On the basis of this control paradigm, voice source models should be able (i) to produce F0 variation, (ii) to produce the voiceless to voiced sound to glottal stop contrast, and (iii) to produce the typical vocal fold vibration patterns and the typical acoustic features of the voice quality continuum breathy-modal-pressed.

The physiological differences of the male vs. female voice source as well as the resulting glottal flow waveshapes and acoustic features are described in the literature (e.g. Holmberg et al., 1988; Titze, 1989c; Hanson and Chuang, 1999; Karlsson, 1992; Karlsson and Liljencrants, 1996). These settings and features are listed in detail in section 2 of this paper as a basis for our modeling experiment (section 4). It should be noted that female voices in general are found to be more breathy than male voices, which makes it interesting to model male vs. female voice quality together with modeling the pressed-modal-breathy voice quality continuum. The physiological or voice source settings as well as the resulting glottal flow waveshapes and acoustic features for breathy, modal, and pressed voice quality for both male and female speakers are described in the literature, e.g.Klatt and Klatt (1990), Karlsson and Liljencrants (1996) and these settings and features are listed in section 3 of this paper as a further basis of our modeling experiment.

The articulation-based speech synthesis system used for our simulation experiment is a geometric-kinematic model. No detailed neuromuscular and no detailed biomechanical modeling of the vocal tract articulators (i.e. tongue, lips, lower jaw, velum, larynx) or of the voice source is aimed for in our model. Our vocal tract model comprises a *geometric articulator model* (Birkholz et al., 2006; Birkholz and Kröger, 2006)) driven by a *kinematic articulatory control model* via a set of phonetically motivated *articulatory control parameters* (Kröger and Birkholz, 2007). Our voice source model comprises a *geometric voice source model* as introduced by Titze (1984, 1989a) representing the membranous part of the glottis extended by a permanent bypass for airflow representing the cartilaginous part of the glottis (Kröger, 1997, 1998; Birkholz, 2005). Both parts of the voice source model are driven by a *kinematic voice source control model* via a set of phonetically (i.e. acoustically and articulatorily) motivated *voice source control parameters* (introduced below). In addition the *pulmonary control parameter* lung pressure p_l is introduced. A lung pressure value above phonation threshold pressure (Titze, 1988; Chan and Titze, 2006) is needed for the initiation and maintenance of vocal fold vibration.

While the vocal tract model and its control has already been described in earlier publications (see above), the geometric voice source model and its control is described in detail in section 2 of this paper. In addition the (time invariant) parameterization of the voice source model for a typical male and for a typical female speaker is given. Furthermore a set of (time variant) voice source control parameters capable of adjusting the larynx for the realization of different voice qualities for a male as well as for a female speaker is introduced.

2 The voice source control model

2.1 Structure of the voice source model and its parameters

In our modeling approach, voice source *static* parameters are speaker-specific and time-invariant, i.e. voice source static parameters are capable of adjusting the voice source model with respect to a specific speaker's voice source geometry, while time-variant voice source *dynamic control* parameters specify phonatory gestures generating speech sounds and speech prosody. The set of voice source static parameters directly results from the design of the membranous and cartilaginous part of the voice source model while the set of voice source dynamic control parameters are designed mainly with respect to phonetic criteria for controlling vocal fold activity during speech. It is the task of these dynamic control parameters (i) to generate all segmental voiceless-voiced and voiced-voiceless transitions (i.e. glottal adduction and glottal abduction gestures), (ii) to generate different degrees of stress (sentence stress pattern), intonation contours (sentence intonation pattern), and different loudness levels (soft-normal-loud), and (iii) to generate different voice qualities. The time course of voice source control parameters is generated by the voice source control model.

The *speaker-specific voice source static parameters* for adjusting the voice source with respect to a specific speaker are (i) the *effective length of the membranous and cartilaginous part of the glottis* L_m and L_c and (ii) the *thickness of the glottal constriction* T. It should be emphasized that "effective length" means not "real length" of the vocal folds as occurs if the vocal folds are not vibrating. The effective length of the vocal folds is the midsagittal length of the glottis during vibration (see Figure 1). This effective length is shorter than the real length of the vocal folds, because the vocal folds are bent during vibration (Hollien and Moore, 1960; Titze, 1989c). The real length of the vocal folds is approximately 16 mm for a male and 10 mm for a female speaker. Within the chest register (which is most important for speech) the effective length of the membranous part L_m is different for males and females (as well as the real length) but in addition the effective length varies with frequency. Length values for L_m as estimated after Hollien (1960); Hollien and Moore (1960);

Titze (1989c) for a male and a female voice source and are subsumed in a formula given by Titze (1989a, p.195) for a typical male voice. It can be assumed that the length of the membranous part is shorter by a scale factor of $1/1.6 = 0.625$ for female voices (Titze, 1989c, p.1700). The resulting length values are listed in Table 1.

Table 1. Effective length of membranous part of the glottis L_m and thickness of glottal constriction T as function of fundamental frequency F0 for a standard male and a standard female speaker for the chest register. Estimation after a formula given by Titze (1989b) for the male voice and by using a male to female scaling factor of 0.625 (Titze, 1989c).

MALE VOICE	low	mid-low	mid	mid-high	high
F0 [Hz]	70	90	120	160	200
L_m [mm]	6.2	7.6	8.0	10.2	11.5
T [mm]	11.5	9.5	9.0	7.5	6.3
FEMALE VOICE	low	mid-low	mid	mid-high	high
F0 [Hz]	120	150	200	270	350
L_m [mm]	5.0	6.4	7.2	8.3	9.4
T [mm]	5.6	4.7	3.9	3.4	3.0

The length of the cartilaginous part of the glottis L_c can be estimated as 3 mm for male and 2.5 mm for female speakers (Hirano, 1983; Titze, 1989c). Thickness of the glottal constriction T is different for males and females as well and also depends on frequency (Hollien, 1960; Titze, 1989c). Thickness values for T as estimated after Hollien (1960) and Titze (1989c) are also subsumed in a formula (Titze, 1989a, p.195) and given in Table 1 for the whole range of the chest register (low, mid-low, mid, mid-high, high) for a typical male and female voice. Differences in length and thickness between the male and female voice source can be scaled by the anatomy related scale factor $1/1.2 = 0.83$ for the cartilaginous part of the glottis (Titze, 1989a). It should be noted that the lower scale factor of 0.625 for the membranous part of the glottis (see above) indicates that the vocal folds are a gender-specific human organ.

