Real-time numerical solution of Webster's equation on a non-uniform grid

K. van den Doel and U.M. Ascher

***Abstract*— We present a numerical scheme for the real-time solution of the discretized one dimensional linearized acoustics equation (Webster's equation) augmented with dissipative terms, in a tube with a spatially and temporally varying cross section. The resulting algorithm produces similar results as the Kelly-Lochbaum model but has several advantages over it: the tube length can change continuously, the area function can change in time and can be governed by a dynamical model itself, the scheme is of second order accuracy (which can be interpreted as using conical elements) and the spatial and temporal discretization does not need to be uniform. We show how the model can be used to build a realistic real-time audio synthesis method for articulatory speech synthesis by coupling it to a lip radiation model, a dynamical wall model, and a glottal excitation model.**

I. Introduction

A widely used physical model for sound propagation in ducts such as occurring in the vocal tract and in wind instruments is a one-dimensional tube described by an area function. Excitations are placed in the tube and their propagation is approximated by a digital ladder (or waveguide) filter. The celebrated Kelly-Lochbaum (KL) model [1] approximates the tube by a set of cylindrical elements with length $L_e = c/F_s$, where c is the velocity of sound and F_s is the audio sampling rate. The KL model leads to a very efficient implementation but has some limitations, most of which have been addressed in some form in the literature.

First, the length of the tube is quantized and cannot be changed smoothly. This prohibits the synthesis of vocal sounds which involve lip protrusion or lowering the larynx, or instruments such as a slide trumpet. Several solutions have been proposed to alleviate this difficulty. One approach [2] is to change the length by changing the audio sampling rate. This approach requires the use of sample rate converting filters as described in detail in [3]. Another solution [4], [5] is to replace at least one of the segments by an all-pass fractional length delay filter. As the tube changes in length, sections have to be added and removed. This approach has been refined and generalized in [6], [7]. A related approach [8] specific to speech is to attach variable length models of the larynx and lip sections to the KL model.

Second, a piecewise *linear* approximation for the mostly smoothly varying area function using conical elements seems more logical than the piecewise constant approximation with cylindrical elements. Higher order conical elements have been proposed in [6], [9].

Finally, the KL model does not allow a non-uniform spatial discretization. If the area function changes rapidly in some regions and slowly in others then a non-uniform grid is natural for cheap accurate simulation.

In this paper we present a new method for the real-time numerical simulation of wave propagation in a 1D acoustical tube that solves all these problems simultaneously. We perform a real-time numerical integration of the 1D acoustics equation in the tube (the linearized Navier-Stokes equation, closely related to Webster's equation), which is a partial differential equation (PDE) in time and space. A finite volume discretization of the PDE is employed, and the resulting system is numerically integrated in time using a generalization of the leap-frog method. Large damping can occur at constrictions [10], and this introduces stiffness into the system; however, an implicit treatment of the damping terms avoids potential stability problems. The resulting algorithm is functionally equivalent to a KL filter, yet does not suffer from the limitations mentioned above. In addition, we believe the approach to be closer to the physics of the tube and as such more easily extendable.

The computational cost is typically about 2 times higher than the KL method for comparable applications, but still quite small for applications such as speech synthesis or musical instrument modeling, yielding an efficient real-time audio synthesis algorithm.

Our approach is similar in spirit to the method proposed in [11], [12], though our numerical integration method is quite different.

The remainder of this paper is organized as follows. In Section II we discuss the 1D acoustics equations in a tube of varying cross section. In Section III we present a scheme for the discretization of the classical 1D wave equation in first order form and show that it is equivalent to the leap-frog scheme. In Section IV we generalize this discretization to the acoustics equation in a tube of varying cross-section. In Section V we use the model to construct a relatively complete model of the vocal tract for articulatory speech synthesis. In Section VI we present numerical results for applying the scheme to the modeling of vowels generated by specific vocal tract shapes. Conclusions are presented in Section VII.

II. The Acoustics Equation

Let us consider a tube of length L and area function $A(x,t)$ with $0 \le x \le L$ along the axis of the tube. We assume that the physical quantities of pressure $\hat{p}$, air density $\hat{\rho}$, and air velocity $\hat{u}$ depend only on x and t. We introduce the scaled variables $p(x,t)$ (dimensionless) and $u(x,t)$ (dimension of area) for pressure or density deviation and volume-velocity,

Department of Computer Science, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. E-mail: {kvdoel|ascher}@cs.ubc.ca.
Supported in part under NSERC Discovery Grant 84306.

by $p = \hat{\rho}/\rho_0 - 1$, $u = A\hat{u}/c$, where ρ_0 is the mass density of air and c is the speed of sound. The linear lossless acoustics of the tube is governed by the equation of continuity and the linearized Navier-Stokes equation, see for example [13], [12], [14]

$$(u/A)_t + cp_x \;=\; 0, \qquad (1a)$$

$$(Ap)_t + cu_x \;=\; -A_t, \qquad (1b)$$

where the subscripts t and x denote partial derivatives. We shall assume that A_t is small and hence products of A_t with u or p can be dropped (as the variables u and p are also considered small).

While Eq. 1 is on solid theoretical footing, realistic models must also account for losses. The latter are much harder to model within a one dimensional model as essentially higher dimensional phenomena such as boundary layer effects play an important role here. Since the losses in many applications are frequency dependent [14], we add a damping term at the right hand side of (1a) of the form

$$-d(A)u + D(A)u_{xx}.$$

For monochromatic waves of the form $e^{i\omega(t-x/c)}$ this results in a more flexible frequency dependent damping term of the form

$$-[d(A) + D(A)\omega^2/c^2]u. \qquad (2)$$

We are therefore led to consider the PDE

$$(u/A)_t + cp_x \;=\; -d(A)u + D(A)u_{xx}, \qquad (3a)$$

$$(Ap)_t + cu_x \;=\; -A_t. \qquad (3b)$$

Note that near constrictions in the tract the damping coefficients can become large, whereas in other areas they may be small. This may give the model different properties in different spatial regions. For ease of presentation we shall consider the boundary conditions (BC)

$$u(0,t) = u_g(t), \quad p(L,t) = 0, \qquad (3c)$$

where u_g is a prescribed volume velocity source. Coupling to a dynamical excitation model such as the two mass Flanagan-Ishizaka model [10] and a radiation model can be easily achieved by modifying these boundary conditions, as we show explicitly in Section V. Note that $A(x,t)$ could itself be described by a dynamical model, resulting in a dynamical wall model; see Section V.

