Tuning of vocal tract model parameters for nasals using sensitivty functions

W. Kreuzer and C. H. Kasess
Acoustics Research Institute, Austrian Academy of Sciences,
Wohllebengasse 12–14, A-1040 Vienna, Austria

March 6, 2015

Abstract

Determining the cross-sectional areas of the vocal tract models from the linear predictive coding or autoregressive-moving-average analysis of speech signals from vowels has been of research interest for several decades now. To tune the shape of the vocal tract to given sets of formant frequencies iterative methods using sensitivity functions have been developed. In this paper the idea of sensitivity functions is expanded to a three-tube model used in connection with nasals and the energy-based sensitivity function is compared with a Jacobian-based sensitivity function for the branched-tube model. It is shown that the difference between both functions is negligible if the sensitivity is taken with respect to the formant frequency only. Results for an iterative tuning a three-tube vocal tract model based on the sensitivity functions for a nasal (/m/) are given. It is shown that besides the polar angle, the absolute value of the poles and zeros of the rational transfer function also needs to be considered in the tuning process. To test the effectiveness of the iterative solver the steepest descent method is compared with the Gauss-Newton method. It is shown, that the Gauss-Newton method converges faster if a good starting value for the iteration is given.

PACS numbers: 43.70.Bk, 43.72Ct Keywords: Acoustics, vocal tract simulation, speech production

The following article appeared in JASA Vol.137, No.2 and may be found at Kreuzer W. and Kasess H.C., J. Acoust. Soc. Am. 137, 1021-1031.

1 INTRODUCTION

Computational models for speech production and analysis have been of research interest since the 1960s (e.g. Badin *et al.*, 2005; Fant, 1970; Mathur *et al.*, 2006; Story *et al.*, 1996; Wakita, 1973). The types of models range from parametric models (e.g. Badin *et al.*, 2005), tube or transmission-line models (e.g. Flanagan, 1972; Kelly and Lochbaum, 1962; Lim and Lee, 1993; Markel and Gray, 1976) to more elaborated wave guide and (3D)-FEM/BEM models (e.g. Matsuzaki *et al.*, 2014).

Particularly popular and simple models most suitable for analysis and production of vowel speech are one-tube models. These models assume that the vocal tract (VT) can be represented by a straight segmented tube and, under certain conditions, the flow inside the n-th segment can be approximated by a combination of two polynomials $U_n^+(z)$ and $U_n^-(z)$, which represent one wave traveling from the glottal end to the lips and one wave traveling in the opposite direction. The transfer function of the vocal tract for these models is given by the ratio of the flow at the lips and the flow at the glottis. During the years, several modifications and improvements of one-tube models have been made, for example, by adding radiation conditions at the lips and/or the glottis (e.g. Deng *et al.*, 2004) or using non-rigid boundary conditions for the tube walls (e.g. Makarov and Sorokin, 2004). The estimation of the parameters of such one-tube models has received a lot of attention during the last four decades. Wakita (1973) or Markel and Gray (1976) showed that if the boundary conditions at both ends of the tube model are set in a specific way, it is possible to estimate cross-sectional areas of a one-tube model using linear predictive coding (LPC) analysis of the speech signal of a vowel. Other authors propose a least squares optimization comparing the transfer function of the tube model with a transfer function based on an LPC (all pole transfer function) or autoregressive-moving-average (ARMA, pole-zero transfer function) model of the signal, or by using pre-defined articulatory codebooks (Schroeter and Sondhi, 1994). In Heinz (1967); Mrayati *et al.* (1988); Radolf (2007), and Story (2006) techniques were presented to tune vocal tract area functions for a one-tube model based on acoustic sensitivity functions. These functions describe the sensitivity of a particular formant frequency to changes in the cross-sectional areas of a segment of the tube model.