From the phonetic viewpoint it is advantageous to control the voice source by using two main *voice source dynamic control parameters*, i.e. *vocal fold tension* and *glottal aperture* (i.e. degree of *glottal ab-/adduction*) (see section 1 of this paper). Thus, two different types of *voice source gestures* *or phonatory control gestures* are mainly used in articulation-based speech synthesis systems, i.e.(i) *vocal fold tension gestures* and (ii) *glottal abduction and adduction gestures* (see Gordon and Ladefoged, 2001; Kröger, 1993; Kröger, 1998; Kröger and Birkholz, 2007). In order to identify vocal fold tension and glottal aperture in terms of our voice source model it is now necessary to describe our voice source model in more detail. The voice source model of the membranous and cartilaginous part of the glottis is given in Figure 1 (see also Titze, 1984, 1989a; Kröger, 1997; Birkholz, 2005). We can directly identify (i) *effective length of the glottis for the membranous part L_m* and *effective length of the glottis for the cartilaginous part L_c*, (ii) *thickness of the glottal constriction T*, (iii) *cross-sectional area of the glottis for the membranous part* (in the case of vocal fold rest position, also called pre-phonatory state) A_m and for the *cartilaginous part A_c*, (iv) *displacement of posterior membranous part of (symmetrically positioned) vocal folds in rest position x_m* (i.e. displacement of vocal processes, sometimes also called degree of glottal ab-/adduction) and *displacement of the arytenoid cartilages x_c*, (v) *displacement of upper* (or superior) posterior part and *of lower* (or inferior) *posterior membranous part of vocal folds x_{mu}* and x_{ml} with $x_m = (x_{mu} + x_{ml}) * 0.5$ as *mean displacement of the posterior membranous part of the vocal folds* and with $x_{md} = x_{ml} - x_{mu}$ as difference between displacement of upper and lower posterior membranous part of the vocal folds. x_{md} is called *converging rest-position displacement*; $x_{md} > 0$ means: converging vertical glottal shape in rest position as is given in Figure 1b; $x_{md} < 0$ means: diverging vertical glottal shape in rest position. [2] In the case of vocal fold vibration additional parameters need to be introduced. The model can be controlled directly by (i) *fundamental frequency* F0 (i.e. frequency of vocal fold vibration), (ii) *amplitude of vocal fold vibration x_v* (which is equal for the upper and lower part of the vocal folds in this model), and (iii) *phase difference between upper and lower part of vocal*

[2] Please notice that all $x_{m...}$-parameters represent displacements or differences of displacements at the (posterior) *end* of the membranous part of the glottis (because the cross-sectional area of this part has a triangular form, Figure 1a), while m_c represents the displacement in the *middle* of the cartilaginous part of the glottis since the cross-sectional area of this part has a trapezoidal form (Figure 1a). Thus, the relationship between cross-sectional glottal area and vocal fold or cartilage *displacement* (which is half of the vocal fold or cartilage *distance*) is $A_c = 2 * x_c * L_c$ for the cartilaginous part and is $A_m = x_m * L_m$ for the membranous part of the glottis (in rest position or pre-phonatory state).

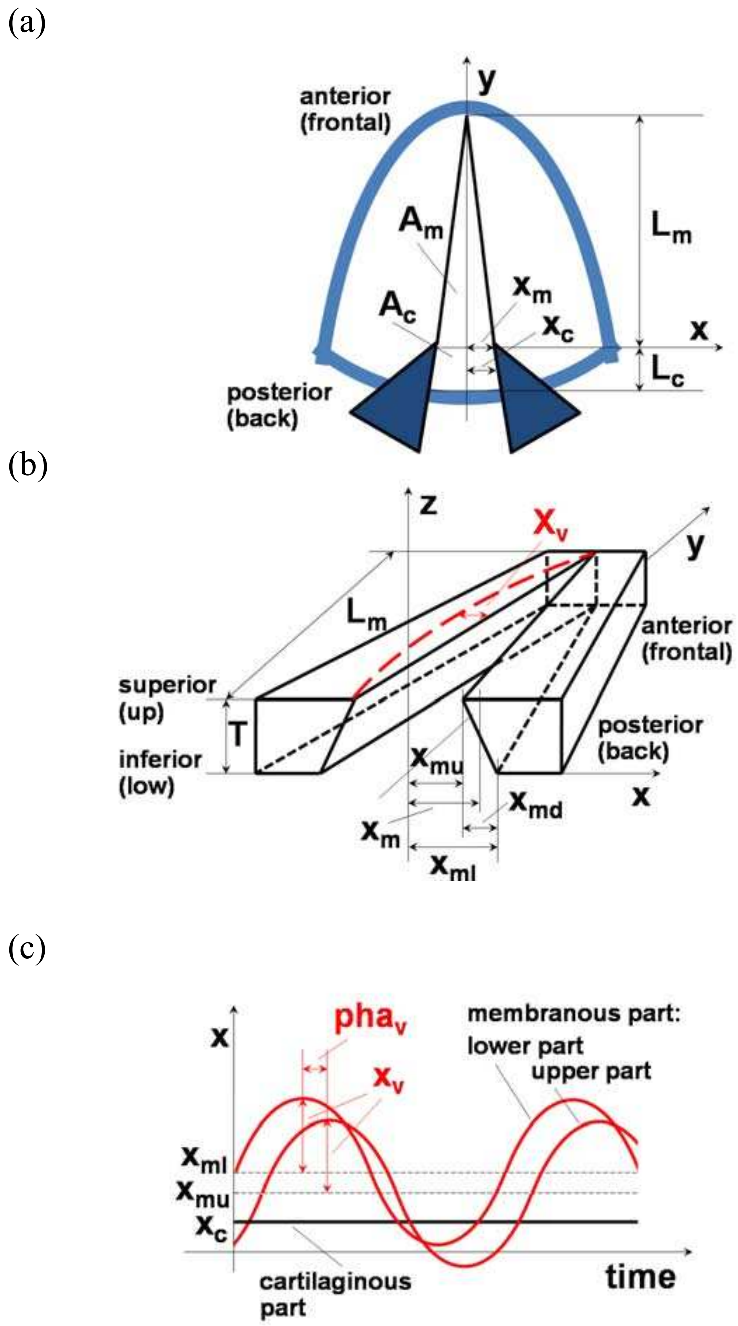

Figure 1. The geometric voice source model. a) View of the vocal folds from above (or horizontal section through the larynx), visualizing the membranous as well as the cartilaginous part of the vocal folds. The filled triangles represent the arytenoid cartilages, the anterior grey arc represents the thyroid cartilage, and the posterior grey arc represents the posterior part of the cricoid cartilage. b) Schematic view of the vocal folds, i.e. of the membranous part of the glottis and its parameterization (after Titze 1984). c) Displacement over time of the upper and lower membranous part of the vocal folds relative to its its rest positions x_{ml} and x_{mu} (dashed lines) during vocal fold vibration. The displacement of the arytenoids cartilages x_c over time is indicated by a black bold line. Vibration patterns in space b) and in time c) are indicated as grey lines. Vibration pattern in space in b) is indicated only for the upper edge of the left vocal fold.

folds pha_v. Some additional remarks concerning our voice source model should be given here: (i) x_m and x_{md} are directly controlled in our voice source model. Thus x_{mu} and x_{ml} are calculated as $x_{mu} = x_m + x_{md} * 0.5$ and $x_{ml} = x_m - x_{md} * 0.5$ (see Figure 1b). (ii) Positive values of x_m directly quantify the degree of abduction for the membranous part of the vocal folds. Negative values for x_m are possible and quantify the degree of adduction and thus quantify the *degree of medial compression* on the vocal folds. This is important if a glottal stop or if pressed voice qualities are modeled.