The system (3) can be written as a second order PDE for a potential ϕ by writing

$$p = \frac{\phi_x}{A} - 1, \quad u = -\frac{1}{c}\phi_t, \qquad (4)$$

which leads to

$$\left(\frac{\phi_t}{A}\right)_t - c^2\left(\frac{\phi_x}{A}\right)_x = -d(A)\phi_t + D(A)\phi_{txx}. \qquad (5)$$

This reduces to Webster's equation if A is independent of time and if there is no damping. Note that the variable ϕ is not the usual acoustic potential; we introduce it here only to show the connection to Webster's equation.

III. The classical wave equation

Let us first consider the classical wave equation, i.e. $A = 1$ in (1). We now want to discretize (1) using a finite volume approach and short differences so that the well-known leap-frog scheme [15], [16] is obtained for (5) upon suitable elimination.

Consider a grid in space, $x_0 < x_1 < \ldots < x_J < x_{J+1}$, with $\Delta x_{j+1/2} = x_{j+1} - x_j$, and likewise a grid in time $0 = t_0 < t_1 < \ldots$, with $\Delta t_{n+1/2} = t_{n+1} - t_n$. Let also $x_{j+1/2} = (x_j + x_{j+1})/2$, $\Delta x_j = x_{j+1/2} - x_{j-1/2}$ and likewise for $t_{n+1/2}$ and Δt_n; see Fig. 1. We can integrate (1a) on a grid volume

$$\begin{aligned} 0 &= \frac{1}{\Delta t_n \Delta x_j}\int_{t_{n-1/2}}^{t_{n+1/2}}\int_{x_{j-1/2}}^{x_{j+1/2}} (u_t + cp_x)\,dtdx \\ &= \frac{u_j^{n+1/2} - u_j^{n-1/2}}{\Delta t_n} + c\,\frac{p_{j+1/2}^n - p_{j-1/2}^n}{\Delta x_j}. \qquad (6a)\end{aligned}$$

This formula is exact by the Gauss divergence theorem if we interpret $u_j^{n+1/2}$ etc. as line integrals, e.g.

$$u_j^{n+1/2} = \frac{1}{\Delta x_j}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_{n+1/2})dx,$$

which is a second order approximation for the value at edge midpoint $(x_j, t_{n+1/2})$.

Likewise, we integrate (1b) on *another* nearby grid volume

$$\begin{aligned} 0 &= \frac{1}{\Delta t_{n+1/2}\Delta x_{j+1/2}}\int_{t_n}^{t_{n+1}}\int_{x_j}^{x_{j+1}} (p_t + cu_x)\,dtdx \\ &= \frac{p_{j+1/2}^{n+1} - p_{j+1/2}^n}{\Delta t_{n+1/2}} + c\,\frac{u_{j+1}^{n+1/2} - u_j^{n+1/2}}{\Delta x_{j+1/2}}. \qquad (6b)\end{aligned}$$

Again, 2nd order accurate point-wise values can be obtained upon interpretation of these quantities as midpoint values at the edges of the control volume. From (6) it is clear that this is a staggered scheme, as the volume velocity variables u live on the grid nodes in space and in-between the grid nodes in time, whereas the pressure variables p live in-between the spatial grid nodes and on the temporal grid nodes.

Now, using the centered approximations of (4)

$$u_j^{n+1/2} = -\frac{\phi_j^{n+1} - \phi_j^n}{c\Delta t_{n+1/2}}, \qquad p_{j+1/2}^n = \frac{\phi_{j+1}^n - \phi_j^n}{\Delta x_{j+1/2}} - 1, \quad (7)$$

we can substitute (7) into the equations (6) and obtain the leap-frog scheme approximating (5) without the damping term for ϕ alone. On a uniform grid this leap-frog scheme is the usual

$$\frac{\phi_j^{n+1} - 2\phi_j^n + \phi_j^{n-1}}{\Delta t^2} = c^2\frac{\phi_{j+1}^n - 2\phi_j^n + \phi_{j-1}^n}{\Delta x^2}.$$

The scheme (6) inherits various conservation properties from the differential system, including symplecticity and multisymplecticity on a uniform grid; see for example [17], [18].

IV. Discretization of the acoustic equation

A. The basic discretization

We now develop a discretization for (3) using the same notation. The variables u and p continue to live on separate

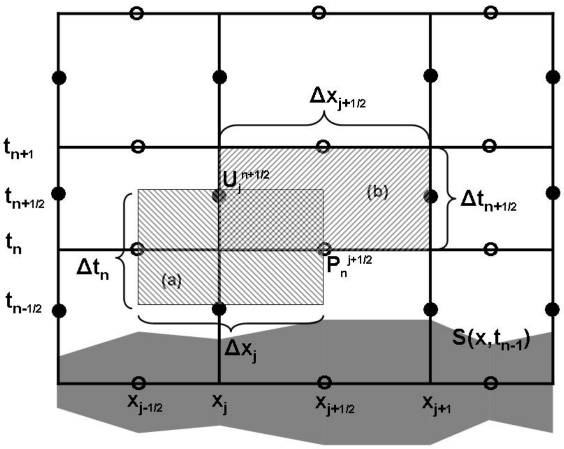

Fig. 1. The space-time domain is discretized with a non-uniform staggered grid. The area function $A(x,t)$ (drawn for time t_{n-1}) is known at all grid points and our finite volume scheme approximates it by a piecewise linear function between grid points. The pressure variables (open circles) and velocity variables (closed circles) live on separate space-time grid locations. The acoustics equations (3a) and (3b) are integrated on the indicated cells labeled (a) and (b), resp.

space-time grid nodes as indicated in Fig. 1 and are considered constant in the grid cells around the nodes. The area function A on the other hand is considered given externally and is assumed to be defined everywhere.

