Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system

Peter Birkholz and Dietmar Jackèl

Institute for Computer Sciences, University of Rostock
Albert-Einstein-Str. 21, 18055 Rostock, Germany (Europe)
email: {piet,dj}@informatik.uni-rostock.de

Abstract

A time domain simulation of acoustic propagation in the vocal tract requires the spatial and temporal discretization of the equations of motion and continuity. In the classic transmission line model of the vocal tract with lumped elements, the spatial discretization is provided by the piece-wise constant area function. The temporal finite-difference approximation of the differential equations can, however, vary from one implementation to the other (e.g., [4] vs. [5]). In this study, we have adopted a *general* finite-difference scheme that depends on a parameter θ where $0 \leq \theta \leq 1$. As special cases, this general method includes the trapezoid rule ($\theta = 0.5$) as well as the implicit ($\theta = 1$) and explicit ($\theta = 0$) finite-difference schemes. We have examined how formant frequencies and bandwidths of simulated vowels are effected by the choice of θ. The experiments were conducted for the sampling rates of 44.1 kHz and 88.2 kHz and compared with the accurate and thus desirable frequencies and bandwidths measured in frequency domain simulations of the vocal tract. It can be shown that optimal values for θ are slightly *above* 0.5 depending on the sampling rate.

1. Introduction

In articulatory speech synthesis, the vocal tract is frequently represented by an inhomogeneous transmission line with lumped elements [5, 4]. In this transmission line model (TLM), the vocal tract is approximated as a series of abutting cylindrical tube sections as illustrated in Fig. 1 (a). Each individual section is represented as a lumped element of the transmission line as depicted in Fig. 1 (b).

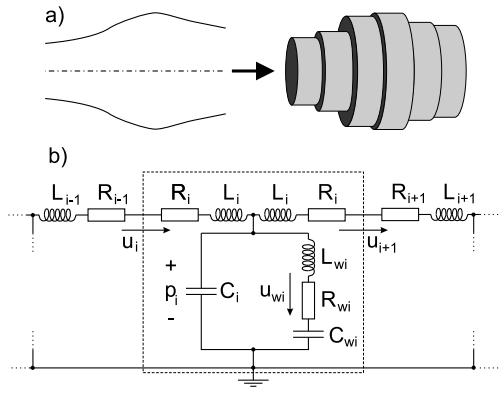

Figure 1: a) Discrete tube representation of a part of the vocal tract. b) Lumped transmission line network for one tube section (dashed box).

In this analogy, electrical current corresponds to volume velocity u and voltage corresponds to acoustic pressure p. The inductivities L_i and the capacities C_i represent respectively the inertance and the compliance of the air in the tube section i. The resistances R_i account for energy losses due to viscous friction and the series connection of $R_{\mathrm{w}i}$, $L_{\mathrm{w}i}$ and $C_{\mathrm{w}i}$ represents the vocal tract wall impedance. Vocal tract losses arising from heat conduction at the wall are not included in this network, because they are essentially negligible in the frequency region of interest [3]. The derivation of the transmission line element values can be found, for instance, in [3]. We summarize the values for the network components as follows:

$$\begin{array}{ll} L_i = \overline{\rho} l_i/(2A_i) & L_{\mathrm{w}i} = M_{\mathrm{w}}/(l_i S_i) \\ R_i = [S_i l_i/(2A_i^2)]\sqrt{\overline{\rho}\omega\mu/2} & R_{\mathrm{w}i} = B_{\mathrm{w}}/(l_i S_i) \\ C_i = A_i l_i/(\overline{\rho} c^2) & C_{\mathrm{w}i} = (l_i S_i)/K_{\mathrm{w}} \end{array} \quad (1)$$

S_i, A_i and l_i are the perimeter, area and length of the tube section i, respectively. ω is the radian frequency, $\overline{\rho}$ is the ambient density, c is the speed of sound and μ the coefficient of viscosity. The remaining parameters M_{w}, B_{w} and K_{w} are the mass, resistance and stiffness of the vocal tract walls per area, respectively. In our simulation, we have chosen the values that were measured for the relaxed cheek in [6], namely $M_{\mathrm{w}} = 21\ \mathrm{kg/m^2}$, $B_{\mathrm{w}} = 8000\ \mathrm{kg/m^2 s}$ and $K_{\mathrm{w}} = 845000\ \mathrm{kg/m^2 s^2}$.

