Simulation of Losses Due to Turbulence in the Time-Varying Vocal System

Peter Birkholz, Dietmar Jackèl, and Bernd J. Kröger

Abstract—**Flow separation in the vocal system at the outlet of a constriction causes turbulence and a fluid dynamic pressure loss. In articulatory synthesizers, the pressure drop associated with such a loss is usually assumed to be concentrated at one specific position near the constriction and is represented by a lumped nonlinear resistance to the flow. This paper highlights discontinuity problems of this simplified loss treatment when the constriction location changes during dynamic articulation. The discontinuities can manifest as undesirable acoustic artifacts in the synthetic speech signal that need to be avoided for high-quality articulatory synthesis. We present a solution to this problem based on a more realistic distributed consideration of fluid dynamic pressure changes. The proposed method was implemented in an articulatory synthesizer where it proved to prevent any acoustic artifacts.**

Index Terms—**Articulatory speech synthesis, fluid dynamics, kinetic pressure loss, vocal system.**

I. Introduction

CONCATENATIVE synthesis is currently the leading approach to high-quality text-to-speech synthesis. Despite its success in generating close-to-natural sounding speech, other approaches to speech synthesis continue to be pursued. In the long term, the most promising of them seem to be articulatory speech synthesis [1]. It is not subject to any fundamental limitations and has applications that are beyond pure text-to-speech synthesis, e.g., articulatory driven facial animation [2] and visual support in second language learning and the therapy of speech disorders [3], [4]. Furthermore, phonetic education and research can benefit from articulatory speech synthesis [5].

Complete articulatory synthesizers are very complex systems, because they require appropriate models to simulate all different aspects of speech production, including models for the generation and propagation of sound, for the anatomy of the vocal system, and for the control of the articulators. There are only few articulatory synthesizers that include all or most of these aspects in one integrated system [5]–[10]. Our synthesizer [10]–[12] has been developed since 2001 and includes implementations for all of the aforementioned components. It is able to generate arbitrary utterances including all sounds of German. The utterances are well comprehensible but still perceived as synthetic. Currently, we try to improve the acoustic quality of the synthesizer so far that it is comparable with the best concatenative synthesizers. The segmental acoustic quality is tightly coupled to the aeroacoustic simulation of the vocal system. This paper deals with a special aspect of the aeroacoustic simulation, namely the simulation of fluid dynamic energy losses due to flow separation and turbulence.

Losses associated with turbulence basically occur in the vicinity of constrictions in the vocal system. Typical constrictions are formed with the vocal folds (the glottis) and with the supraglottal articulators during the production of consonants. When the contraction or expansion of the conduit at a constriction is rather abrupt than gradual, the flow may detach from the tube walls and create regions of turbulence that dissipate energy [13]. The energy losses are manifested as additional pressure losses in the flow [14].

The *contraction* towards the glottis or a supraglottal constrictions in the vocal tract is usually gradual such that no flow separation needs to be assumed in these regions [15]. Therefore, when losses due to friction at the tube wall are neglected, the pressure drop is governed by Bernoulli's equation

$$\Delta p = p_1 - p_2 = \varrho \left(v_2^2 - v_1^2 \right) /2,$$

where p_1 and v_1 are the pressure and flow velocity upstream from the constriction, p_2 and v_2 are the pressure and flow velocity *in* the constriction, and ϱ is the ambient density of air. Here, a decrease of the static pressure p results in an increase of the kinetic pressure $\varrho v^2/2$ without energy dissipation. Also, when the expansion of the constriction is gradual such that the airflow stays laminar, most of the kinetic pressure is transformed back into static pressure and energy losses are minimal. However, the usual case is that the air leaves the constrictions as a jet that creates turbulent air motion and dissipates most of the kinetic energy of the flow in the constriction [16], [17]. So, the pressure drop occurs at the entry of the constriction, but the actual loss occurs at the exit. Part of the energy of the turbulent air motion is usually transformed back into acoustic energy that manifests as noise sources for aspiration (glottal constriction) or frication (supraglottal constrictions).

In acoustic models of the vocal system, the losses caused by flow separation are typically modeled by *lumped* pressure drops or resistances in the corresponding constrictions, both for the glottis (e.g., [18]–[22]) and supraglottal constrictions (e.g., [8], [19], [20], [23], [24]). These supplemental resistances will be referred to as *kinetic resistances* in the following. In contrast to

Manuscript received January 5, 2006; revised September 14, 2006. This work was supported by the German Research Foundation under Research Training Group 466 and Grant JA 1476/1-1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thierry Dutoit.

P. Birkholz and D. Jackèl are with the Institute for Computer Science, University of Rostock, 18059 Rostock, Germany (e-mail: piet@informatik.uni-rostock.de; dj@informatik.uni-rostock.de).

B. J. Kröger is with the Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen (UKA) and Technical University Aachen (RWTH), D-52074 Aachen, Germany (e-mail: bkroeger@ukaachen.de).

Digital Object Identifier 10.1109/TASL.2006.889731

1558-7916/$25.00 © 2007 IEEE

the glottis, the location of a supraglottal constriction is not fixed in relation to the vocal tract tube. Thus, the position of the corresponding kinetic resistance depends on the articulation and is subject to permanent changes during speaking. In a discrete tube model of the vocal system, which is typically applied for articulatory speech synthesis, a changing kinetic resistance position may cause acoustic distortions under certain circumstances, because it involves sudden changes of the acoustic variables at the concerned places. In this paper, we examine this problem and propose a solution that considers fluid dynamic pressure variations not only at discrete constriction locations but all along the vocal tube.

This paper is organized as follows. Section II introduces our articulatory synthesizer with emphasis on the aeroacoustic simulation. In Section III, we discuss the problems resulting from the application of a lumped kinetic pressure loss in a time-varying vocal tract and their causes. Section IV presents a new method for the consideration of fluid dynamic factors and its implementation in a transmission line circuit model of the vocal system. Conclusions are drawn in Section V.

II. Articulatory Synthesizer

The problem discussed in this paper was encountered during the synthesis of a variety of utterances with an articulatory speech synthesizer that we developed during the last few years. However, the fundamental problem, which is described in detail in Section III, is relevant to all simulations of the time-varying vocal system in the time-domain that include losses due to turbulence—independent of the particular implementation. This section gives a brief overview of our synthesizer. Currently, the system is able to generate arbitrary utterances containing all sounds of German from an organized pattern of articulatory gestures as input. The next subsections describe the individual modules of the synthesizer with the focus on the acoustic model.