Integrating (3a) as before over the cell indicated by (a) in Fig. 1 yields the expression

$$\begin{aligned}\frac{u_j^{n+1/2}/\tilde{A}_j^{n+1/2}-u_j^{n-1/2}/\tilde{A}_j^{n-1/2}}{\Delta t_n}+c\,\frac{p_{j+1/2}^n-p_{j-1/2}^n}{\Delta x_j}\\ =\frac{1}{\Delta t_n\Delta x_j}\int_{t_{n-1/2}}^{t_{n+1/2}}\int_{x_{j-1/2}}^{x_{j+1/2}}[-d(A)u(x,t)\\ +D(A)u_{xx}(x,t)\ ]dxdt \qquad (8)\end{aligned}$$

with

$$\frac{1}{\tilde{A}_j^{n+1/2}}=\frac{1}{\Delta x_j}\int_{x_{j-1/2}}^{x_{j+1/2}}\frac{dx}{A(x,t_{n+1/2})}. \qquad (9)$$

To discretize (9) we approximate A as piecewise linear in space to obtain

$$\frac{1}{\tilde{A}_j^{n+1/2}}\approx\left(\frac{1}{A_{j-1/2}^{n+1/2}}+\frac{2}{A_j^{n+1/2}}+\frac{1}{A_{j+1/2}^{n+1/2}}\right)/4. \qquad (10)$$

We can interpret this as using conical segments in our model. However, since we are using a more abstract discretization, this is not to be equated with the conical segments used in [6], [9].

Higher accuracy could be obtained if so desired, for example if A has discontinuities.

In order to discretize the volume integral of $-d(A)u$ in (8) we consider u at $t_{n+1/2}$ rather than t_n, thus obtaining an implicit scheme that remains stable even for very large values of d without requiring very small time steps. Large values of d can occur for example near constrictions. We then integrate over x, giving $-\widehat{d}_j^{n+1/2}u_j^{n+1/2}$, where

$$\widehat{d}_j^{n+1/2}:=\left(d(A_{j-1/2}^{n+1/2})+2d(A_j^{n+1/2})+d(A_{j+1/2}^{n+1/2})\right)/4. \qquad (11)$$

For the term $D(A)u_{xx}$ in (8) we first discretize $u_{xx}\approx {u''}_j^{n+1/2}$ with

$${u''}_j^{n+1/2}:=\frac{u_{j+1}^{n+1/2}-u_j^{n+1/2}}{\Delta x_j\Delta x_{j+1/2}}-\frac{u_j^{n+1/2}-u_{j-1}^{n+1/2}}{\Delta x_j\Delta x_{j-1/2}},$$

for $j=1,\dots,J-1$. (For $j=J$ we define for convenience ${u''}_J=0$.) Then we proceed as for the term $d(A)u$, leading to $\widehat{D}_j^{n+1/2}{u''}_j^{n+1/2}$, where we have used a similar notation for $\widehat{D}$ as in (11). We finally obtain

$$\begin{aligned}\frac{u_j^{n+1/2}/\tilde{A}_j^{n+1/2}-u_j^{n-1/2}/\tilde{A}_j^{n-1/2}}{\Delta t_n}+c\,\frac{p_{j+1/2}^n-p_{j-1/2}^n}{\Delta x_j}+\\ \widehat{d}_j^{n+1/2}u_j^{n+1/2}-\widehat{D}_j^{n+1/2}{u''}_j^{n+1/2}=0. \qquad (12a)\end{aligned}$$

Likewise, integrating (3b) over the cell labeled (b) in Fig. 1 yields

$$\begin{aligned}\frac{\hat{A}_{j+1/2}^{n+1}p_{j+1/2}^{n+1}-\hat{A}_{j+1/2}^n p_{j+1/2}^n}{\Delta t_{n+1/2}}+c\,\frac{u_{j+1}^{n+1/2}-u_j^{n+1/2}}{\Delta x_{j+1/2}}=\\ -\frac{\hat{A}_{j+1/2}^{n+1}-\hat{A}_{j+1/2}^n}{\Delta t_{n+1/2}}, \qquad (12b)\end{aligned}$$

with

$$\widehat{A}_{j+1/2}^n:=\left(A_j^n+2A_{j+1/2}^n+A_{j+1}^n\right)/4,$$

where we have again approximated A as piecewise linear, as in (10) To correspond to the BC (3c) we set $x_0=0,\ x_{J+1/2}=L$. At a given time t_n, $n\geq 0$, assume we already know $\{p_{j+1/2}^n\}_{j=0}^J$ and $\{u_j^{n-1/2}\}_{j=0}^J$. For $n=0$ we obtain this from the initial data. One time step involves the following:

1) Use (12a) to obtain $\{u_j^{n+1/2}\}_{j=1}^J$. This involves the inversion of just a tridiagonal matrix which can be done efficiently using the Thomas algorithm at about three times the cost of solving a diagonal system. If $D=0$ no matrix inversion is needed at all. Note that the tridiagonal system involves only the u nodes.
2) Set $u_0^{n+1/2}=u_g(t_{n+1/2})$ by the left BC.
3) Use (12b) to obtain $\{p_{j+1/2}^{n+1}\}_{j=0}^{J-1}$. This can be done without any matrix algebra.
4) Set $p_{J+1/2}^{n+1}=0$ by the right BC.

For stability, the Courant-Friedrich-Levy (CFL) condition [19], [15] requires that Δt_n be small enough so that information cannot propagate beyond a finite interval of size Δx_j around a grid line x_j. Defining the CFL number

$$\eta=\max_{i,n}\frac{c\Delta t_n}{\Delta x_i},$$

a heuristic condition for stability is $\eta<1$. In practice we will often take Δt_n to be constant, and the CFL condition then becomes a constraint on the smallest grid element Δx_j in the system. In more general situations (see Section VI) we find that stability depends also on the area function $A(x,t)$, and instability may occur sometimes already for η close to but still below 1. The damping term does not make matters worse here because it is discretized implicitly.

Though (12) is strictly correct only for stationary grids, we can dynamically change the length $L(t)$ of the tract by changing Δx_J in time, as long as we do it slowly.

To compare the computational cost of this algorithm with that of the KL model in situations where the KL model can also be used, i.e. on a uniform grid, with a time independent area function A and without any damping, we write our scheme for this special case:

$$u_j^{n+1/2} = u_j^{n-1/2} - \eta \tilde{A}_j (p_{j+1/2}^n - p_{j-1/2}^n), \tag{13a}$$

$$p_{j+1/2}^{n+1} = p_{j+1/2}^n - \eta (u_{j+1}^{n+1/2} - u_j^{n+1/2}) / \hat{A}_{j+1/2}. \tag{13b}$$

After precomputing $\eta \tilde{A}_j$ and $\eta / \hat{A}_{j+1/2}$ this amounts to 2 multiplications and 4 addition/subtractions per grid point, to be compared to 1 multiplication and 3 addition/subtractions for the Kelly-Lochbaum model.