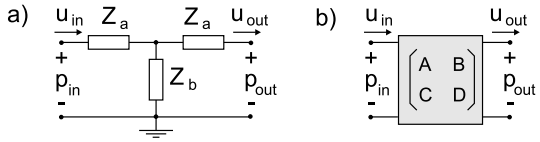

Figure 2: a) T-network and b) system representation of one individual tube section.

A simplified representation of a tube section is the T-network shown in Fig. 2 (a) or the corresponding two-port system shown in Fig. 2 (b). The entire vocal tract consists of the concatenated two-port systems of all tube sections. In our simulation, the entire network is connected to a current source at the glottis and is terminated by a radiation impedance at the mouth. According to Flanagan [3], the radiation impedance is well approximated by a parallel R-L circuit as depicted in in Fig. 3, where

$$L_{\mathrm{R}} = \frac{8\overline{\rho}}{3\pi\sqrt{\pi A_N}}, \qquad R_{\mathrm{R}} = \frac{128\overline{\rho}c}{9\pi^2 A_N}, \quad (2)$$

and A_N is the area of the last tube section (mouth area).

Depending on the requirements, the vocal tract network can be simulated in either the time domain or the frequency domain. A

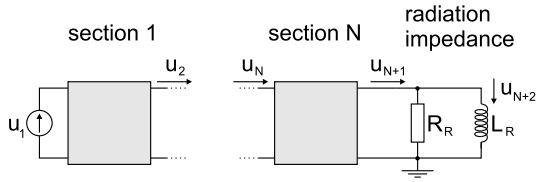

Figure 3: Entire network for the vocal tract consisting of N tube sections. Glottal excitation is represented by the volume velocity source u_1.

frequency domain simulation (FDS) is needed for the strict evaluation of the system in terms of formant frequencies and bandwidths, because this simulation method does not introduce errors (dispersion effects) due to a temporal discretization and can properly account for the frequency dependent resistance R_i in the transmission line. However, a time domain simulation (TDS) is desirable for speech synthesis, because it can account for the dynamic aspects of speech production in a natural way. The basis for the TDS is the temporal discretization of the differential equations that describe the behaviour of the vocal tract network. Therefore, the differential equations must be approximated by difference equations. Well-known methods for the finite-difference approximation are the simple implicit method, the simple explicit method and the trapezoid rule. All three methods are special cases of a *general* discretization scheme that depends on a parameter θ. In this study we examined the influence of different θ-values (and thus different discretization schemes) on the bandwidths and frequencies of vowel formants for two sampling rates. The values obtained in the TDS were compared to the desirable values from a FDS in order to find the value for θ that gives the best spectral match. The procedures for the TDS and the FDS are briefly described in the following two sections and the examined cases are explained in Sec. 4. We shall discuss the results in Sec. 5 and draw the conclusions in Sec. 6.

2. Time domain simulation

The relationships between pressures and volume velocities for a single tube section can be readily derived from Fig. 1 (b):

$$\dot{p}_i = \frac{1}{C_i}(u_i - u_{i+1} - u_{\mathrm{w}i}) \quad (3)$$

$$p_{i-1} - p_i = \dot{u}_i(L_{i-1} + L_i) + u_i(R_{i-1} + R_i) \quad (4)$$

$$p_i = L_{\mathrm{w}i}\dot{u}_{\mathrm{w}i} + R_{\mathrm{w}i}u_{\mathrm{w}i} + \frac{1}{C_{\mathrm{w}i}}\int u_{\mathrm{w}i}dt \quad (5)$$

Equation (5) describes the vibration of the vocal tract walls and the Eqs. (3) and (4) express the conservation of mass and momentum. For the volume velocities u_{N+1} and u_{N+2} through the radiation impedance we obtain (cf. Fig. 3)

$$p_N = u_{N+1}R_N + \dot{u}_{N+1}L_N + (u_{N+1} - u_{N+2})R_{\mathrm{R}}, \quad (6)$$

$$\dot{u}_{N+2}L_{\mathrm{R}} = (u_{N+1} - u_{N+2})R_{\mathrm{R}}. \quad (7)$$

With regard to the forthcoming discretization we take the time derivative of Eq. (5) so that the integral disappears from the right-hand side. When the slow variations of the component values $L_{\mathrm{w}i}$, $R_{\mathrm{w}i}$ and $C_{\mathrm{w}i}$ are neglected, we get

$$\dot{p}_i = L_{\mathrm{w}i}\ddot{u}_{\mathrm{w}i} + R_{\mathrm{w}i}\dot{u}_{\mathrm{w}i} + \frac{1}{C_{\mathrm{w}i}}u_{\mathrm{w}i}. \quad (8)$$