A. *Acoustic Model*

Acoustically, the vocal system is modeled as a branched nonuniform tube that includes the vocal tract, the glottis, the nose cavity, the paranasal sinuses, and the subglottal tract. This tube system is approximated in terms of incremental contiguous tube sections. Each tube section has an individual length and a uniform elliptical cross section given by its area and perimeter. In contrast to circular cross sections, two tube sections with elliptical cross sections of the same area can have different perimeters that partly determine the acoustic resistances. We use 13 tube sections for the subglottal tract, two sections for the glottis, 40 sections for the vocal tract, 19 sections for the nose cavity, and four sections for the paranasal sinuses. Fig. 1 shows an example of the piecewise constant area function of the tube system that represents the vowel [o]. The tube sections for the nasal cavity are flipped upside down, and the sections of the paranasal sinuses are displayed as circles at their coupling locations. Within the tube we assume plane wave propagation.

There are different techniques for the simulation of sound propagation in a discrete tube model. The most common techniques are based on wave digital filters (e.g., [23], [25]), on the direct numerical simulation of the transmission line circuit model (TLM) of the vocal tract (e.g., [18], [26]), or on a hybrid time-frequency simulation of the vocal system (e.g., [20], [27]). Each method has its individual strengths and weaknesses. Our acoustic simulation is based on the direct numerical simulation of the TLM with lumped elements. The TLM is easily interpretable and very descriptive, because it is directly based on the analogy between acoustic and electrical systems. Furthermore, it makes no restrictions regarding the lengths of the individual tube sections.

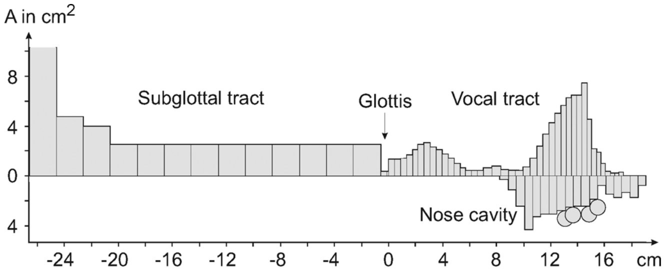

Fig. 1. Discrete area function of the vocal system for the vowel [o] in our articulatory synthesizer. The cross-sectional areas of the nose cavity are flipped upside down. The circles represent four paranasal sinuses.

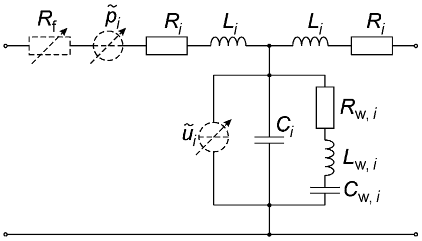

Fig. 2. Two-port network for one tube section. The network elements are described in the text.

1) Transmission Line Circuit Model: Using the TLM with lumped elements, the whole vocal system can be represented in a uniform fashion. Each tube section is represented by a two-port network as in Fig. 2. In this circuit analogy, voltage is equivalent to pressure, and current is equivalent to volume velocity. L_i is the inertance of the mass of air in the tube section i, C_i represents its compressibility, and R_i accounts for energy lost to viscous friction at the tube walls. The $R_{\mathrm{w},i} - L_{\mathrm{w},i} - C_{\mathrm{w},i}$ circuit models the elasticity of the vocal tract walls. The optional elements $\tilde{u}_i, \tilde{p}_{i,}$ and R_{f} constitute a volume velocity source, a pressure source and a resistance for the kinetic pressure drop at the main constriction and will be discussed below. The derivation of the element values can be found in [28], and they are summarized for our synthesizer in [10], [11].

The paranasal sinuses are modeled as discrete Helmholtz resonators. They can be represented by the same network as in Fig. 2 with the difference that only one port is connected to the transmission line. Here, R_i and L_i represent the resistance and inductance of the air in the resonator neck, and C_i is the capacity of the air in the resonator cavity.

The network for the whole vocal system results from the concatenation of the two-ports for the individual tube sections and is shown in Fig. 3. At the mouth and the nostrils, the network is terminated with a radiation impedance that was realized as a parallel R–L-circuit [18]. At the peripheral end of the subglottal

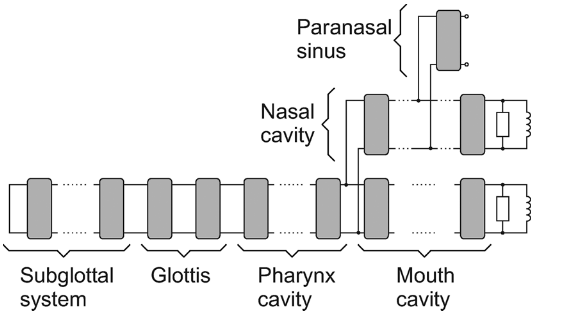

Fig. 3. Transmission line circuit model for the entire vocal system. Each gray box represents a two-port network as in Fig. 2 for one tube section.

tract, the network is terminated with a short circuit that approximates the high acoustic capacity of the pulmonary alveoli. A pressure source in the first subglottal section simulates the pulmonary pressure. Voiced excitation is automatically obtained by forced oscillations of the cross-sectional areas of the upper and lower glottal tube sections.

2) Noise Source Model: Voiceless excitation of the vocal tract is caused by the complex interaction between the sound field and nonacoustic air motions such as turbulence. Accurate and approved physical models to simulate this interaction in the context of articulatory speech synthesis do not yet exist. All present models for voiceless excitation are based on a rather gross simplification of the real physical process.[1] They simply model flow-induced noise as random fluctuations in either pressure or velocity, usually by means of lumped noise sources (e.g., [20], [31]–[33]). Despite their simplicity, some of these models are able to produce remarkably good results. The crucial point for good results is a proper parameterization of the noise sources in terms of their number, positions, levels, and spectra. Despite a wealth of theoretical and experimental research, there is no general agreement about these details. However, there is extensive agreement about the basic mechanisms of turbulence noise generation at a constriction in the vocal tract [16], [33], [34]. Stevens [16] summarizes at least three such mechanisms.