V. Applications to Speech Synthesis

We have applied the methods described thus far to acoustics simulation for articulatory speech synthesis. We use the ArtiSynth [20], [21] simulation environment, which allows real-time interaction, visualization and audio synthesis. In this application the area function $A(x,t)$ is extracted dynamically from the simulation of the motions of 3D tissues, including a finite element model of the tongue, so that $A(x,t)$ (and the total length L of the tract) is updated at every time step of the tissue simulation. The details are described in [22]. For the time domain simulation of the acoustics (which runs in a separate thread from the simulation of the deformable tissues) we use a uniform spatial grid and a temporal sampling rate of 44100Hz corresponding to a step size of $\Delta t = 1/44100$s. Two configurations are employed, one with $N = 38$ finite half-volumes and one with $N = 18$, corresponding to $J = 19$ and $J = 9$ resp. The spatial grid size (which spans two finite volumes) is $\Delta x = 2L/N$ with L the length of the tube, assumed to vary slowly. We use $c = 350$m/s and consider several vocal tract shapes with $L \geq 16.5$cm. At the smallest tract length we have $\eta_{38} = 0.91$ and $\eta_{18} = 0.44$. For realistic modeling of the production of vowel sounds we have coupled the wave equation described above to a realistic lip radiation model, a dynamical wall vibration model, and a glottal excitation which we modeled using either a dynamical coupling to the Ishizaka-Flanagan two-mass model [10] or the simpler Rosenberg model [23].

A. Radiation Model

Following Flanagan [24], we model the radiation at the lip by a radiation impedance. This can be done by introducing a new time-dependent variable $w(t)$ and modifying the right boundary condition (3c) to

$$u(L,t) = \frac{\rho c}{R_R} p(L,t) + w(t), \tag{14a}$$

where w satisfies

$$\frac{dw}{dt} = \frac{\rho c}{L_R} p(L,t), \tag{14b}$$

with

$$L_R = \frac{8\rho}{3\pi\sqrt{\pi A(L,t)}}, \quad R_R = \frac{128\rho}{9\pi^2 A(L,t)}.$$

With some straightforward algebra this can be seen to be equivalent to the continuum limit of the time-domain radiation impedance described in [11]. When discretizing we terminate the grid at u_J and update by

$$w^{n+1/2} = w^{n-1/2} + \Delta t \frac{\rho c}{L_R} p_{J-1/2}^n, \tag{15a}$$

$$u_J^{n+1/2} = \frac{\rho c}{R_R} p_{J-1/2}^n + w^{n+1/2}. \tag{15b}$$

Note that the order is important, imposing the BC at the new time level. If we update u_J first instead using w at a previous time (corresponding to an explicit method) then we find computationally that the system becomes frequently unstable.

Finally, we render the time derivative of u_J (corresponding to the radiated pressure) in real-time using the JASS [25] audio synthesis system.

B. Wall Vibration Model

A dynamical wall model is obtained by substituting in (3)

$$A(x,t) \to A(x,t) + C(x,t) y(x,t),$$

where $A(x,t)$ on the right hand side is considered to be slowly varying, $C(x,t)$ is the associated (slowly varying) circumference, and $y(x,t)$ is a small outward displacement of the wall. We then linearize (i.e., drop products of y with u or p) and obtain (3b) with the right hand side replaced by $-\partial(Cy)/\partial t$. The right hand side of (12b) is then modified by replacing $A \to Cy$. The wall displacement y is modeled as a damped mass-spring system (independently at every location x, so we are neglecting wall bending resistance), with the driving force determined by the pressure. This results in the ODE's in time

$$M\ddot{y}(x,t) + B\dot{y}(x,t) + Ky(x,t) = p(x,t), \tag{16}$$

with

$$M = M_0/\rho c^2, \quad B = B_0/\rho c^2, \quad K = K_0/\rho c^2,$$

and we have used the same values as in [11]

$$M_0 = 21\text{kg/m}^2, \quad B_0 = 8000\text{kg/m}^2, \quad K_0 = 845000\text{kg/m}^2\text{s}^2.$$

To discretize (16), note that (12b) requires the values of y on the temporal grid nodes, both on and in-between the spatial grid nodes. However, since the driving force (16) on y depends on the pressure variables, and the latter live only on the in-between spatial nodes, those variables y_j that live on the spatial nodes are not independent and can be set to $y_j = (y_{j-1/2} + y_{j+1/2})/2$. Inspecting (3b) we see that the right hand side becomes simply

$$-\frac{(Cy)_{j+1/2}^{n+1} - (Cy)_{j+1/2}^n}{\Delta t}. \tag{17}$$

The discrete time stepping for (16) is performed after updating the pressures, using the auxiliary variable $z = \dot{y}$:

$$\begin{aligned} y^{n+1}_{j+1/2} &= y^n_{j+1/2} + z^n_{j+1/2}, \\ z^{n+1}_{j+1/2} &= [M z^n_{j+1/2} + \Delta t(p^{n+1}_{j+1/2} - K y^{n+1}_{j+1/2})]/(M + \Delta t B) \end{aligned} \tag{18}$$

C. Modeling Losses

Assuming a hard walled, circular tube, the losses at the wall can be modeled by a frequency dependent damping coefficient [10]

$$d(\omega) = \sqrt{\frac{2\pi\mu\omega}{\rho}}/A^{3/2}, \tag{19}$$

where the coefficient of viscosity is $\mu = 1.86^{-5}\mathrm{kg/ms}$ and the air density is $\rho = 1.14\mathrm{kg/m}^3$. In [26] an empirically constructed linear function of f was used instead, with the same dependence on A. We shall also assume this area dependence and write for the two damping coefficients

$$d = \tilde{d}A^{-3/2}, \; D = \tilde{D}A^{-3/2}. \tag{20}$$

This leads to a frequency dependent damping given by (2). We determine the coefficients $\tilde{d}$ and $\tilde{D}$ by first matching (20) to (2) at two reference frequencies of 250Hz and 2000Hz, with the intention of matching them approximately over a speech-relevant frequency range, resulting in

$$\tilde{d} = 1.6\mathrm{m/s} \text{ and } \tilde{D} = 0.002\mathrm{m}^3/\mathrm{s}.$$

In Fig. 2 we show the two damping models. In practice we have found a need to increase the damping to reproduce realistic formant bandwidths, as discussed in Section VI.