Our general approach for the finite-difference approximation reads

$$f[n] - f[n-1] = \theta\Delta t\dot{f}[n] + \overline{\theta}\Delta t\dot{f}[n-1], \quad (9)$$

where n is the sampling index, Δt is the time step, θ is a constant between 0 and 1, and $\overline{\theta} = 1 - \theta$. The function f stands for p or u. With regard to Eq. (9), the simple implicit finite-difference scheme corresponds to $\theta = 1$, the simple explicit scheme to $\theta = 0$ and the trapezoid rule to $\theta = 0.5$.

For a better readability we introduce the following abbreviations, where n is the *current* sampling index and $n-1$ is the *immediately elapsed* sampling index:

$$f[n] \equiv f \qquad f[n-1] \equiv f'$$
$$\dot{f}[n] \equiv \dot{f} \qquad \dot{f}[n-1] \equiv \dot{f}'$$

The first and second time derivative can then be written as

$$\dot{f} = \frac{1}{\Delta t\theta}(f - f') - \frac{\overline{\theta}}{\theta}\dot{f}', \quad (10)$$

$$\ddot{f} = \frac{1}{\Delta t^2\theta^2}(f - f') - \frac{1}{\Delta t\theta}\left(\frac{\overline{\theta}}{\theta} + 1\right)\dot{f}' - \frac{\overline{\theta}}{\theta}\ddot{f}'. \quad (11)$$

We now combine the equations of continuity (3) and wall vibration (8) to a single time-discrete equation. Therefore, we expand $\dot{u}_{\mathrm{w}i}$ and $\ddot{u}_{\mathrm{w}i}$ in (8) by means of (10) and (11) and then solve this equation for $u_{\mathrm{w}i}$. We obtain

$$u_{\mathrm{w}i} = \dot{p}_i\alpha_i + \beta_i, \quad (12)$$

where

$$\alpha_i = 1/\left(\frac{L_{\mathrm{w}i}}{\Delta t^2\theta^2} + \frac{R_{\mathrm{w}i}}{\Delta t\theta} + \frac{1}{C_{\mathrm{w}i}}\right)$$

$$\beta_i = \alpha_i\left[u'_{\mathrm{w}i}\left(\frac{L_{\mathrm{w}i}}{\Delta t^2\theta^2} + \frac{R_{\mathrm{w}i}}{\Delta t\theta}\right) + \dot{u}'_{\mathrm{w}i}\left(\frac{L_{\mathrm{w}i}}{\Delta t\theta}(\overline{\theta}/\theta + 1) + R_{\mathrm{w}i}\frac{\overline{\theta}}{\theta}\right) + L_{\mathrm{w}i}\frac{\overline{\theta}}{\theta}\ddot{u}'_{\mathrm{w}i}\right]$$

When we put Eq. (12) in (3), expand $\dot{p}_i$ and then solve for p_i, we get

$$p_i = D_i + E_iu_i - E_iu_{i+1}, \quad (13)$$

where

$$D_i = p'_i + \Delta t\overline{\theta}\dot{p}'_i - E_i\beta_i$$
$$E_i = \Delta t\theta/(C_i + \alpha_i)$$

The equation of motion (4) and the Eqs. (6) and (7) are discretized analogously by expanding the first derivatives of the volume velocities as

$$p_{i-1} - p_i = u_i\left(\frac{L_{i-1,i}}{\Delta t\theta} + R_{i-1,i}\right) - L_{i-1,i}\left(\frac{u'_i}{\Delta t\theta} + \frac{\overline{\theta}}{\theta}\dot{u}'_i\right),$$

$$p_N = u_{N+1}\left(R_N + R_R + \frac{L_N}{\Delta t\theta}\right) - u_{N+2}(R_R) - L_N\left(\frac{u'_{N+1}}{\Delta t\theta} + \frac{\overline{\theta}}{\theta}\dot{u}'_{N+1}\right), \quad (14)$$

$$-u_{N+1}(R_R) + u_{N+2}\left(\frac{L_R}{\Delta t\theta} + R_R\right) = L_R\left(\frac{u'_{N+2}}{\Delta t\theta} + \frac{\overline{\theta}}{\theta}\dot{u}'_{N+2}\right),$$