1) The airflow emerging from the constriction forms a jet and creates turbulent velocity fluctuations in the region downstream from the constriction exit.
2) Irregularities in the constriction geometry cause random velocity fluctuations within the constriction.
3) The rapid airflow impinges at an obstacle or surface oriented normal to the flow which generates fluctuating forces on the medium that in turn constitute a source of sound. This mechanism is the most efficient one in terms of the produced sound pressure [16] compared to the first and second mechanism. The efficiency is furthermore dependent on the orientation of the obstacle and the distance between the jet nozzle and the obstacle. Jets that impinge at a normal direction to the obstacle result in a greater fluctuating force on the medium than those that impinge at smaller angles (i.e., $<90°$). As the distance to the obstacle becomes greater, the jet widens and the particle velocity and sound source strength drops [33].

Random velocity fluctuations according to the first two mechanisms constitute a flow monopole source and can be modeled with a volume velocity source in the TLM. Fluctuating forces according to the third mechanism constitute a dipole source and can be modeled with a pressure source (cf. Fig. 2 for the source positions in the TLM). In [35], we proposed a noise source model based on the above mechanisms that defines quantitative relations between the flow conditions in the constriction and the parameters of the noise sources. The parameters were derived empirically by means of analysis-by-synthesis experiments (i.e., we tried to match real and synthetic fricative spectra by systematic parameter variation). The model can be summarized as follows.

At any time, we assume that there is at most one supraglottal constriction in the vocal tract from which a turbulent jet emerges. The noise caused by this jet is modeled with one volume velocity source $\tilde{u}$ at the exit of the constriction (flow separation point), and one pressure source $\tilde{p}$ in the section, where the jet is assumed to hit an obstacle or surface. The suitability of this approach for the synthesis of fricatives is also substantiated by Narayanan and Alwan [33]. The easiest way to determine the origin of the jet would be to select the tube section with the smallest cross-sectional area. This may, however, be ambiguous when more than one constriction exists in the vocal tract for the same articulation. The fricative $/\int/$, for example, has two constrictions, one at the tongue tip and one at the teeth, that are separated by a sublingual cavity [36]. In this case, the tongue constriction is supposed to be the origin of the jet. Currently, we search for one potential jet constriction in each of two disjunct parts of the vocal tract: one posterior part from the glottis to the tongue tip and one anterior part from the tongue tip to the lips. The constriction in each part is composed of one or more consecutive tube sections whose areas are below the minimum area in that part plus a small threshold of 0.2 cm^2. Thus, a constriction may be composed of a cascade of tube sections when they form a narrow channel. Whether the posterior or the anterior constriction is chosen as origin for the jet is decided on the basis of the cross-sectional area *and* the length of the constrictions. A constriction is more likely to be chosen, when it is as long and as narrow as possible [10]. For the fricative $/\int/$, this will most probably be the tongue constriction.

As soon as one of the constrictions was chosen, the model determines the positions of the monopole and the dipole source. The monopole source is always placed in the most anterior section of the constriction, where we assume the flow to detach. The dipole source is always placed at the location of a representative obstacle in the path of the jet. It is placed at the teeth, when the distance from the flow separation point (FSP) to the teeth is shorter than 4 cm, as it is for the alveolar and postalveolar fricatives. Otherwise, e.g., for velar fricatives, it is placed 0.5 cm downstream from the FSP. In the latter case, we assume the vocal tract walls to act as the obstacle. When the FSP is at or downstream from the teeth (/f/ and /v/), the dipole source is placed at the position of the lips.

The noise sources are only activated under certain flow conditions. A widely used criterion for the generation of turbulence in articulatory synthesizers is the squared Reynolds number Re^2

[1] To our knowlege, the only attempts to model the physical process underlying noise generation in the vocal tract are due to Sinder [29] and Krane [17]. Their models are based on theoretical results in aerodynamic theory by Howe [30].

in the constriction (e.g., [5], [8], [20], [32]). When Re^2 is below a certain threshold Re^2_{crit}, no noise is generated. Otherwise, the radiated noise sound pressure is proportional to $Re^2 - Re^2_{\text{crit}}$. In our model, we use this fundamental dependence to determine both the amplitude $\tilde{u}$ of the monopole source and the amplitude $\tilde{p}$ of the pressure source according to the empiric formulas

$$\begin{aligned} \Phi &= \begin{cases} \alpha \cdot \left(Re^2 - Re^2_{\text{crit}}\right), & \text{for } Re > Re_{\text{crit}} \\ 0, & \text{otherwise} \end{cases} \\ \tilde{p} &= \Phi \cdot \eta \cdot e^{-d/\tau} \\ \tilde{u} &= \beta \cdot \Phi \end{aligned}$$

where $Re = v_{\text{c}} d_{\text{c}}/\nu$ is the Reynolds number of the flow in the constriction, and $Re_{\text{crit}} = 1800$ is the critical Reynolds number. v_{c} is the velocity in the narrowest tube section of the constriction, d_{c} its diameter, and ν the kinematic viscosity. Φ can be interpreted as the strength of a dipole source that evolves, when the obstacle in the path of the jet is immediately downstream from the constriction and oriented normal to the flow. The factor η accounts for the attenuation of the noise, when the jet hits the obstacle at an angle smaller than 90°. We set $\eta = 1$ when the incisors act as obstacle and $\eta = 0.5$ for the vocal tract walls or the lips. With these adjustments, nonstrident fricatives like /f/ and /x/ generate less noise than strident fricatives like /s/ and / ʃ /. The factor $e^{-d/\tau}$ causes a decrease of the noise strength with an increasing distance d between the FSP and the obstacle. The reference $\tau = 1.23$ cm and the constants $\alpha = 4 \cdot 10^{-6}$ Pa and $\beta = 5 \cdot 10^{-8}$ m^5/Ns were determined empirically.

The spectra of the noise sources in our model are shaped with a second-order Butterworth low-pass filter. So they are flat up to a certain break frequency and have a slope of −12 dB/oct above this frequency. For the monopole source we use a constant break frequency of 1100 Hz. The resulting spectrum is thereby very similar to a real measured monopole spectrum (cf. Fig. 3 in [33]). For the dipole sources, we assume a variable break frequency $f_{-3\text{ dB}} = kv_{\text{c}}/d_{\text{c}}$, where $k = 0.4$ is a constant [16].

Finally, turbulence needs time to establish as soon as the appropriate conditions are met. This means, it takes some time from the point, when the Reynolds number exceeds its critical value, to the point, when the noise sources achieve their full strength. We simulate this by passing the noise source amplitudes $\tilde{p}$ and $\tilde{u}$ through a recursive single pole low-pass filter that delays the onset and offset of noise generation due to its group delay. At the same time, this filter prevents acoustic distortions owing to a "hard" noise onset and offset.