The vocal fold vibration amplitude x_v is controlled (or calculated) indirectly, depending on the rest position displacement of the membranous part of the glottis x_m and on subglottal pressure p_s (i.e. $x_v = f(x_m, p_s)$ where $p_s = p_l$ in our voice source model) (see Birkholz, 2005, p.40). Vocal fold vibration amplitude x_v increases proportional to p_s^2 (see Titze, 1989b, p.104) and (Birkholz, 2005, p.40) and decreases with increasing glottal ab-/adduction, which leads to an absence of vocal fold vibration if the degree of vocal fold abduction is high (as is needed for example for the realization of voiceless sounds). Other influences on vocal fold vibration amplitude x_v are neglected in this preliminary version of the voice source model.

Now the *voice source dynamic control parameters* vocal fold tension and glottal ab-/adduction can be quantified in terms of our voice source model. (i) It is assumed that vocal fold tension mainly influences fundamental frequency F0 and thus F0 is taken directly as voice source control parameter instead of vocal fold tension. But it should be kept in mind that F0 can be seen here as a direct equivalent to vocal fold tension since in our model F0 has a direct influence on effective length and thickness of the vocal folds (Table 1). (ii) The voice source control parameter *glottal abduction and/or glottal adduction* is quantified by at least four voice source parameters, i.e. displacement of membranous and cartilaginous part of the glottis x_m and x_c, converging rest position displacement x_d, and phase difference of lower versus upper edge of the vocal folds pha_v. The later two parameters (converging rest position displacement and phase difference of lower vs. upper part of the vocal folds) are held constant during glottal ab- and adduction gestures (see below) but the distances of the membranous part and of the cartilaginous part of the glottis x_m and x_c may vary independent from each other even within a single abduction or adduction gesture. This results from different types of movements which can occur for the arytenoids, i.e. rotation and translation. Both types of movements can be superimposed in a single abduction or adduction gesture. Figure 2 illustrates that (i) translation leads to comparable change in x_m and x_c while (ii) rotation mainly leads to changes in x_m while x_c remains constant.

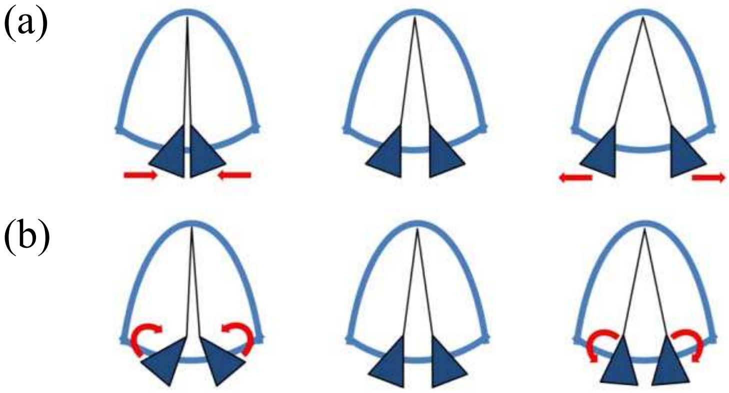

Figure 2. (a) Inward and outward translation. (b) Inward and outward rotation of the arytenoid cartilages occurring in glottal abduction gestures.

2.2 Gestures for controlling voice source activity during speech

While F0 gestures can easily be generated for example by using the Fujisaki approach (Birkholz, 2005, p.111), the situation is more complex for the positioning of the membranous and the cartilaginous part of the glottis for phonation or for segmental voiceless-voiced and voiced-voiceless changes. This laryngeal positioning is realized by glottal abduction and glottal adduction gestures (see Figure 3).

A prominent example for a glottal adduction gesture is the transition of the glottal opening from breathing or from a voiceless sound towards modal phonation. This gesture starts with a wide opening of the glottis where no vocal vibration occurs. The gesture ends with a closure of the glottis (x_m and x_c near 0) and a typical converging vertical shape of the glottis ($x_d > 0$). The best or most effective parameter set for modal phonation can be found by trial and error phonation during modeling voice and speech acquisition (i.e. early phonation stage Kröger et al.,

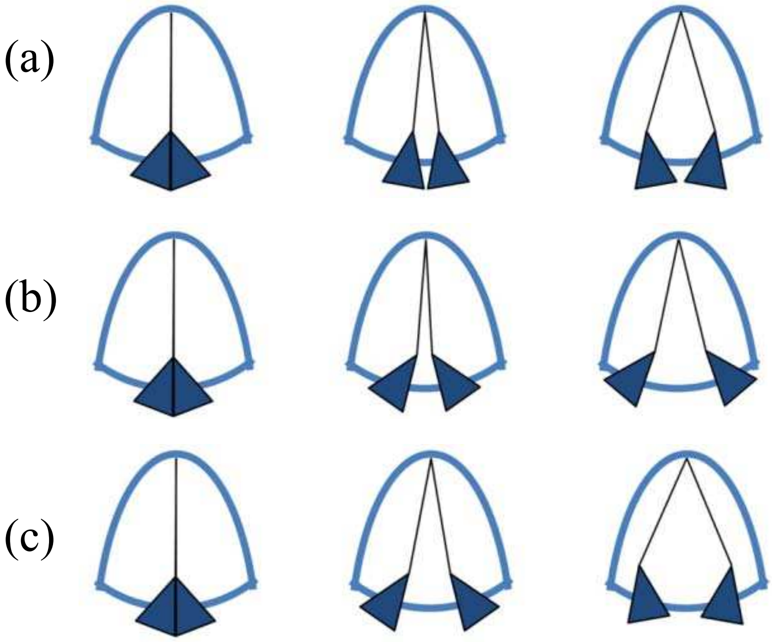

Figure 3. Three abduction gestures with different contributions of outward translation, inward rotation, and outward rotation. Resulting net cross-sectional increase (i.e. resulting overall degree of abduction) is comparable for gestures (a) and (b) but x_m is always smaller during abduction gesture (b) which means, that vocal fold vibration ends at higher values for glottal aperture in this gesture in comparison with gesture (a). Thus gesture (b) is used in our approach for modeling smooth voiced-voiceless transitions and voiceless-voiced transitions. If a higher degree of glottal aperture is needed as can be reached in gesture (b) (e.g. for breathing), a further outward rotation of the arytenoids may occur at high degrees of glottal aperture (c).

2009). If normal phonation is maintained during the whole utterance the parameters x_d and pha_v can be maintained even during within-sentence abduction-adduction gestures needed for the realization of voiceless segments. In addition for a natural sounding segmental voiceless-voiced transition it is important that vocal fold vibration already starts *before* the adduction gesture has reached its target, i.e. the glottal closure for effective modal phonation. Thus glottal abduction gestures should be able to generate smooth voice onsets and voice offsets (Gobl and Chasaide, 1988). Consequently, the vibrational amplitude x_v of the vocal folds needs to be carefully controlled within glottal abduction gestures. Since x_v (beside p_l) mainly depends on x_m in our voice source model, a carefully designed control of x_m (together with x_c) is needed. For smooth voice onsets and voice offsets this leads to a translational movement combined with an inward rotational movement of the arytenoid cartilages during voiceless-voiced adduction (Figure 2) as well as vice versa for voiced-voiceless abduction gestures in order to keep x_m small even for medium degrees of abduction (Birkholz, 2005, p. 40).[3]

2.3 Acoustic relevance of voice source parameters

Following the description of our voice source model (see 2.1), six voice source parameters (L_m, L_c,T, x_m, x_c, x_d) describe the geometry of the vocal folds in its rest position (pre-phonatory state) while three parameters (F0, x_v and pha_v) describe vocal fold vibration. A shortcoming of this parameterization is that the waveshape of the cross-sectional glottal area and of the glottal volume flow, which results from vocal fold vibration, cannot be estimated directly. But especially the waveshape of glottal volume flow and of its first time derivative directly reflects important acoustic features of the voice source signal and thus of voice quality (Klatt and Klatt, 1990). Thus we will introduce now a parameterization of the waveshape of glottal flow and its first time derivative. This parameterization is related to the LF model (Fant et al., 1985). At the end of this section the influence of different voice source parameters on these *flow waveshape parameters* is discussed.