where $L_{i-1,i} = L_{i-1} + L_i$ and $R_{i-1,i} = R_{i-1} + R_i$. When the p_i ($1 \le i \le N$) are substituted by Eq. (13) and u_1 is provided as the excitation function, then a linear system can be defined for computing the unknown volume velocities $u_2 \dots u_{N+2}$ by means of Eqs. (14). The linear system of equations has a tridiagonal coefficient matrix and can be solved with great efficiency using the Thomas-algorithm [2, p. 59]. By means of the volume velocities

u_i and the equations given in this section, the following quantities must be computed for all indices i at the current sampling point: $\dot{u}_i$, p_i, $\dot{p}_i$, $u_{\mathrm{w}i}$, $\dot{u}_{\mathrm{w}i}$ and $\ddot{u}_{\mathrm{w}i}$. They are needed in order to compute the coefficient matrix for the linear system in the following time step.

Since the resistance due to viscous friction in Eq. (1) is frequency dependent, it can not be used in this form in the TDS. Therefore we use, analogous to Maeda [5], the Hagen-Poiseuille formula for the flow resistance in our TDS: $R_i = 4\mu l_i \pi / A_i^2$.

3. Frequency domain simulation

For the vocal tract simulation in the frequency domain we recall that each tube section can be represented by a two-port network as in Fig. 2. The impedances Z_{a} and Z_{b} for a tube section i are

$$\begin{aligned} Z_{\mathrm{a}i} &= R_i + j\omega L_i \\ Z_{\mathrm{b}i} &= (R_{\mathrm{w}i} + j\omega L_{\mathrm{w}i} + 1/(j\omega C_{\mathrm{w}i})) \parallel C_i. \end{aligned}$$

The input-output relations for such a two-port network can be written in matrix form as

$$\begin{pmatrix} p_{\mathrm{out}} \\ u_{\mathrm{out}} \end{pmatrix} = \begin{pmatrix} 1 + Z_{\mathrm{a}}/Z_{\mathrm{b}} & -2Z_{\mathrm{a}} - Z_{\mathrm{a}}^2/Z_{\mathrm{b}} \\ -1/Z_{\mathrm{b}} & 1 + Z_{\mathrm{a}}/Z_{\mathrm{b}} \end{pmatrix} \begin{pmatrix} p_{\mathrm{in}} \\ u_{\mathrm{in}} \end{pmatrix}.$$

The matrix K_{tot} for the entire vocal tract is the product of the individual matrices K_i from the mouth opening to the glottis: $K_{\mathrm{tot}} = K_N K_{N-1} \ldots K_2 K_1$. The input-output relations for the *entire* network in Fig. 3 are then

$$\begin{pmatrix} u_{N+1} Z_{\mathrm{R}} \\ u_{N+1} \end{pmatrix} = K_{\mathrm{tot}} \begin{pmatrix} p_{\mathrm{g}} \\ u_1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} p_{\mathrm{g}} \\ u_1 \end{pmatrix},$$

where p_{g} is the pressure across the volume velocity source and $Z_{\mathrm{R}} = R_{\mathrm{R}} \parallel j\omega L_{\mathrm{R}}$ is the radiation impedance. The volume velocity transfer function can be derived as

$$H(\omega) = \frac{u_{N+1}(\omega)}{u_1(\omega)} = \frac{1}{A(\omega) - C(\omega) Z_{\mathrm{R}}(\omega)}. \qquad (15)$$

4. Investigations

We have investigated the first four formants of the vowels /a/, /ae/, /e/ and /u/. The area functions for these vowels were produced by the articulatory model in [1] and discretized in 32 tube sections of equal length. The entire vocal tract length varied from 15.02 cm to 17.53 cm depending on the vowel. Our first experiments have shown that not all values for θ between 0 (explicit scheme) and 1 (implicit scheme) are sensible. For $\theta < 0.5$, all simulations became unstable and for $\theta > 0.54$ the bandwidths of the higher formants became unnaturally high. Therefore, we have restricted our simulations to the following limited set of values: $\theta \in \{0.5, 0.51, 0.52, 0.53, 0.54\}$. $\theta = 0.5$ corresponds to the trapezoid rule and values above 0.5 shift the weight from the trapezoid rule to the implicit finite-difference scheme. Each combination of a vowel and a θ-value was simulated at the sampling rates 44.1 kHz and 88.2 kHz.