As stated before, the introduction of noise sources in the case of turbulence only models the acoustic result of a complex aeroacoustic process. In this process, energy from the potential irrotational flow field in the vocal tract is consumed for the formation and convection of vortices that in turn spend some of their energy for the generation of noise [17]. The energy consumption for the formation and convection of the vortices manifests as a kinetic pressure loss in the turbulent region. The pressure loss is approximately equal to the dynamic pressure in the constriction, i.e., $\Delta p \approx \varrho v^2_{\text{c}}/2$. The corresponding pressure drop is usually regarded to be concentrated at one specific place in the vocal tract [8], [19], [20], [23], [24]. Using the TLM, it can be modeled with a supplemental kinetic resistance $R_{\text{f}} = \varrho |u_{\text{c}}|/2A^2_{\text{c}}$ in the T-network of the tube section forming the constriction (cf. Fig. 2), where A_{c} is the cross-sectional area of the constriction, and u_{c} is the volume velocity through the constriction. In the first version of our synthesizer, we chose to implement the kinetic pressure loss this way, too. However, the drawbacks of this approach with a time-varying vocal tract (Section III) brought us to devise and implement a new method to consider kinetic losses (Section IV).

3) Numerical Simulation: The TLM is a linear network that can be described by a system of coupled ordinary differential equation. For the digital simulation, these equations are approximated by difference equations and solved simultaneously at a rate of 44 kHz [10], [11]. The speech signal is approximated as the first derivative of the volume velocities through the radiation impedances at the nostrils and the mouth opening.

For the temporal discretization, we use the theta method [37], which is a generalization of Euler's method and the trapezoid rule. With this method, we are able to influence the formant bandwidths by numerical damping. In [11], we have shown that the influence of the frequency-dependent boundary-layer resistance in the vocal tract can be simulated by carefully choosing the parameter θ of the discretization scheme.

B. Anatomic Models

We use different anatomic submodels to generate the tube geometry for the subglottal tract, the glottis, the vocal tract, and the nasal cavity. The geometry of the subglottal tract and the nasal cavity is directly given by the respective discrete area and perimeter functions. The subglottal geometry is completely static and a coarse approximation of the model proposed by Ishizaka *et al.* [38]. Apart from the first three tube sections, the nasal tube is also static and modeled after Dang and Honda [39], [40]. The first three sections represent the region above the velum and change their cross sections corresponding to the state of the velum.

The vocal tract is modeled with a 3-D geometric model shown in Fig. 4(a) and described in [12]. The transformation from the vocal tract geometry to the cross sections of the discrete supraglottal tube sections starts with the calculation of the center line. The vocal tract model is then intersected with 64 equidistant planes orthogonal to the center line. The center line and the parallel projection of the section planes is shown in Fig. 4(b). From each cross section of the vocal tract model, the area and the perimeter are calculated. The resulting piecewise linear area and perimeter functions [Fig. 4(c) and (d)] are then discretized to yield the tube section parameters. We use 16 equal sampling intervals/tube sections for the region between the glottis and the velopharyngeal port and from there on to the mouth 24 intervals with decreasing length. Thereby we get a finer spatial sampling in the region of the incisors and lips, which is especially important for an accurate representation of the geometry of fricative constrictions in this area. Fig. 4(c) and (d) shows the piecewise linear area and perimeter functions for the vowel [a] obtained with our model.

The representation of the vocal folds and their motion is based on the geometric glottis model by Titze [41]. The input parameters for the model are the degree of abduction at the posterior

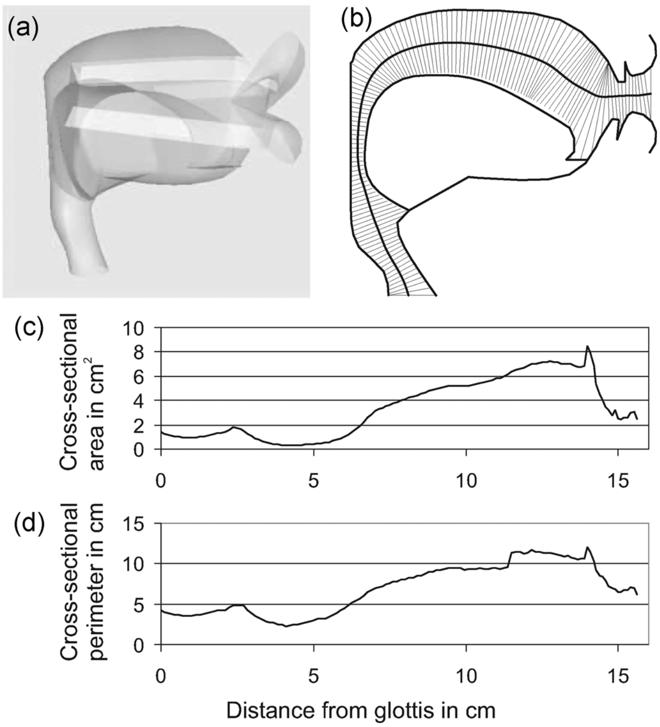

Fig. 4. Three-dimensional geometric model of the vocal tract. (a) Rendering for the vowel [a]. (b) Center line and section planes. (c) Area function. (d) Perimeter function.

upper and lower edge of the vocal folds, the fundamental frequency, the subglottal pressure, and the phase lag between the upper and lower edge of the vocal folds. The 3-D shape of the model is transformed into the two corresponding elliptical tube sections in the discrete tube model—one for the lower part of the glottis and one for the upper part. The original model by Titze [41] was only designed to simulate and analyze phonation. There, the amplitude of vocal fold motion was only a function of subglottal pressure and fundamental frequency. We have introduced a further dependence of the amplitude on the degree of abduction, such that the vibration gradually ceases when the vocal folds are sufficiently abducted, as in voiceless fricatives or plosives [10].

C. Control Model

The control model of our articulatory synthesizer is based on the concept of articulatory gestures [42], [43]. Each utterance must be specified in terms of a gestural score, which is an organized pattern of gestures. The control model transforms a gestural score into a continual sequence of parameter values for the vocal tract and vocal fold parameters. These parameters are in turn transformed into the discrete tube geometry by means of the 3-D models for the vocal tract and the glottis. Our gestural model is somewhat simpler than the task-dynamic approach by Saltzman and Munhall [43] and the model by Kröger *et al.* [44] but nevertheless effective in generating plausible coarticulatory motion patterns. A detailed description of the model is given in [12].