The glottal flow waveshape can be subdivided into glottal open and closed period, and the open period itself can be subdivided into opening and closing period (Figure 4).

If the cartilaginous part of the glottis is completely closed, no glottal flow occurs in the closed period. The *leak flow* fl_{lk} reflects (i) the flow passing the cartilaginous part of the glottis together with (ii) the flow passing that part of the membranous glottis which never closes during the glottal (vibration) cycle. T_0 is the *duration of a complete glottal cycle* ($F0 = 1/T_0$). T_c represents the time instant of glottal closure and thus the *duration of the glottal open phase* equals T_c while the *duration of the closed phase* can be calculated as $T_0 - T_c$. T_p represents the time instant of *peak flow* (or maximum flow) fl_{pk}. Thus the *duration of opening phase* equals T_p while the *duration of the closing phase* can be calculated

[3] A strong increase in vocal fold tension is a further mechanism for controlling voicelessness (cf. Hanson and Stevens, 2002) beside increasing abducion. This mechanism will be considered in future modeling studies.

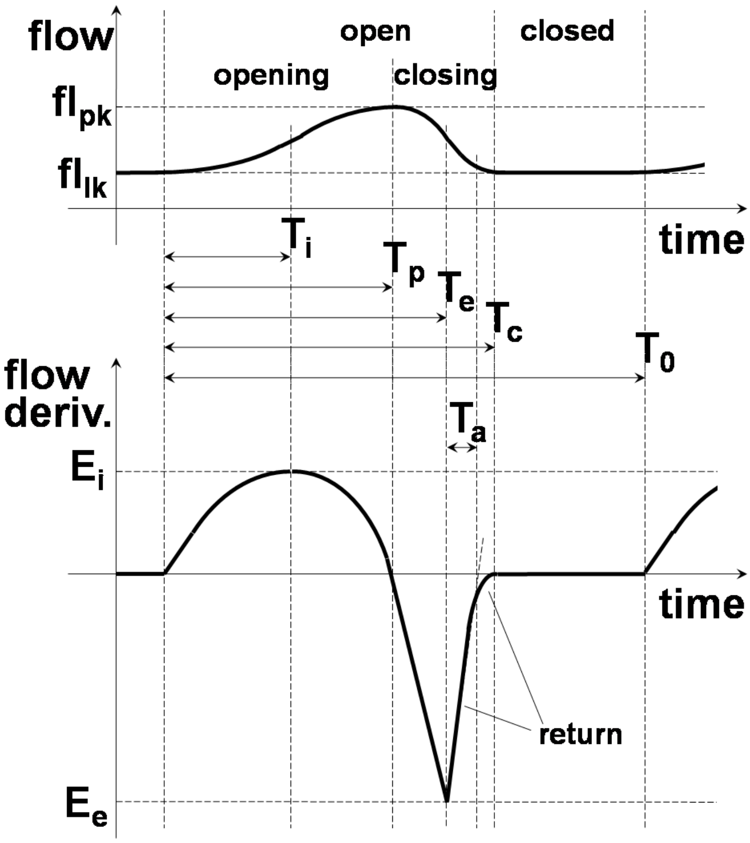

Figure 4. Typical waveshape of glottal flow and its first time derivative for one glottal vibration cycle in modal phonation. A description of the parameters is given in the text.

as $T_c - T_p$. During the opening phase the time instant of maximum *increase* in glottal flow T_i can be identified and the maximum increase in flow is quantified as E_i. During the closing phase, in an analogous way the time instant of maximum *decrease* in glottal flow can be identified. This is the time point T_e, at which the voice source produces the *maximal vocal tract (acoustic) excitation*. The maximum (acoustic) excitation amplitude is given as E_e. The time interval between T_e and T_c is called the return phase and the *decay time* T_a as defined in Figure 4 is a measure for the *abruptness of glottal closure* (please note: $T_a < T_c - T_e$). If decay time T_a is zero, the glottis is closed in the most abrupt way by cutting down the flow to leak flow immediately at the time instant of maximum vocal tract excitation T_e. In this case the time derivative of glottal flow abruptly decreases to zero at T_e. The more the decay time T_a increases, the less abrupt the glottal closure takes place. In this case the decay time T_a occurs between $T_c - T_e > T_a > 0$.

The *relative time parameters of the glottal flow waveshape* are always quantified relative to the length of the glottal cycle T_0 and thus relative to fundamental frequency. These parameters are RG, RE, OQ and RA in terms of the LF-model (Karlsson and Liljencrants, 1996): The *opening quotient* T_p/T_0 equals $1/2 * RG$, the *excitation time quotient* T_e/T_0 equals RE, the *open quotient* T_c/T_0 is also labeled OQ, and the *relative decay time* T_a/T_0 equals RA. In addition two glottal flow time parameters are quantified absolutely without being related to the voice fundamental. These are *opening frequency* FG, which equals $1/(2\pi T_p)$, and *decay frequency* FA, which equals $1/(2\pi T_a)$. The *open quotient* T_c/T_0 as well as the *flow pulse skewness* factor $(T_c/T_p) - 1$ (also labeled as *speed quotient* SP) give information about the waveshape of the vocal fold vibration. The relative location of the acoustic excitation time point with respect to the glottal opening period $(T_e/T_p) - 1$ is labeled as *excitation skewness* factor RK (Karlsson and Liljencrants, 1996), which is only slightly lower than flow pulse skewness since the time instant of maximal excitation T_e occurs at the end of the glottal pulse. *Peak flow* fl_{pk}, *leak flow* fl_{lk} (also called *dc-flow*), and the difference between peak flow and leak flow $fl_{pk} - fl_{lk}$ (also called *ac-flow*) are *aerodynamic parameters of the glottal flow wave shape* which can be used to adjust the vibrational amplitude of the vocal fold model and which can be used to adjust the aperture of the membranous and cartilaginous part of the vocal fold model x_m and x_c, if lung pressure is known. An interpretation of glottal flow parameters in terms of acoustics was put forward by Carlson et al. (1989). The main results are that fl_{pk} as well as E_i represent the amplitude or spectral energy of the voice fundamental. E_e (also sometimes labeled as EE) represents the spectral energy above FG. Both energy or amplitude levels should only be interpreted if they can be related to each other as $E(fl_{pk})/EE$ or E_i/E_e. The decay time T_a determines the spectral tilt of glottal flow (as well as of its derivative). Medium T_a around 0.15 ms (and higher values) leads to a medium spectral tilt of -12dB/octave for glottal flow. Medium T_a to minimum T_a around 0.0 ms leads to a less steep spectral tilt of about just 6dB/octave for glottal flow. Medium to maximum T_a of around 0.6 ms leads a steeper spectral tilt of about -18dB/octave for glottal flow.