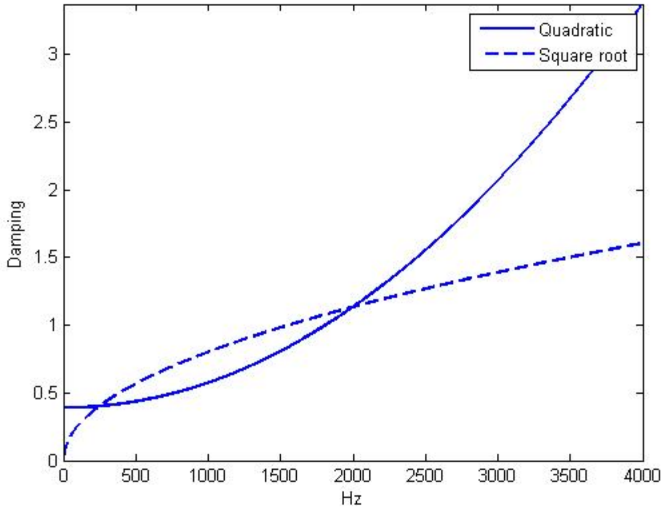

Fig. 2. Frequency dependent damping using the $\sqrt{\omega}$ proportional hard wall model, and the quadratic damping function resulting from our model.

D. Glottal Model

The glottal excitation u_g was generated parametrically according to the Rosenberg model [23], or by a dynamic coupling to the Ishizaka-Flanagan two-mass model [10]. This model computes pressure oscillations as well as glottal flow and is dynamically driven by lung pressure and tension parameters in the vocal chords. The vocal chord model is coupled to the discretized acoustics equation in the vocal tract. We follow the implementation described in [27] using the same Δt as for the propagation modeling. At every time step we provide the glottal model with the pressure derived from $p_{1/2}$, compute the volume velocity u_g, and then update the tube variables u_j and $p_{j+1/2}$.

VI. Results

A. Formant Accuracy

To test the accuracy of the method, we have computed the formants (resonance frequencies) for six area functions A and compared this with an almost exact solution in the frequency domain on a very fine spatial mesh. For these calculations we set the damping coefficients d and D to zero, as their effect on the formant frequencies is negligible (less than 0.1%) for the values resulting from the damping model described in Section V-C. All other parameters are set to the values described in Sections V-A and V-B.

In the time domain simulation we calculate the formant frequencies by calculating $u(J)$ for 2^{15} time samples or 0.743s for an impulse u_g at time $t = 0$, taking the Fourier transform, and finding the maxima in the power spectrum using quadratic interpolation.

To obtain a solution in the frequency domain, we consider (12) with the right hand side replaced by (17). We set $d = D = 0$, use the boundary condition (14) for w, and use (18) for y. We then take the limit $\Delta t \to 0$ and write the resulting ODE system in terms of the Fourier transforms of the spatial grid variables. These are denoted by capitalization, e.g. the Fourier transform of u is $U(\omega)$, with ω the radial frequency. Equation (14b) allows us to elimininate $W = P\rho c/i\omega L_R$ from (14a) and Y can be written in terms of P as $Y_{j+1/2} = P_{j+1/2}/(K + iB\omega - M\omega^2)$. With these substitutions (12) and the BC can be written as

$$\begin{aligned} \imath\omega U_j &= -\frac{c}{\Delta x}\tilde{A}_j(P_{j+1/2} - P_{j-1/2}), \; j = 1, \ldots, J-1 \\ \imath\omega P_{j-1/2} &= -\frac{c}{\Delta x}Q_{j-1/2}\hat{A}^{-1}_{j-1/2}(U_j - U_{j-1}), \; j = 1, \ldots, J \\ u_J &= (\frac{1}{R_R} + \frac{1}{i\omega L_R})P_{J-1/2}, \end{aligned}$$

where

$$Q_{j-1/2} = (1 + \frac{C_{j-1/2}\hat{A}^{-1}_{j-1/2}}{K + i\omega B - M\omega^2})^{-1}.$$

Using the vector notation $X_j = (P_{j-1/2}, U_j)^T$, and using the boundary conditions, we can rewrite this as

$$X_{j+1} = K_j X_j \quad \text{for} \quad j = 1, \ldots, J-1, \tag{21a}$$

$$U_J = a_J P_{J-1/2}, \quad P_{1/2} = -b_{1/2}(U_1 - U_g), \tag{21b}$$

with $a_j = c\tilde{A}_j / i\Delta x\omega$, $a_J = (\frac{1}{R_R} + \frac{1}{i\omega L_R})$, $b_{j+1/2} = cQ_{j+1/2}\hat{A}^{-1}_{j+1/2}/i\Delta x\omega$, and

$$K_j = \begin{pmatrix} a_j & 0 \\ 1 & b_{j+1/2} \end{pmatrix}^{-1} \begin{pmatrix} a_j & -1 \\ 0 & b_{j+1/2} \end{pmatrix}.$$

We can solve (21a) in the form $X_J = KX_1$, with $K = K_{J-1} \cdots K_1$, and this together with the conditions (21b) allow us to solve for U_J in terms of U_g, i.e. $U_J = H(\omega)U_g$. We then extract the formant frequencies from the maxima of the transfer function $||H(\omega)||$.

This scheme was implemented using a very fine spatial grid with $J = 499$, and $J = 999$. The difference in formant frequencies between the two J values was less than 0.1%, which is why we refer to this as the "exact solution". Six vowel tract shapes for Russian vowels taken from [26] were used as area functions, depicted in Fig. 3. The time-domain and frequency domain results for the first three formant frequencies are listed in Table I. The relative error is largest for the third formant and smallest for the first, and it depends on the area function. For the first formant there is little difference between the two grids, but for the second formant the $N = 38$ results are twice as accurate as the $N = 18$ results and the difference is a factor four for the third formant.

Fig. 4 displays these data graphically. For comparison we have also indicated the measured values of these formants and the simulated values reported by Fant [26]. Our results for $N = 38$ are very close to the measured values, except for the third formant of [e] which is somewhat too high. For $N = 18$ the third formant is too low in all cases except for the [e]. However, informal listening tests indicate that this is not an audible discrepancy.