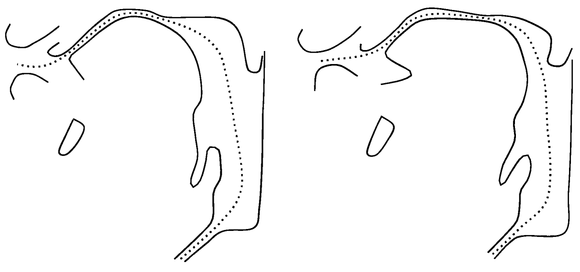

Fig. 5. MRI tracings of the fricative [ʃ] in [i]-context (left) and [u]-context (right) [45]. The dotted lines represent the vocal tract center lines.

III. Problems With a Lumped Kinetic Pressure Loss in the Supraglottal Tract

The articulatory synthesizer described in the previous section was tested with a variety of single static sounds, syllables, and complex utterances containing all types of sounds of standard German including fricatives and plosives. In some utterances, the synthesis of fricatives and plosives was accompanied by the generation of acoustic artifacts in form of short disturbing click sounds, especially during phoneme-to-phoneme transitions. The noise sources could be excluded as causes for these artifacts, because they also appeared for the same utterances when all noise sources were internally "turned off." A careful observation of the time-varying pressure and volume velocity distribution in the vocal system during the simulations revealed that these clicks were generated during the change of the tube section with the smallest cross-sectional area. As we discussed in Section II-A, it is always the supraglottal tube section with the smallest area that causes an additional pressure loss due to the lumped kinetic resistance R_f. When the position of this resistance changes in the TLM, discontinuities in the acoustic variables are generated both at the old and the new position. When the constrictions are sufficiently narrow and R_f correspondingly high, these discontinuities manifest as the click sounds in the synthetic speech signal. For constriction areas greater than approximately 0.3 cm^2, no audible artifacts are generated.

We identified three causes for a switch-over of the kinetic resistance position.

1) The lengths of the cavities upstream and downstream from the constriction of a fricative change due to coarticulation. These length changes result from the protrusion/retraction of the lips, from lifting/lowering of the larynx, and from changes of the tongue body position. Fig. 5 shows midsagittal MRI tracings of the fricative [ʃ] in the context of the vowels [i] and [u]. Obviously, the shape of the front and back cavity as well as the overall length of the vocal tract is different in both situations. When [ʃ] is produced in the nonsense utterance [iʃu], the length of the vocal tract and so of the individual tube sections might change during the constriction interval of the fricative. This can result in a change of the tube section containing the kinetic resistance, as it is illustrated in Fig. 6. The left column shows a constriction in the vocal tract continually moving from the right to the left and the corresponding sections of the discrete tube model. When the constriction position reaches

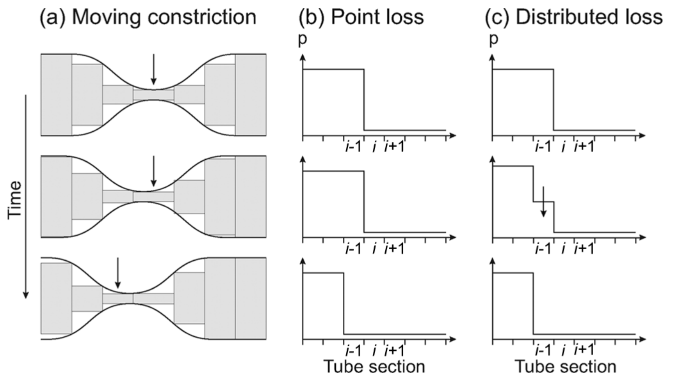

Fig. 6. (a) Constriction continually moving from right to left. The arrows point to the tube sections with the smallest cross-sectional areas. (b) Sudden switch-over of the pressure drop point. (c) Gradual change of the pressure drop.

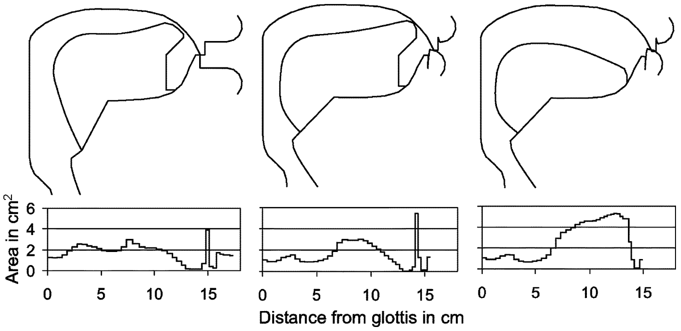

Fig. 7. Vocal tract profiles and area functions for [ʃ] (left), [v] (right) and the transition between [ʃ] and [v] (middle) in the word [ʃva].

the boundary between two sections, the narrowest tube section (arrow) suddenly changes. The stylized change of the static pressure distribution in this tube using a lumped kinetic resistance is shown in the middle column. Between the second and the third time instant, a sudden change of the pressure distribution occurs that can cause a transient in the acoustic signal.

2) The main supraglottal constriction of the vocal tract is ambiguous for the aerodynamic-acoustic algorithm. This can have two reasons. On one hand, there may be multiple constrictions in the vocal tract for the same articulation as discussed in Section II-A2 for the fricative / ʃ /. In this case, it is not clear where the main kinetic pressure drop occurs—at the incisors or at the tongue tip. On the other hand, a fricative constriction may extend over several (more than one) tube sections with approximately equal cross-sectional areas. If so, even small coarticulatory changes of the area function during the constriction interval suffice to cause a switch-over of the kinetic resistance location. Points 1) and 2) are especially relevant for rather short tube sections, as in our simulation. With longer tube sections (e.g., 0.5 cm or greater), shifts of the constriction between adjacent sections may be less frequent. However, acoustic simulations with short sections are more accurate for high frequencies. For the following third point, the tube section length is irrelevant.
3) The constriction location switches between the points of articulation of two consecutive fricatives or plosives in a consonant cluster. This case is illustrated for the consonant cluster [ʃv] in the word [ʃva] in Fig. 7. The left sagittal profile shows the consonant [ʃ] with one constriction at the tongue tip and one at the incisors, the right profile shows the consonant [v] with only one labio-dental constriction, and the profile in the middle shows the instant during the transition, where the main constriction moves from the tongue tip to the incisors. These vocal tract profiles were generated with the gestural control model of our articulatory synthesizer [12]. Fig. 8(a) shows the synthetic speech signal for the utterance [ʃva], where the transient sound caused by the change of the kinetic resistance position between [ʃ] and [v] is well visible.