From the simulation experiments done by Titze (1989a) it can be concluded even for our extended model that decay time T_a mainly increases with increasing vocal fold abduction x_m. Open quotient T_c/T_0 increases with vocal fold abduction but also with increasing converging glottal shape x_d and slightly with decreasing phase delay pha_v. But further simulation experiments will be done in this paper in order to highlight the relation between geometrical-articulatory and acoustic parameters of the voice source in more detail.

3 Laryngeal behavior of male and female breathy, modal, and pressed voice

Modal (or normal) phonation in the mid or mid-low region of the chest register means that the vocal folds (membranous part of the glottis) as well as the cartilaginous part of the glottis are nearly closed (x_m and x_c around 0). The peak-flow (fl_{pk}) and leak flow (fl_{lk}) are around 0.38 l/s (l=liter) and 0.12 l/s for males and around 0.22 l/s and 0.09 l/s for females (at lung pressure around 630 Pa for males and around 570 Pa for females, Holmberg et al. (1988)) indicating that a noticeable chink (i.e. a leak between the cartilaginous part of the glottis) exists for males as well as for female voices in the case of normal phonation. The flow waveform typically shows an open quotient T_c/T_0 of approximately 0.55 for male and around 0.75 for female voices (Holmberg et al., 1988; Klatt and Klatt, 1990; Karlsson and Liljencrants, 1996, and Table 2). The glottal pulse is slightly skewed with $(T_e/T_p) - 1$ as well as $(T_c/T_p) - 1$ of approximately 0.38 for male voices and 0.43 for female voices (Karlsson and Liljencrants, 1996). The spectrum of glottal flow has an average tilt of approximately 12 dB/octave (Klatt and Klatt 1990), and T$_a$ is approximately 0.15 msec for male voices and around 0.2 msec for female voices (Karlsson and Liljencrants, 1996) and Table 2. Changing from modal to *pressed voice quality* means that the membranous and cartilaginous part and the glottis are now tightly closed and that a medial compression occurs at the vocal processes perpendicular to the length of the vocal folds (Klatt and Klatt, 1990). That leads to a narrowing of the glottal pulse and thus to a decrease in open quotient T_c/T_0 of about 0.41 for male voices, Karlsson and Liljencrants (1996) and Table 2). Glottal closure is more abrupt than in modal phonation. T_a is about 0.1 msec for male and female voices (Karlsson and Liljencrants, 1996) and Table 2 and consequently the spectral tilt is about -6 dB/octave for this change

Table 2. Parameters of glottal flow shape and acoustic relevant parameters for male and female pressed (=P), modal (=M), and breathy (=B) voice after Karlsson and Liljencrants (1996). All parameters have been discussed in the text. SP is an abbreviation for skew of pulse (see text).

		MALE			FEMALE		
		P	M	B	P	M	B
$F0$ [Hz]	$(1/T_0)$	128	126	131	261	246	254
$1/2RG$ [%]	(T_p/T_0)	27.0	38.0	40.0	45.0	43.0	48.0
FG [Hz]	$(1/2T_p)$	229	167	164	289	284	266
RE [%]	(T_e/T_0)	39.2	52.0	60.5	67.6	65.8	71.0
OQ [%]	(T_c/T_0)	41.0	54.0	65.0	71.0	71.0	79.0
Ta [ms]		0.102	0.160	0.349	0.124	0.202	0.320
RA [%]	(T_a/T_0)	1.31	2.02	4.57	3.24	4.96	8.12
FA [Hz]	$(1/(2\pi T_a))$	1620	1001	461	1290	810	503
SP [%]	$((T_c/T_p) - 1)$	39.5	37.7	51.1	50.0	51.8	48.5
RK [%]	$((T_e/T_p) - 1)$	39.5	37.7	51.0	49.9	51.7	48.3

in voice quality. Leak flow (fl_{lk}) is about zero here due to a completely closed glottis during the closed period of the glottal cycle for male as well as for female voices. Changing from modal to *breathy voice quality* means that the arytenoids are well separated at their posterior end, but that the vocal processes (i.e. anterior end of the arytenoids) are sufficiently approximated so that the vocal folds are capable of vibrating. While vocal folds close simultaneously along their length during modal phonation, this is not the case in breathy phonation. Here the vocal folds first close at the anterior end and then the closure propagates towards the posterior end of the vocal folds (Figure 1a), leading to an overall flow waveshape with a rounded corner at the end of the return phase (Figure 4). That leads to an increase of T_a in comparison to modal voice (T_a is approximately 0.35 ms for male and female voices, Karlsson and Liljencrants (1996) and Table 2) and furthermore leads to an increase in the spectral tilt to approximately -18 dB/octave (Klatt and Klatt, 1990). It remains to mention that the anterior-posterior propagation of closure, which occurs for breathy voice, should not be confused with the inferior-superior propagation of glottal closure, which occurs for all voice qualities in chest register. Glottal closure always starts at the inferior part of the glottis and propagates towards the superior part and glottal opening starts from the inferior towards the superior part. This behavior also occurs for a strongly converging vocal fold configuration (large x_d). But in that case the closure does not occur along the whole inferior-superior dimension but only in the superior part of the glottal constriction. Notice that a converging pre-phonatory glottal shape is important in order to allow subglottal pressure to push apart the vocal folds from the inferior part of the glottal constriction and thus to start and to maintain vocal fold vibration.

4 Experiment: Synthesis of a male and female breathy, modal, and pressed voice quality

4.1 Motivation and hypothesis

The voice source model parameters of our voice source model and typical glottal flow parameters for pressed, modal, and breathy phonation for a male and female voice are introduced above (section 2 and section 3). It is the goal of this experiment to find adjustments of the model parameters of our voice source model such that it is capable of reproducing glottal volume flow waveshapes, which are comparable to those described by Karlsson and Liljencrants (1996) for male and female modal, breathy, and pressed voice quality (Table 2).

4.2 Method

A continuum of male and female phonatory or voice source states was generated by varying the degree of glottal abduction. A continuum was chosen since it is known that male and female voices differ in breathiness (Holmberg et al., 1988; Kröger, 1989; Klatt and Klatt, 1990). All states were generated by using a small male vocal tract positioned for a constant /a:/ (Birkholz and Kröger, 2006). F0 was adjusted to 126 Hz for the male and to 255 Hz for the female voice in order to produce voice source signals which are comparable with those reported by Karlsson and Liljencrants (1996) (see Table 2). In both sets of states (male and female), glottal abduction is varied from high negative values (representing a tightly closed glottis with medial compression on the vocal folds and thus with no occurrence of vocal fold vibration) to high positive values (voiceless sound condition and again no occurrence of vocal fold vibration). The cartilaginous part was assumed to abduct with a displacement value of x_c which is comparable to the displacement value of the upper (or superior) posterior part of the membranous part of vocal folds x_{mu} (see Figure 1). The variation of x_{mu}, which represents the continuum of varying degree of glottal abduction is listed in Table 3. Starting with $x_{mu} = 0$, two cases can be separated: (i) increasing adduction towards $x_{mu} = -0.4$ mm for the male voice and towards $x_{mu} = -0.33$ mm for the female voice (see Table 3) and (ii) increasing abduction towards $x_{mu} = 1.0$ mm for the male and towards $x_{mu} = 0.83$ mm for the female voice (see Table 3 and Figure 5). For higher degrees of ab- or adduction no vocal fold oscillation occurs which is capable of exciting the vocal tract acoustically.