	F_1 (1.7%)			F_2 (2.3%)			F_3 (2.4%)		
	FD	N=38	%	FD	N=38	%	FD	N=38	%
u	294	296	0.7	630	627	0.5	2392	2350	1.8
o	568	554	2.4	944	913	3.2	2426	2398	1.1
a	716	690	3.6	1184	1122	5.2	2558	2499	2.3
e	458	446	2.6	2038	1991	2.3	2998	2906	3.1
i	272	271	0.4	2324	2291	1.4	3298	3144	4.7
ɨ	328	326	0.6	1762	1736	1.5	2390	2356	1.4
	F_1 (2%)			F_2 (4%)			F_3 (8%)		
	FD	N=18	%	FD	N=18	%	FD	N=18	%
u	294	296	0.7	630	630	0	2392	2151	10
o	568	558	1.8	944	952	0.8	2426	2286	5.8
a	716	692	3.3	1184	1160	2.0	2558	2398	6.3
e	458	450	1.8	2038	1924	5.6	2998	2748	8.3
i	272	278	2.2	2324	2156	7.2	3298	2935	11
ɨ	328	318	3.1	1762	1614	8.4	2390	2185	8.6

TABLE I

The formant frequencies computed from the tract shapes for the six Russian vowels reported by Fant. The highly accurate frequency domain results (FD) are compared to the time domain results for $N = 28$ finite volumes and for $N = 18$. We also indicated the percentage errors for the individual shapes and the average error.

B. Formant Bandwidth

As described in Section V-C, we have first tuned the two damping parameters $\tilde{d}$ and $\tilde{D}$ to approximately reproduce the hard-wall loss formula (19) over a relevant frequency range. We have then manually scaled both these parameters to bring them close to the formant bandwidths reported in [26] which, though obtained by simulation, were claimed to be within 50% of actual values. We found a reasonable scaling factor to be 4 for $N = 38$ and 8 for $N = 18$. The spectra of the vowels are displayed in Fig. 5. They were computed using the time domain simulation on the $N = 38$ grid as described in Section VI-A. The 3dB bandwidth was then computed. In Table II we list the bandwidths for the first four formants for these values of the damping parameters. As can be seen, they are mostly in reasonable agreement with the values reported by [26], with the exception of the first formant for [u] and the fourth formant for [e], which are too wide for $N = 38$.

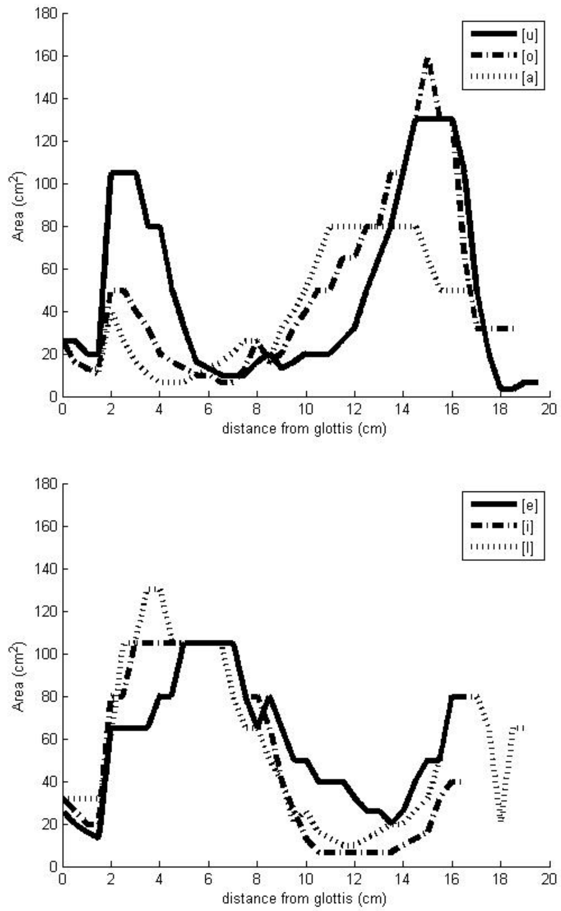

Fig. 3. The area functions for the six Fant vowels.

C. Tract Motion

To test the synthesis method we have created a standalone simulator, which produces audio in real-time and allows the user to dynamically change the parameters of the model. These include the area function and the tract length, and are changed through sliders while the sound is being produced. The simulator is made available online at [28], in the configurations $N = 18$ and $N = 38$. Three control windows are displayed, for the tract model, the Rosenberg glottal model

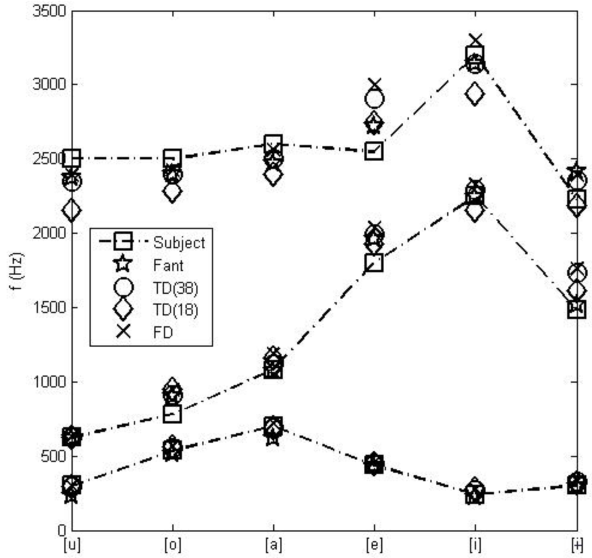

Fig. 4. The formants of the Russian vowels [u] [o] [a] [e] [i] [ɨ] as computed by our time-domain (using grids with $N = 18$ and $N = 38$) and frequency domain methods. For comparison we show the measured values and the simulation results from Fant [26].