In order to determine the formant frequencies and bandwidths, the time domain simulations were excited with a volume velocity impulse. The discrete impulse responses $u_N[n]$ were recorded for 0.743 s. From each impulse response, the magnitude spectrum was calculated by means of the discrete Fourier transform. Estimates of the formant frequencies were obtained by parabolic interpolation around the corresponding peaks in the magnitude spectra. For each formant, the 3 dB bandwidth was calculated. The *exact* frequencies and bandwidths of the formants were determined by means of the FDS for comparison. Therefore, we have computed the transfer functions according to Eq. (15) with a resolution of 1 Hz and evaluated them analogously to the TDS-spectra. The comparison of the TDS and FDS formants allowed us to assess the θ-value(s) giving a best spectral match. Since the structure of the acoustic network is identical for both the TDS and FDS, all differences in the transfer functions can be attributed only to the temporal discretization and the different expressions for the resistance components.

		$\theta = 0.5$	$\theta = 0.52$	$\theta = 0.54$	FDS
/a/	B1	16.6	22.0	27.0	28.2
	B2	36.3	47.2	57.7	49.5
	B3	132.9	199.3	265.1	147.8
	B4	10.1	137.9	285.4	34.4
/ae/	B1	12.1	16.2	20.5	19.7
	B2	109.8	136.0	161.7	117.6
	B3	70.6	122.3	173.8	85.5
	B4	33.7	134.9	240.6	52.8
/e/	B1	16.1	18.8	21.8	26.6
	B2	13.3	56.5	101.7	23.8
	B3	186.1	277.1	384.0	216.7
	B4	56.2	174.0	324.8	73.5
/u/	B1	22.1	25.4	28.9	31.5
	B2	5.8	12.3	18.7	22.2
	B3	2.2	40.7	79.0	16.1
	B4	1.6	90.2	172.4	15.3

Table 1: Formant bandwidths of the vowels for different values of θ at 44.1 kHz. The reference values from the FDS are given for comparison.

5. Results and Discussion

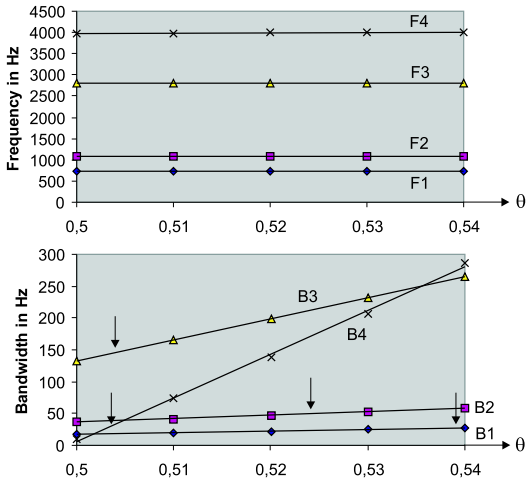

Figure 4: Discrete measurements and regression lines for the formant frequencies and bandwidths over θ for the vowel /a/. The arrows indicate the θ-values, where the bandwidths are equal to the exact bandwidths according to the FDS.

The tables 1 and 2 show the measured data for the sampling rate 44.1 kHz. From the TDS, the formant frequencies and bandwidths are listed for $\theta \in \{0.5, 0.52, 0.54\}$. The reference values from the FDS are listed in the right column. Both the formant frequencies and the bandwidths show a nearly linear dependence on θ in the ex-

		$\theta = 0.5$	$\theta = 0.52$	$\theta = 0.54$	FDS
/a/	F1	735.9	738.4	740.9	737.1
	F2	1093.8	1094.3	1094.6	1095.9
	F3	2808.4	2811.8	2812.8	2848.8
	F4	3976.1	3990.2	4006.2	4086.1
/ae/	F1	663.1	665.5	667.9	666.3
	F2	1717.6	1719.8	1721.5	1726.4
	F3	2464.0	2464.6	2463.6	2489.9
	F4	3531.7	3538.2	3540.7	3608.0
/e/	F1	339.6	342.6	345.5	342.4
	F2	2318.6	2322.7	2327.8	2339.5
	F3	3031.1	3034.3	3038.4	3076.4
	F4	3724.5	3726.9	3720.4	3813.6
/u/	F1	310.6	314.4	318.1	312.6
	F2	874.4	875.9	877.2	874.2
	F3	2206.0	2208.0	2208.3	2226.8
	F4	3380.7	3386.3	3388.4	3446.5

Table 2: Formant frequencies of the vowels for different values of θ at 44.1 kHz. The reference values from the FDS are given for comparison.

amined interval. The results for $\theta = 0.51$ and $\theta = 0.53$, which are not listed in the tables, confirm this trend. Fig. 4 shows the measured formant frequencies and bandwidths over θ together with the regression lines for the vowel /a/. The gradients of the lines (the regression coefficients) are clearly smaller for the formant frequencies than for the bandwidths. Fig. 5 shows the regression coefficients of the bandwidths of all vowel formants over the corresponding formant frequencies. It can be seen that the bandwidth grows faster with increasing θ when the formant frequency is higher. A similarly regular dependence of the formant frequency regression coefficients was not observed.