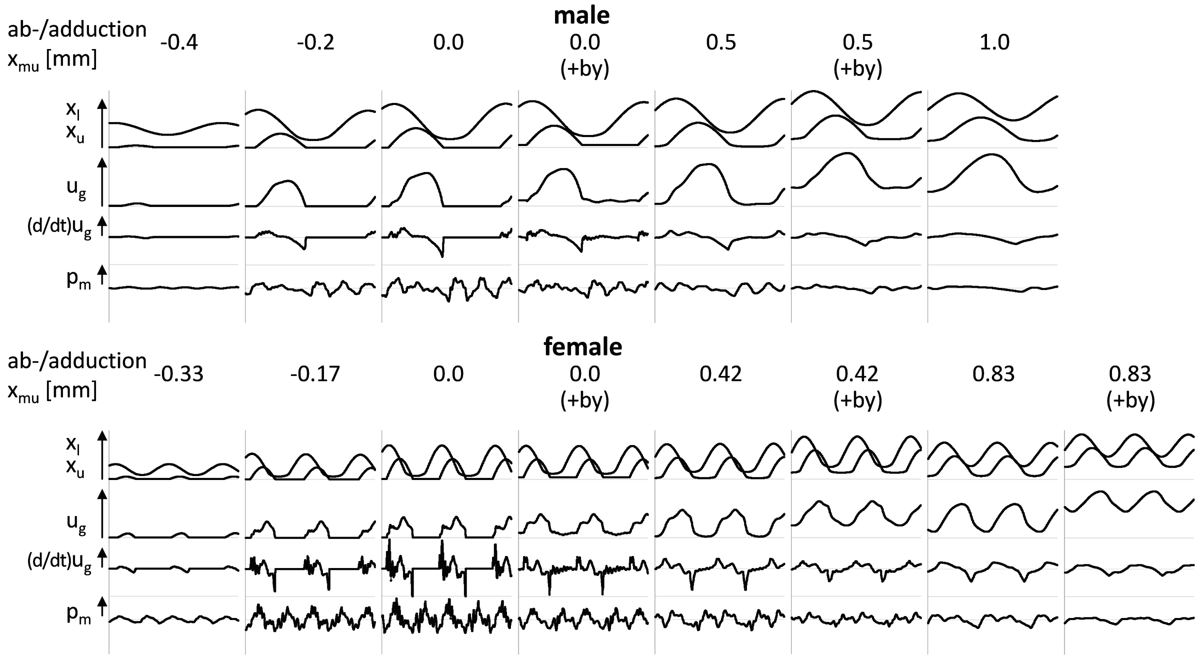

Figure 5. Waveshape of lower (inferior) and upper (superior) part of the vocal folds (distance x_l, x_u), of glottal flow u_g, of its first time derivative $(d/dt)u_g$, and of sound pressure radiated from the mouth p_m for approximately one glottal vibration cycle in the case of the male voice (upper row) and for approximately two glottal cycles in the case of the female voice (lower row) (time window is 10 ms in each case). Each row represents the continuum from pressed over normal towards breathy voice. The value given for each column is degree of ab-/adduction x_{mu} in mm (cf. Table 2).

A slight inward rotation of the arytenoids was introduced by setting

Table 3. Parameters of glottal flow waveshape for (a) male and (b) female voices for different degrees of glottal ab-/adduction. The label (+by) indicates simulations including the cartilaginous part of the glottis (extended Titze 1984 model, see Birkholz 2005; the cartilaginous part is also labeled "aerodynamic bypass" or "+by"). All other cases represent simulations using the pure non-extended Titze (1984) model. No data indicate that no audible vocal tract excitation is generated in that case.

MALE

ab-/adduct. $(x_m u)$	-0.4	-0.2	0.0	0.0 (+by)	0.5	0.5 (+by)	1.0	1.0 (+by)
$F0$ [Hz] $(1/T_0)$	127	126	126	125	124	126	125	-
$1/2RG$ [%] (T_p/T_0)	17.2	28.3	30.5	29.8	43.4	46.4	57.5	-
FG [Hz] $(1/2T_p)$	367	223	206	210	143	135	109	-
RE [%] (T_e/T_0)	24.7	44.0	48.6	48.6	63.1	62.7	80.2	-
OQ [%] (T_c/T_0)	31.3	46.3	50.3	50.0	71.1	72.5	93.6	-
T_a [ms]	0.476	0.181	0.068	0.087	0.399	0.658	1.036	-
RA [%] (T_a/T_0)	6.03	2.29	0.85	1.09	4.96	8.26	13.00	-
FA [Hz] $(1/(2\pi T_a))$	334	877	2340	1824	399	242	154	-
SP [%] $((T_c/T_p)-1)$	81.7	63.6	65.0	67.6	64.0	51.1	62.9	-
RK [%] $((T_e/T_p)-1)$	43.3	55.6	59.3	62.9	45.5	35.0	39.6	-
fl_{pk} [l/s]	0.018	0.180	0.272	0.292	0.378	0.460	0.453	-
fl_{lk} [l/s]	0.000	0.000	0.000	0.044	0.019	0.167	0.139	-
$fl_{pk}-fl_{lk}$ [l/s]	0.018	0.180	0.272	0.248	0.359	0.293	0.314	-

FEMALE

ab-/adduct. $(x_m u)$	-0.33	-0.17	0.0	0.0 (+by)	0.42	0.42 (+by)	0.83	0.83 (+by)
$F0$ [Hz] $(1/T_0)$	256	255	255	253	256	255	256	257
$1/2RG$ [%] (T_p/T_0)	23.8	33.2	35.5	34.8	47.7	49.7	50.6	49.6
FG [Hz] $(1/2T_p)$	538	383	359	364	269	256	253	259
RE [%] (T_e/T_0)	33.1	45.7	49.7	49.4	66.3	67.6	70.1	67.1
OQ [%] (T_c/T_0)	37.2	47.4	51.4	51.7	73.3	76.3	81.4	80.2
T_a [ms]	0.159	0.065	0.045	0.057	0.162	0.312	0.420	0.476
RA [%] (T_a/T_0)	4.07	1.66	1.16	1.45	4.14	7.95	10.76	12.24
FA [Hz] $(1/(2\pi T_a))$	1003	2447	3509	2791	985	510	379	334
SP [%] $((T_c/T_p)-1)$	56.1	42.6	44.7	48.8	53.7	53.5	60.9	61.8
RK [%] $((T_e/T_p)-1)$	39.0	37.4	39.8	41.1	39.0	36.0	38.5	35.3
fl_{pk} [l/s]	0.039	0.125	0.160	0.198	0.241	0.312	0.297	0.408
fl_{lk} [l/s]	0.000	0.000	0.000	0.039	0.014	0.132	0.070	0.251
$fl_{pk}-fl_{lk}$ [l/s]	0.039	0.125	0.160	0.159	0.227	0.180	0.227	0.157