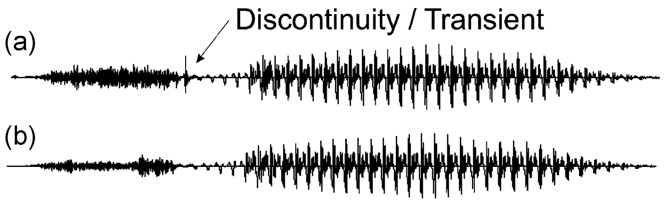

Fig. 8. (a) Speech signal for the utterance [ʃva] using a single lumped kinetic resistance. (b) Same utterance synthesized with the proposed loss model.

IV. Proposed Method for the Consideration of Fluid Dynamic Losses

A. Basic Principles

In this section, we present a simple but effective way to incorporate fluid dynamic pressure losses that prevents the generation of acoustic artifacts due to changing constriction locations. Our solution is based on two principles for the junctions between adjacent tube sections.

1) When the downstream section of any two adjacent tube sections has a *smaller* area than the upstream section, the pressure drop is considered according to Bernoulli's law. Hence, in a converging duct we assume continuity of total pressure $p + \varrho v^2/2$.
2) When the downstream section of any two adjacent tube sections has a *greater* area than the upstream section, all the dynamic pressure in the narrower tube section is lost. Hence, in a diverging duct we assume continuity of static pressure p.

These two principles are not only meant to apply for the supraglottal tube sections, but also for the glottal sections. The first principle is immediately plausible when we exclusively assume *smooth* transitions from a wide cross section to a narrow one. In this case, the flow does not detach from the walls (vena contracta effect) and can be described by Euler's equation of motion, which becomes Bernoulli's law for a stationary flow. The only place where the vena contracta effect is frequently presumed in simulations of the vocal system is at the inlet of the glottis (e.g., [18], [20]), where it causes a somewhat higher pressure drop than predicted by Bernoulli's law. However, in more recent works, the occurrence of this effect is doubted due to the actually smooth area transition between the trachea and the glottis [15], [46]. According to the first principle, the static

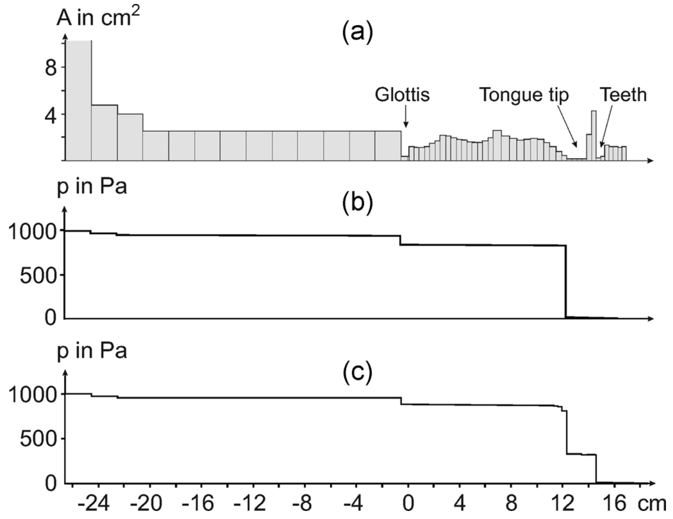

Fig. 9. Static pressure distribution in the vocal tract for the fricative [ʃ]. (a) Area function. (b) Distribution with a single supraglottal kinetic resistance. (c) Distribution with the proposed model.

pressure now drops gradually due to the gradual decrease of the cross-sectional areas approaching a constriction.

The second principle assumes that the kinetic pressure completely dissipates when the flow passes the junction from a narrow to a wide tube section. At the outlet of the glottis, this is a good approximation for the losses caused by flow separation and turbulence. Actually, less than 20% (and over most of the glottal cycle less than 10%) of the kinetic pressure in the glottis is recovered after exit [47]. Some researchers nevertheless simulate this pressure recovery on the basis of the conservation of momentum [18], [22], [48]. With regard to the glottis, the proposed principles also reflect the fact that the flow separation point changes during a glottal cycle. According to Pelorson *et al.* [15], [46], flow separation occurs very near the outlet of the glottis when the vocal folds open (convergent shape) while the flow-separation point moves upstream to the glottis inlet when the vocal folds close (slightly divergent shape). Our glottis model consists of an upper and a lower tube section with the areas A_{lower} and A_{upper}. According to the proposed principles, the flow detaches at the glottal outlet, when $A_{\text{lower}} > A_{\text{upper}}$, and at the junction between the two sections otherwise.

According to the second principle, the kinetic pressure is also lost at every junction from a narrow to a wide tube section in the supraglottal tract. When both adjacent tube sections are relatively wide (> 0.5 cm^2) the pressure loss due to this principle is essentially negligible compared with other distributed losses. However, when the upstream section is relatively narrow, the kinetic pressure loss becomes more dominant and approximates well the energy loss of the lumped kinetic resistance of the earlier models.

Fig. 9(a) shows the area function for the consonant [ʃ] with a wide open glottis. When both of the above principles are applied, the static pressure distribution depicted in Fig. 9(c) evolves. Acoustic pressure disturbances caused by noise sources were excluded for this simulation. We observe major pressure drops at the glottis, the tongue tip, and the teeth. At the tongue tip constriction, this drop is relatively gradual, because the contraction of the tube is gradual, too. In contrast, the same simulation with a single supraglottal kinetic resistance in Fig. 9(b) causes only two hard pressure losses at the glottis and the tongue tip. When the constriction location(s) move and the proposed method is used, the change in acoustic variables will not happen suddenly as in Fig. 6(b), but continual as in Fig. 6(c).

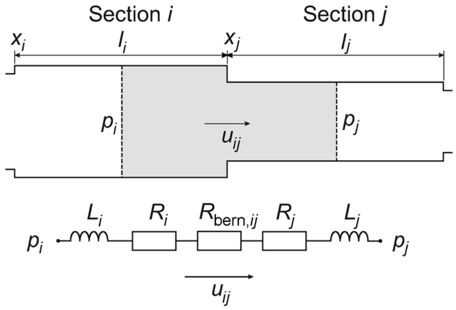

Fig. 10. Discretization of Euler's equation of motion. The gray region between the centers of the two adjacent tube sections is represented by the corresponding electrical network (bottom picture).