	F_1			F_2		
	Fant	Default	Scaled	Fant	Default	Scaled
u	69	63/22	196/61	50	28/12	92/74
o	54	21/15	50/48	65	31/32	74/88
a	57	21/17	46/53	72	40/45	79/110
e	39	15/14	32/49	95	44/34	62/74
i	60	25/26	46/89	75	11/10	22/43
ɨ	43	18/16	37/42	125	183/89	958/143

	F_3			F_4		
	Fant	Default	Scaled	Fant	Default	Scaled
u	110	13/12	45/86	115	15/15	52/109
o	100	33/37	63/101	135	36/33	15/136
a	101	70/70	111/158	175	182/168	248/600
e	170	190/86	139/162	325	1231/168	1291/306
i	240	221/89	330/301	230	107/89	169/287
ɨ	77	44/31	70/82	134	84/39	132/147

TABLE II

The 3dB bandwidths of the first four formants are displayed. We show the values reported by Fant, the values obtained with our default parameter setting obtained by matching the hard wall form (19), and the values after scaling damping parameters with a factor $4/8$. Results (separated by "/") are for the $N = 38$ grid and the $N = 18$ grid.

and the two-mass glottal model. The sliders marked *u_xx mult* and *u_mult* are used to further scale the damping parameters $\tilde{d}$ and $\tilde{D}$, which are set to the values determined in Section VI-B. The slider labeled *wall coeff* allows control over the wall coupling by scaling the right hand side of (16). By moving the *length* slider smooth sound changes can be observed, which is achieved by changing Δx in the simulation. Note that for the $N = 38$ version the [u] and [e] vowels become unstable if the length is decreased to below 16cm, even though the CFL number 0.94 is still below 1. For the $N = 18$ simulator, the maximum CFL number attainable (at the current audio sampling rate) is 0.46 and as expected no instabilities due to a length decrease are observed. It is of course possible to stabilize the $N = 38$ model by increasing the audio sampling rate.

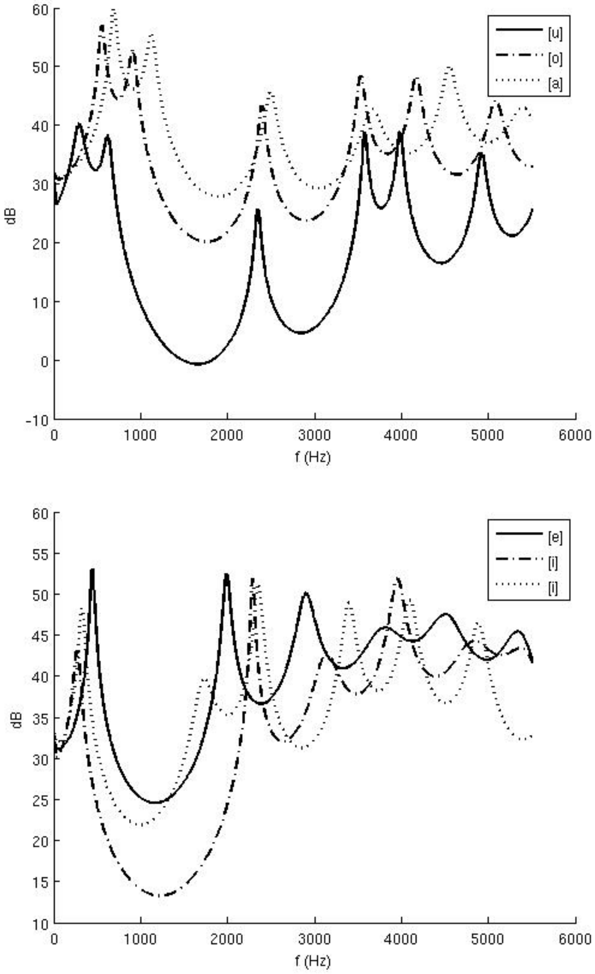

Fig. 5. Spectra (norm of the velocity-velocity transfer function) for the six Russian vowels. The default values for the damping parameters were used. Grid size was $N = 38$.

The remaining sliders represent 40 equidistant control points (not related to the spatial grid employed for the simulation) for the area function, with *A(39)* representing the mouth. The area functions for the six Russian vowels can be loaded by pressing the buttons, and the spectrum can be displayed by pressing the *Formants* button.

The two-mass glottal model is turned on by default and can be controlled through the three sliders which set the q-factor, the lung pressure, and the glottal rest area, see [10] for the meaning of these parameters. If the (nonlinear) two-mass model goes unstable, *NaN* is displayed. [1] Pressing the *Reset* button on the main control window resets the model (and also the automatic gain control). To turn off the two-mass model set the lung pressure to 0. To activate the Rosenberg model increase the gain with the corresponding slider.

We observe little difference in quality between the two grids

[1] This happens because the non-linear part of the spring model is treated by an explicit method.

if the two-mass model is used, but when using the Rosenberg model the $N = 18$ simulation clearly sounds better. We believe the reason lies in the excess high frequency energy present in the Rosenberg excitation. In Fig. 6 we display the spectral response of the [a] vowel over the entire frequency domain. It can be seen that the $N = 18$ grid eliminates frequencies above about $6000Hz$ whereas the cutoff for $N = 38$ lies at about 12000Hz. As the one-dimensional tube model (and therefore the PDE model considered here) is not valid for frequencies above roughly 6000Hz, their inclusion on the finer grid therefore results in an unnatural sound. If more accurate higher formants are required it is also possible to increase the audio sampling rate

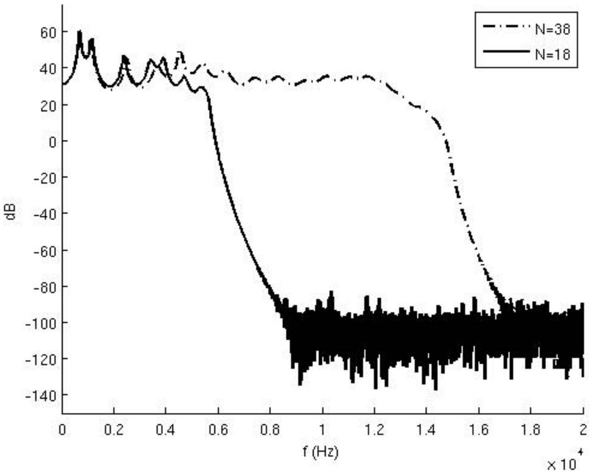

Fig. 6. The spectrum of the vowel [a] for $N = 18$ finite volumes and for $N = 38$. Though the $N = 18$ formants are less accurate, the falloff around 6000Hz results in a more natural sound if the glottal excitation contains energy above this frequency.