$x_c = x_{mu} + 0.25mm$ for the male and for the female voice if $x_{mu} \geq 0$. A further set of voice source states was generated in the case of abduction with $x_c = 0$ (case: no cartilaginous part i.e. non-extended Titze 1984 model; the other case is designated as "including aerodynamic bypass", abbreviated as "+by" in Table 3) for the male and female voice source. In all sets of states the glottis converges in the vertical direction from its inferior to its superior part (Figure 1b) and a convergence ratio $x_d/x_v = 3.0$ is assumed (this ratio varies between 1.0 and 5.0 according to Titze, 1989a). The glottal convergence parameter x_d was hold constant for all voice source states in all sets of states. Phase delay pha_v was set to 72° (see Titze, 1984) and kept constant within all sets. Lung pressure p_l was set to 630 Pa for males and to 570 Pa for females (see Holmberg et al., 1988, p. 525). Resulting waveshapes of the vocal folds, of glottal flow and its time derivative are shown in Figure 5. A quantitative estimation of time and flow parameters of our simulated voice qualities is given in Table 3.

4.3 Results

The parameter values listed in Table 3 indicate that vocal tract excitation occurs for glottal adduction until $x_{mu} = -0.4$ mm for male and until -0.33 mm for female voice and for glottal abduction until 1.0 mm for male voices (while in the case of an added cartilaginous part or "bypass" vocal tract excitation ends between 0.5 mm and 1 mm) and until 0.83 mm for the female voice. Concerning the glottal flow waveshape parameters and its acoustic consequences it can be stated: (i) Values of open quotient OQ increase from maximal adduction (negative values of x_{mu}) towards maximal abduction (positive values of x_{mu}) and OQ is always higher for the female than for the male voice for a given degree of ab-/adduction. These results are in accordance with natural data (e.g. Karlsson and Liljencrants, 1996). It can be concluded that an increase in glottal open period and consequently a decrease in glottal closed period indicates a change of voice quality towards breathy voice and vice versa a decrease towards pressed voice.

(ii) Values of opening quotient $1/2RG$ as well as values of reciprocal opening frequency $1/FG$ increase with increasing abduction and indicate an increasing opening period. But due to the parallel increase in open period this increase of opening quotient does not automatically result in an increase of pulse skewness. Values of pulse skewness SP of glottal flow resulting from our simulations are too high for the male voice and do not show the expected increase with decreasing adduction as it occurs in natural data (see Karlsson and Liljencrants, 1996). Only in the case of the female voice skewness values are comparable to natural data (ibid.). Thus in the natural case an increase in pulse skewness SP is expected for male and female voices if adduction increases (i.e. if the voice quality changes towards pressed voice). This tendency occurs in our simulations at least for the increase of glottal adduction between $x_{mu} = -0.2$ mm to $x_{mu} = -0.4$ mm in the case of the male voice and for the increase in glottal adduction between $x_{mu} = -0.17$ mm to $x_{mu} = -0.33$ mm in the case of the female voice.

(iii) As well as for open quotient also values of excitation time quotient RE increase from maximal adduction towards maximal abduction and these values are always higher for the female voice than those of the male voice for a given degree of ab-/adduction. These trends are in accordance with natural data (Karlsson and Liljencrants, 1996). Thus an increase of the excitation time quotient as well as an increase of the open quotient indicate a change of voice quality towards breathy voice and vice versa.

(iv) Excitation skewness *RK* values are too low in the case of the female voice in comparison to natural data (ibid.), but of the right order of magnitude in the case of the male voice. Excitation skewness naturally increases with increasing adduction. This is reflected in our simulations at least in the case of the change from breathy to normal voice (glottal adduction between $x_{mu} = 0.5$ mm and $x_{mu} = 0$ mm) for the male as well as for the female voice.

(v) Decay time parameters *TA*, *RA*, and reciprocal *FA* increase during glottal abduction while no clear tendencies can be found during glottal adduction for both the male and female voice. Values of decay *TA* are lower than 1.5 ms at $x_{mu} = 0$ for male as well as for female voices and above 1.5 ms in the case of abduction and this trend is in accordance with literature (Karlsson and Liljencrants, 1996; Klatt and Klatt, 1990). Only for strong adduction ($x_{mu} = -0.4$ mm for male and -0.33 mm for female voice) a too high decay time value occurs in our simulations. The normally occurring decrease in *TA*, *RA* and reciprocal *FA* indicates

a shallower spectral tilt of the voice source signal and thus an increase in amplitude of higher harmonics with increasing adduction.
(vi) Peak flow fl_{pk} and leak flow fl_{lk} are increasing during increasing abduction for all positive values of glottal ab-/adduction (i.e. in the case of normal to breathy voice). In the case of added bypass these values are even higher for a given degree of ab-/adduction. AC-flow $fl_{pk} - fl_{lk}$ is between approximately 250 l/s and 350 l/s for the male voice and between 150 l/s and 230 l/s for the female voice if glottal ab-/adduction values are positive (i.e. voice quality between normal and breathy voice). AC amplitude is still high in the case of high positive ab-/adduction values indicating that the audible signal fades away during abduction before the glottal vibration ends. These tendencies are in accordance with natural data (Holmberg et al., 1988). Thus vocal tract excitation fades away earlier than glottal vibration (amplitude) with increasing abduction from normal to breathy voice. For negative values of glottal ab-/adduction (i.e. in the case of the normal to pressed voice continuum) peak flow and AC flow decreases with increasing adduction and no leak flow occurs, which is in accordance with natural data as well (ibid.). Decreasing peak flow as well as AC flow in this case indicates the tendency of decreasing glottal opening with increasing adduction.

5 Discussion

The main result of this study is that our voice source model is flexible enough to cover the parameter ranges of voice source parameters that are defined by the LF model (Fant et al., 1985) for male and female pressed, modal, and breathy voice. Open quotient as well as opening and excitation time quotient as come out from our simulations of female and male voice cover the whole range of natural data (Karlsson and Liljencrants, 1996) and show the expected tendencies during abduction and adduction. Decay time, which reflects the spectral tilt of the source signal, also covers the range of natural data (ibid.) and shows the expected trends during abduction. Only in the case of strong adduction near the end of the breathy-normal-pressed continuum does decay time again increase, thus showing a too high value. Flow pulse skewness as well as excitation skewness values generated by the model do not show clearly the tendencies given in natural data. This may result from the strong ripple occurring with the glottal flow waveshape in the model as a consequence of glottal source vocal tract acoustic interaction (Bavegard and Fant, 1994) which makes it difficult to estimate the time instant of peak flow T_p in the case of the model flow waveshapes.

From the viewpoint of our model it would be possible in addition to vary not only glottal ab-/adduction as was done above but in addition to vary other parameters like phase delay pha_v (Titze, 1989a, p.197) as well as glottal convergence x_d (ibid.). But our experiments indicate that the parameter range for glottal flow parameters is sufficiently modeled by varying glottal ab-/adduction. Furthermore, varying bypass opening area A_c by varying width x_c and length L_c of the cartilaginous part of the glottis has an effect mainly on leak flow and decay time. But the value range for the acoustically important decay time is already sufficiently modeled by changing abduction of the membranous part of the voice source model. It can be concluded from these findings that the Titze (1984) model itself is capable of generating a wide spectrum of different glottal flow shapes and thus is capable of generating a wide value range of acoustically important flow parameters. But the addition of a bypass is important to adjust the overall flow and thus the aerodynamics of the mode which is, for example, important for the generation of frication noise in the vocal tract.