B. Implementation

With regard to the TLM, the proposed principles to simulate the fluid dynamic losses can be implemented as shown in Fig. 10. Let A_i and A_j be the cross-sectional areas of two consecutive tube sections i and j. According to the first principle, an additional resistance

$$R_{\text{bern},ij} = |u_{ij}|\frac{\varrho}{2}\left(\frac{1}{A_j^2} - \frac{1}{A_i^2}\right)$$

must be inserted at the boundary between the tube sections, whenever $A_i > A_j$, where u_{ij} is the volume velocity between the sections, and ϱ is the ambient density of air. A derivation of this resistance can be found in [49]. According to the second principle, this resistance must not be inserted when $A_i \leq A_j$.

For the implementation of the TLM, it is convenient to split up the additional resistance in its two summands and assign the summand containing the term $1/A_i^2$ to tube section i and the other one to section j. Doing this, each tube section can be represented by a two-port network as in Fig. 11, where

$$R_i = \begin{cases} R_{\text{fric},i} + |u_{\text{in},i}|\varrho/\left(2A_i^2\right), & \text{when } A_i < A_{\text{pred}} \\ R_{\text{fric},i}, & \text{otherwise} \end{cases}$$

and

$$\overline{R}_i = \begin{cases} R_{\text{fric},i} - |u_{\text{out},i}|\varrho/\left(2A_i^2\right), & \text{when } A_{\text{succ}} < A_i \\ R_{\text{fric},i}, & \text{otherwise.} \end{cases}$$

$u_{\text{in},i}$ and $u_{\text{out},i}$ are the volume velocities entering and leaving the tube section, A_{pred} and A_{succ} are the cross-sectional areas of the tube sections upstream and downstream of i, $R_{\text{fric},i}$ is the resistance representing viscous friction at the tube wall, and the remaining network elements are the same as in Fig. 2. Modeling the tube sections in this way, bifurcations in the vocal system, for instance at the velopharyngeal port, can be represented analogous to Fig. 3.

With the integration of the kinetic resistances, the TLM is actually no more a linear system, because R_i and $\overline{R}_i$ are now

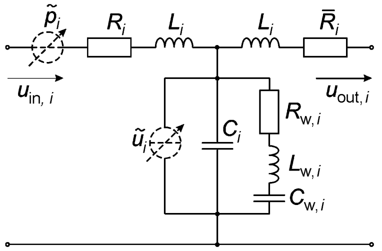

Fig. 11. Two-port network for a single tube section in the extended TLM.

a function of the volume velocity. However, for the numerical simulation of this modified TLM in the time-domain, the R_i and $\overline{R}_i$ can simply be calculated with the known volume velocity values from the previous time step without the risk of numerical instabilities.

The application of the proposed principles to treat fluid dynamic losses prevents the emergence of acoustic artifacts as in Fig. 8(a). Fig. 8(b) shows the utterance [ʃva] synthesized with the proposed method. Despite other minor visual waveform differences, the transient in Fig. 8(a) constitutes the only audible difference between the two waveforms. The method was tested with a number of synthetic utterances containing both fricatives and fricative clusters in vocalic context. The audio files of the test examples can be found on the Internet at http://wwwicg.informatik.uni-rostock.de/~piet/fdl /examples.html. The utterances were also synthesized without the activation of the noise sources such that potential acoustic artifacts covered with noise could have been detected in the speech signal waveform. However, no such cases were found.

V. Conclusion

The consideration of losses due to turbulence in the time-varying vocal system has been examined. It has been shown that the application of a single lumped kinetic resistance in a supraglottal constriction may cause acoustic artifacts when the constriction location changes during dynamic articulation. For high-quality articulatory speech synthesis, these artifacts must be avoided. Therefore, we have introduced a new method that considers nonlinear fluid dynamic pressure changes along the whole vocal tube and implemented it in an articulatory synthesizer based on the transmission line circuit model. The new method is based on two principles that define the continuity of either total pressure or static pressure at the boundary of adjacent vocal tract tube sections. The method is consistent with the fluid dynamic simulation of the glottal constriction in other works and was substantiated for the application in the supraglottal tract. It is also well suited for vocal tract configurations with multiple or ambiguous constrictions. Tested on a variety of utterances containing fricatives and consonant clusters, the new method was able to prevent any acoustic distortions, and is thus a step towards a higher quality of dynamic articulatory speech synthesis.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: a position paper," in *Proc. 4th ISCA Tutorial Res. Workshop Speech Synth.*, Pitlochry, U.K., 2001, pp. 121–126.

[2] P. Badin, P. Borel, G. Bailly, L. Revéret, M. Baciu, and C. Segebarth, "Towards an audiovisual virtual talking head: 3d articulatory modeling of tongue, lips and face based on mri and video images," in *Proc. 5th Seminar Speech Production: Models and Data and CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, 2000, pp. 261–264.

[3] O. Engwall, P. Wik, J. Beskow, and G. Granström, "Design strategies for a virtual language tutor," in *Proc. 8th Int. Conf. Spoken Lang. Proc.*, Jeju Island, Korea, 2004, vol. 3, pp. 1693–1696.

[4] B. J. Kröger, J. Gotto, S. Albert, and C. Neuschaefer-Rube, "A visual articulatory model and its application to therapy of speech disorders: a pilot study," *ZAS Papers in Linguistics (ZASPiL)*, vol. 40, pp. 79–94, 2005.

[5] A. J. S. Teixeira, R. Martinez, L. N. Silva, L. M. T. Jesus, J. C. Principe, and F. A. C. Vaz, "Simulation of human speech production applied to the study and synthesis of european portuguese," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1435–1448, 2005.

[6] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Amer.*, vol. 115, no. 2, pp. 853–870, 2004.

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

[8] B. J. Kröger, *Ein phonetisches Modell der Sprachproduktion*. Tübingen, Germany: Niemeyer, 1998.

[9] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *J. Acoust. Soc. Amer.*, vol. 70, no. 2, pp. 321–328, 1981.

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

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

[12] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *Int. Conf. Acoust, Speech, Signal Process. (ICASSP'06)*, Toulouse, France, 2006, pp. 873–876.

[13] P. S. Barna, *Fluid Mechanics for Engineers*. London, U.K.: Butterworths, 1957.