VII. Conclusions

We have described a computational scheme for the numerical integration of the linear acoustics equation in a tube with varying cross section. The resulting algorithm is an alternative to the classical Kelly-Lochbaum method and offers several advantages over it: 1) the tract length is not quantized and real-time continuous length changes are allowed, 2) non-uniform spatial grids are allowed, 3) the scheme is of second order accuracy, which can be visualized as using conical sections rather than cylindrical sections, 4) the area function is not required to be quasi-stationary and can move dynamically. The computational cost of the proposed algorithm is somewhat higher (by about a factor 2) than that of the KL model but is still very small in absolute terms on present-day computers. We proposed a two-term damping mechanism, capable of producing frequency dependent damping.

The scheme has been embedded into a relatively complete vocal tract simulation, suitable for synthesizing speech by coupling the model to a dynamical wall model, a radiation model, and an excitation model, and by a specific choice for the damping parameters. Simulations on a coarse ($N = 18$ finite volumes) and a fine ($N = 38$) spatial grid yields formants that are very close to their exact value as computed on a very fine grid using a frequency collocation method.

Area functions of six Russian vowels were used and the results for the formants and damping resulting from our simulation are close to the experimental and simulated values reported in [26]. A real-time simulator was implemented, allowing real-time articulatory control through sliders. In particular, the length of the tract can be changed continuously with smooth changes in the sound.

It was found that the coarse grid, though less accurate in predicting second and third formants, may occasionally be preferred over the dense grid. Using coupling to the two-mass Flanagan-Ishizaka glottal model there is no audible difference between the two grids, whereas using a Rosenberg parametric excitation the coarse grid eliminates high frequencies better, resulting in a more natural sound. In addition the coarse grid has a lower CFL number and hence is more robustly stable (for example if the length is reduced) and it is twice as fast.

References

[1] K. L. Kelly and C. C. Lochbaum, "Speech Synthesis," in Proc. Fourth Int. Congr. Acoust., Paper G42, Copenhagen, 1962, pp. 1–4.

[2] H. Y. Wu, P. Badin, and Y. M. Cheng, "Continuous variation of the vocal tract length in a Kelly-Lochbaum type speech production model," in Proc. Int. Congr. Phonetic Sciences (XIth ICPhS), Tallin, Estonia, 1987, pp. 340–243.

[3] G. T. H. Wright and F. J. Owens, "An optimized multirate sampling technique for the dynamic variation of the vocal tract length in the Kelly-Lochbaum speech synthesis model," IEEE Trans. Speech and Audio Processing, vol. 1, no. 1, pp. 109–113, 1993.

[4] H. W. Strube, "Sampled-data representation of a non-uniformlossless tube of continuously variable length," JASA, vol. 57, no. 1, pp. 256–257, 1975.

[5] U. K. Laine, "Digital modeling of a variable-length acoustic tube," in Proc. Nordic Acoustical Meeting, Tampere, Finland, June 15-17, 1988, pp. 163–168.

[6] V. Välimäki and M. Karjalainen, "Improving the Kelly-Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques," in Proc. Int. Conf. on Spoken Language Processing (ICSLP), Yokohama, Japan, Spet. 18-22, 1994.

[7] S. Mathur, Variable-length vocal tract modeling for speech synthesis, Ph.D. thesis, Dept. of Electrical and Computer Engineering, The University of Arizona, 2003.

[8] C. C. Goodyear, "Incorporating lip protrusion and larynx lowering into a time domain model for articulatory speech synthesis," Computer Speech and Language, vol. 14, no. 3, pp. 211–226, 2000.

[9] P.J.B. Jackson, Characterisation of plosive, fricative and aspiration components in speech production, Ph.D. thesis, Department of Electronics and Computer Science, University of Southampton, , Southampton, UK, 2000.

[10] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Sys. Tech. J., vol. 51, pp. 1233–1268, 1972.

[11] P. Birkholz and D. Jackel, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in INTERSPEECH-2004,8th International Conference on Spoken Language Processing, Korea, 2004, pp. 1125–1128.

[12] P. Birkholz, 3D-Artikulatorische Sprachsynthese, Ph.D. thesis, Universität Rostock, 2005.

[13] S. Temkin, Elements of Acoustics, John Wiley and Sons, Inc., New York, 1981.

[14] H. W. Strube, "The meaning of the Kelly-Lochbaum acoustic-tube model," JASA, vol. 108, no. 4, pp. 1850–1855, 2000.

[15] R. J. Leveque, Finite Volume Methods for Hyperbolic Problems, Cambridge University Press, 2002.

[16] K. W. Morton and D.F. Mayers, Numerical Solution of Partial Differential Equations, Cambridge University Press, 2005, 2nd ed.

[17] U. M. Ascher and R. I. McLachlan, "On symplectic and multisymplectic schemes for the KdV equation," J. Scient. Computing, vol. 25, pp. 83–104, 2005.

[18] T. J. Bridges and S. Reich, "Numerical methods for Hamiltonian PDEs," J. Phys. A: Math. Gen., vol. 39, pp. 5287–5320, 2006.

[19] R. Courant, K.O. Friedrichs, and H. Lewy, "Über die partiellen differenzengleichungen der mathematischen physik," Mathematische Annalen, vol. 100, pp. 32–74, 1928.

[20] S. Fels, F. Vogt, K. van den Doel, J. E. Lloyd, and O. Guenther, "Artisynth: An extensible, cross-platform 3D articulatory speech synthesizer," in Proceedings of Audio Visual Speech Processing (AVSP), Tigh-Na-Mara, Canada, 2005.

[21] S. Fels, F. Vogt, K. van den Doel, J. Lloyd, I. Stavness, and E. Vatikiotis-Bateson, "Artisynth: A biomechanical simulation platform for the vocal tract and upper airway," in International Seminar on Speech Production, Ubatuba, Brazil, 2006.

[22] K. van den Doel, F. Vogt, R. E. English, and S. Fels, "Towards articulatory speech synthesis with a dynamic 3D finite element tongue model," in International Seminar on Speech Production, Ubatuba, Brazil, 2006.

[23] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," JASA, vol. 49, pp. 583–590, 1971.

[24] J. L. Flanagan, Speech Analysis, Synthesis, and Perception, Academic Press, Inc., 1965.

[25] K. van den Doel and D. K. Pai, "JASS: A Java Audio Synthesis System for Programmers," in Proc ICAD, 2001.

[26] G. Fant, Speech sounds and features, Cambridge University Press, 1973.

[27] M. M. Sondhi and J. Schroeter, "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," IEEE Trans on ASSP, vol. 35, no. 7, pp. 955–967, 1987.

[28] Anon, "Online demo, http://members.shaw.ca/anonpaper," 2007.