Our first perceptual impressions of the voice qualities generated by the model do not reflect the fact that the model is capable of generating the whole range of voice qualities along the breathy-normal(modal)-pressed continuum. Voice qualities always sound more or less modal. Only a slight effect towards breathy or pressed voice quality occurs even in the case of strong abduction and strong adduction. Two reasons may be responsible for this impression: (i) No glottal noise excitation source is implemented in our model. But the occurrence of glottal noise is important especially in the case of breathy voices (Klatt and Klatt, 1990). (ii) The waveshape of vocal fold vibration are always sinusoidal in the case of the Titze (1984) model since Titze (1984) by definition only allows sinusoidal waveshapes in his model. Only the amplitude and the phase lag can be varied for the vibration of the upper and lower part of the glottis. The sinusoidal character of waveshapes is reflected by unnaturally high decay time values in the case of high adduction. In order to overcome this disadvantage of the model (i) a glottal noise excitation source should be implemented in order to be able to generate convincing breathy voice quality and (ii) to increase the flexibility of possible waveshapes for vocal fold vibration by allowing to mix sinusoidal waveshapes with other types of waveshapes which in particular model very abrupt glottal closure in order to result in lower decay time values and thus higher excitation skewness. If it is the goal to avoid any presetting of a vocal fold waveshape, of its amplitude and of the phase lag between lower and upper part of the vocal folds, an alternative modeling approach would be to take the geometrical part of this model but to control the vocal fold vibration by a specific modeling of aerodynamic and mechanic forces, i.e. to combine the approach given in this paper with a self-oscillating vocal fold model (e.g. Liljencrants, 1996; Titze and Story, 2002; Drioli, 2005; Tao and Jiang, 2008).

6 Acknowledgements

This work was supported in part by the German Research Council Project Nr. Kr 1439/13-1 and project Nr. Kr 1439/15-1.

References

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., and Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30(3):533–553.

Bavegard, B. and Fant, G. (1994). Notes on glottal source interaction ripple. *STL-QPSR, Royal Institute of Technology Stockholm*, (4):63–78.

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese.* Logos-Verlag.

Birkholz, P., Jackèl, D., and Kroger, B. (2007). Simulation of losses due to turbulence in the time-varying vocal system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1218–1226.

Birkholz, P., Jackèl, D., and Kroger, K. (2006). Construction and control of a three-dimensional vocal tract model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse*, volume 1, pages 873–876.

Birkholz, P. and Kröger, B. (2006). Vocal tract model adaptation using magnetic resonance imaging. In *Proceedings of the 7th International Seminar on Speech Production, Belo Horizonte, Brazil*, pages 493–500.

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I., and Lin, Q. (1989). Voice source rules for text-to-speech synthesis. In *Proceedings of the IEEE Conference on Signal Processing*, pages 223–226.

Chan, R. and Titze, I. (2006). Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *The Journal of the Acoustical Society of America*, 119:2351–2362.

Clark, R., Richmond, K., and King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317–330.

Drioli, C. (2005). A flow waveform-matched low-dimensional glottal model based on physical knowledge. *The Journal of the Acoustical Society of America*, 117:3184–3195.

Engwall, O. (2003). Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication*, 41(2):303–330.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR, KTH Stockholm, Sweden*, 4(1985):1–13.

Gobl, C. and Chasaide, A. (1988). The effects of adjacent voice/voiceless consonants on the vowel voice source: a cross language study. *STL-QPSR, KTH Stockholm, Sweden*, 2-3:23–59.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406.

Hanson, H. and Chuang, E. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, 106:1064–1077.

Hanson, H. and Stevens, K. (2002). A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn. *Journal of the Acoustical Society of America*, 112(3):1158–1182.

Hirano, M. (1983). The structure of the vocal folds. In Stevens, K. and Hirano, M., editors, *Vocal Fold Physiology.*, pages 33–43. University of Tokyo, Japan.

Hollien, H. (1960). Vocal pitch variation related to changes in vocal fold length. *Journal of Speech, Language and Hearing Research*, 3(2):150–156.

Hollien, H. and Moore, G. (1960). Measurements of the vocal folds during changes in pitch. *Journal of Speech, Language and Hearing Research*, 3(2):157.

Holmberg, E., Hillman, R., and Perkell, J. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the Acoustical Society of America*, 84:511–529.

Karlsson, I. (1992). *Analysis and synthesis of different voices with emphasis on female speech. (unpubl.)*. PhD thesis, KTH, Stockholm.

Karlsson, I. and Liljencrants, J. (1996). Diverse voice qualities: models and data. In *TMH-QPSR, KTH Stockholm, Sweden*, volume 2, pages 143–146.

Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

Kröger, B. (1989). *Die Synthese der weiblichen Stimme unter besonderer Berücksichtigung der Phonation. (unpubl.)*. PhD thesis, Department of Phonetics, University of Cologne, Germany.

Kröger, B. (1993). A gestural production model and its application to reduction in German. *Phonetica*, 50(4):213–233.

Kröger, B. (1997). Zur artikulatorischen Realisierung von Phonationstypen mittels eines selbstschwingenden Glottismodells. *Sprache-Stimme-Gehör*, 21:102–105.

Kröger, B. (1998). *Ein Phonetisches Modell der Sprachproduktion.* Niemeyer, Tübingen.

Kröger, B. and Birkholz, P. (2007). A gesture-based concept for speech movement control in articulatory speech synthesis. In Esposito, A., Faundez-Zanuy, M., Keller, E., and Marinaro, M., editors, *Verbal and Nonverbal Communication Behaviours.*, pages 174–189. LNAI 4775, Springer, Berlin.

Kröger, B., Kannampuzha, J., and Neuschaefer-Rube, C. (2009). Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809.

Liljencrants, J. (1996). Analysis by synthesis of glottal airflow in a physical model. . *TMH-QPSR, KTH Stockholm, Sweden*, 2:139–142.

Serrurier, A. and Badin, P. (2008). A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *The Journal of the Acoustical Society of America*, 123:2335–2355.

Stevens, K. (1998). *Acoustic Phonetics.* MIT Press.

Story, B. and Titze, I. (1995). Voice simulation with a body cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97:1249–1260.

Sundberg, J. (1987). *The Science of the Singing Voice.* Dekalb, Illinois.

Tao, C. and Jiang, J. (2008). A self-oscillating biophysical computer model of the elongated vocal fold. *Computers in Biology and Medicine*, 38(11-12):1211–1217.

Titze, I. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area. *The Journal of the Acoustical Society of America*, 75:570–580.

Titze, I. (1988). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America*, 83(4):1536–1552.

Titze, I. (1989a). A four-parameter model of the glottis and vocal fold contact area. *Speech Communication*, 8(3):191–201.

Titze, I. (1989b). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America*, 85(2):901–906.

Titze, I. (1989c). Physiological and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, 85:1699–1707.

Titze, I. and Story, B. (2002). Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112:1064–1076.