[14] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: static considerations," *J. Acoust. Soc. Amer.*, vol. 50, no. 4, pp. 1180–1191, 1971.

[15] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands, and Y. Auregan, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," *J. Acoust. Soc. Amer.*, vol. 96, no. 6, pp. 3416–3431, 1994.

[16] K. N. Stevens, *Acoustic Phonetics*. Cambridge, U.K.: MIT Press, 1998.

[17] M. H. Krane, "Aeroacoustic production of low-frequency unvoiced speech sounds," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 410–427, 2005.

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

[19] P. Badin and G. Fant, "Notes on vocal tract computation," *STL-QPSR*, vol. 2–3, pp. 53–108, 1984.

[20] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 7, pp. 955–967, 1987.

[21] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 97, no. 2, pp. 1249–1260, 1995.

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

[23] J. Liljencrants, "Speech synthesis with a reflection-type line analog," Ph.D. dissertation, R. Inst. Technol., Stockholm, Sweden, 1985.

[24] K. Mawass, P. Badin, and G. Bailly, "Synthesis of french fricatives by audio-video to articulatory inversion," *Acustica*, vol. 86, pp. 136–146, 2000.

[25] P. Meyer, R. Wilhelms, and H. W. Strube, "A quasiarticulatory speech synthesizer for German language running in real time," *J. Acoust. Soc. Amer.*, vol. 86, no. 2, pp. 523–540, 1989.

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

[27] D. R. Allen and W. J. Strong, "A model for the synthesis of natural sounding vowels," *J. Acoust. Soc. Amer.*, vol. 78, no. 1, pp. 58–69, 1985.

[28] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Berlin, Germany: Springer-Verlag, 1965.

[29] D. J. Sinder, "Speech synthesis using an aeroacoustic fricative model," Ph.D. dissertation, Rutgers Univ., New Brunswick, NJ, 1999.

[30] M. S. Howe, "Contributions to the theory of aerodynamic sound, with application to excess jet noise and the theory of the flute," *J. Fluid Mech.*, vol. 71, no. 4, pp. 625–673, 1975.

[31] J. L. Flanagan and L. Cherry, "Excitation of vocal-tract synthesizers," *J. Acoust. Soc. Amer.*, vol. 45, no. 3, pp. 764–769, 1969.

[32] P. Boersma, "Functional phonology," Ph.D. dissertation, Univ. Amsterdam, The Netherlands, 1998.

[33] S. Narayanan and A. Alwan, "Noise source models for fricative consonants," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 328–344, 2000.

[34] C. H. Shadle, "The effect of geometry on source mechanisms of fricative consonants," *J. Phonetics*, vol. 19, pp. 409–424, 1991.

[35] P. Birkholz and D. Jackèl, "Noise sources and area functions for the synthesis of fricative consonants," *Rostocker Inform. Berichte*, vol. 30, pp. 17–23, 2006.

[36] S. S. Narayanan, A. A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 98, no. 3, pp. 1325–1347, 1995.

[37] A. Iserles, *A First Course in th Numerical Analysis of Differential Equations*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[38] K. Ishizaka, M. Matsudaira, and T. Kaneko, "Input acoustic-impedance measurement of the subglottal system," *J. Acoust. Soc. Amer.*, vol. 60, no. 1, pp. 190–197, 1976.

[39] J. Dang and K. Honda, "Morphological and acoustical analysis of the nasal and the paranasal cavities," *J. Acoust. Soc. Amer.*, vol. 96, no. 4, pp. 2088–2100, 1994.

[40] ——, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *J. Acoust. Soc. Amer.*, vol. 100, no. 5, pp. 3374–3383, 1996.

[41] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc. Amer.*, vol. 75, no. 2, pp. 570–580, 1984.

[42] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

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

[44] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *J. Acoust. Soc. Amer.*, vol. 98, no. 4, pp. 1878–1889, 1995.

[45] B. J. Kröger, P. Hoole, R. Sader, C. Geng, B. Pompino-Marschall, and C. Neuschaefer-Rube, "MRT-Sequenzen als Datenbasis eines visuellen Artikulationsmodells," *Hals-Nasen-Ohren Heilkunde*, vol. 52, pp. 837–843, 2004.

[46] X. Pelorson, C. Vescovi, E. Castelli, A. Hirschberg, A. P. J. Wijnands, and H. M. A. Bailliet, "Description of the flow through *in-vitro* models of the glottis during phonation. Application to voiced sound synthesis," *Acta Acustica*, vol. 82, pp. 358–361, 1996.

[47] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 83, no. 4, pp. 1536–1552, 1988.

[48] ——, "The human vocal cords: A mathematical model, part I," *Phonetica*, vol. 28, pp. 129–170, 1973.

[49] P. Birkholz and D. Jackèl, "Simulation of flow and acoustics in the vocal tract," in *Proc. CFA/DAGA '04*, Strasbourg, France, 2004, pp. 895–896.

Peter Birkholz received the M.S. degree and the Ph.D. degree from the Institute for Computer Science, University of Rostock, Rostock, Germany, in 2002 and 2005, respectively.

He has been working as Research Associate at the University of Rostock since 2005. His main topics of reseach include articulatory speech synthesis and speech inversions with a focus on vocal tract modeling. For his dissertation on articulatory speech synthesis, he was awarded the Joachim–Jungius prize 2006 by the University of Rostock and the Klaus–Tschira prize for "understandable science" in 2006.

Dietmar Jackèl received the Ph.D. degree in engineering from the Berlin University of Technology, Germany, in 1987.

He is a Professor of computer science at the University of Rostock, Rostosk, Germany, and holds the chair in interactive graphical systems. His research interests include physically based animation and speech processing. He was a Scientific Advisor at GMD FIRST.

Dr. Jackèl is a member of GI.

Bernd J. Kröger received the M.S. degree in physics from the Rheinische-Wilhelms-University of Münster, Münster, Germany, in 1985 and the Ph.D. degree and his postdoctoral lecture qualification ("Habilitation") in phonetics from the University of Cologne, Cologne, Germany, in 1989 and 1998.

Since 1992, he has been with the Department of Phonetics, University of Cologne, as an Assistant Professor, and since 2001 he has been with the Department of Phoniatrics, Pedaudiology, and Comunication Disorders, Technical University of Aachen, Aachen, Germany, as a Senior Researcher and Associate Professor. His research interests are in the field of general phonetics, articulatory speech synthesis, and neural network modeling.