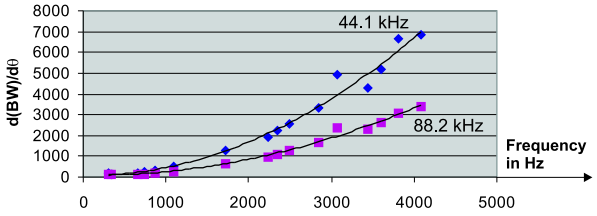

Figure 5: Regression coefficients of all measured formant bandwidths over the formant frequencies for 44.1 kHz and 88.2 kHz.

Concerning the differences between the TDS and FDS, all formant frequencies and bandwidths in the TDS with the trapezoid rule ($\theta = 0.5$) are smaller than the exact values of the FDS (the only exception is F2 of /u/).

The reasons for the bandwidth deviations are not only the temporal discretization but also the different expressions for the resistance R_i. In the FDS, R_i is proportional to $\sqrt{f}$ (Eq. 1) and for this reason, according to Flanagan [3, p. 52], also its contribution to the bandwidths. In the TDS we have approximated viscous friction by means of the *flow* resistance. The flow resistance is frequency independent and causes a smaller contribution to the bandwidths than the boundary layer resistance of the FDS, especially for middle and high frequencies.

The reason for the lower formant frequencies in the TDS is the frequency warping caused by the finite-difference approximation. Maeda [5] has derived a formula that describes this dispersion relation quantitatively. Essentially, frequency warping gets stronger when the temporal and spatial sampling rate is decreased and the higher the formant frequency is.

When θ is increased in excess of 0.5, both the formant frequencies and bandwidths move toward the reference values. Depending on the formant frequencies, the bandwidths in the TDS cross the reference values at certain values of θ. For the vowel /a/, these values (marked by black arrows in Fig. 4) vary between 0.504 and 0.54. The formant frequencies in the TDS reach the reference values only for low frequencies and high θ-values.

For the 88.2 kHz simulations the trends were very similar to those described for the 44.1 kHz simulations. However, the formant frequencies for $\theta = 0.5$ were much closer to the reference frequencies and their regression coefficients were generally smaller compared to the 44.1 kHz simulations. The bandwidths for $\theta = 0.5$ were approximately the same for both sampling rates, but the according regression coefficients were about half of those for the 44.1 kHz simulation (cf. Fig. 5).

6. Conclusions

For a time domain simulation of acoustic propagation in the vocal tract we have examined how formant frequencies and bandwidths change, when the temporal discretization scheme is shifted from the trapezoid rule ($\theta = 0.5$) towards the implicit finite-difference approximation ($\theta = 1$) in small steps. The data show an increase of both formant frequencies and bandwidths for small increments of θ. Thereby they move towards the ideal reference values that were determined by means of a frequency domain simulation. While the increase of the formant frequencies is very small, the bandwidths exceed the reference values for certain values of θ, depending on the corresponding formant frequencies. For the first four formants, this θ-value varies approximately between 0.504 and 0.54. Preliminary listening tests with synthesized vowels have shown a clear preference for $\theta = 0.52$ (compared to $\theta = 0.5$ and $\theta = 0.54$). Essentially, it was shown that a lack of accuracy of formant frequencies and bandwidths in the TDS can partly be compensated by a proper choice of the temporal discretization scheme.

7. Acknowledgments

This research was supported by the Graduate College 466 of the German National Research Council (DFG).

8. References

[1] P. Birkholz and D. Jackèl, "A three-dimensional model of the vocal tract for speech synthesis," in *Proc. 15th ICPhS 2003*, Barcelona, Spain, Aug. 2003, pp. 2597–2600.

[2] J.-J. Chattot, *Computational Aerodynamics and Fluid Dynamics*. Springer Berlin Heidelberg, 2002.

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

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

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

[6] K. Ishizaka, J. C. French and J. L. Flanagan, "Direct determination of vocal tract wall impedance," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, pp. 370–373, 1975.