For nasals like /m/ or /n/ or nasalized vowels a one-tube model is not sufficient because the nasal cavity is coupled to the vocal tract. In Lim and Lee (1996) a branched-tube model to calculate the transfer function of the vocal tract including oral as well as nasal tract was proposed (a similar approach can be found in Schnell and Lacroix (2000)). While the transfer function for most one-tube models (i.e. the ratio of the volume velocities at the glottis and at the lips) can be represented by an all pole transfer function $H(z) = 1/H_1(z)$, the transfer function

for a branched-tube model is given by a rational function $H(z) = H_2(z)/H_1(z)$.

Compared to a one-tube model, the estimation of the cross-sectional area of each segment from the poles and zeros of the transfer function (i.e. the roots of $H_2(z)$ and $H_1(z)$) of an arbitrary signal is more complicated. For a branched-tube model, the total number of poles and zeros is bigger than the number of cross-sections and an approximate solution of a system of non-linear equations is needed. Lim and Lee (1996), for example, proposed to minimize the difference between the polynomial coefficients of the rational transfer function of an ARMA model of the speech segment and the transfer function generated by their branched-tube model. For the parameters of the oral tract the step-down algorithm from Markel and Gray (1976) is used, for the remaining coefficients a least squares optimization method is applied. In Schnell and Lacroix (2000) a method based on inverse filtering the input signal is used. First, the Burg method is applied alternately to the signal to determine the poles and zeros of the transfer function of an ARMA model. In a second step the reflection coefficients for the side branch (in case of nasals this is the oral tract) are determined from the numerator of the ARMA model. In a third step the remaining reflection coefficients are estimated by a special iterative inverse filtering process on the input signal. An alternative for estimating the VT tube parameters directly from the log spectral envelope of the signal without the need of an explicit pole-zero model was presented in Kasess *et al.* (2012). The method is based on a Bayesian scheme which allows adding constraints on the tube parameters, for example, by centering a Gaussian prior for the reflection coefficients around zero, thus preferring a smoother vocal tract.

In this paper the use of sensitivity functions is extended from a one-tube model to the branched-tube model. A comparison with a sensitivity based on the Jacobian of the model is made and these functions are used to iteratively tune the parameters of the tube model (i.e. the cross-sectional areas) based on given poles and zeros.

This paper is structured as follows: In Section 2 a short introduction to the one and multi-branched-tube models is given. In Section 3 the energy-based sensitivity (cf. Story, 2006) is briefly introduced and compared with a Jacobian-based sensitivity. It is shown that if the Jacobian-based sensitivity is defined with respect to the frequency only, both sensitivity functions produce similar results. In Section 4 the iterative method to tune the cross-sectional areas to given poles and zeros of the transfer function is described and the performance of the iteration using steepest descent and Gauss-Newton methods is investigated in Section 5 and demonstrated with numeral examples.

2 VOCAL TRACT MODEL

2.1 One-tube model

Under the assumption of a rigid tube where losses at the walls can be neglected, the transfer function of a segmented tube can be calculated using the solution of the 1D-Helmholtz equation inside each segment. The (analytic) solution is given by the linear combination of two plane waves, one traveling from the right end of the tube (i.e. the lips) to the left end (i.e. the glottis), the other one traveling in the opposite direction. Inside the n-th segment the pressure and the volume velocity are given by

$$p_n(t,x) = \frac{\rho c}{A_n}\left(u_n^+(t,x) + u_n^-(t,x)\right) \tag{1}$$

and

$$u_n(t,x) = u_n^+(t,x) - u_n^-(t,x), \tag{2}$$

where u_n^+ and u_n^- represent the right and left traveling wave in the n-th segment with cross-sectional area A_n. ρ and c are the density of air and the speed of sound inside the tube.

Wakita (1973) defined the VT transfer function of a tube with N segments as the volume velocity at the front end (i.e. lips) divided by the forward volume velocity component at the back end (i.e. the glottis). It is assumed that the front end of the tube (i.e. the lips) is open, while at the back end a virtual segment with a finite non-zero area A_{N+1} is added. Using continuity of the pressure and flow at the segment boundaries, the z-transforms (with $z = e^{2\mathrm{i}\omega\ell/c}$) of the wave components u_n^+ and u_n^- inside the n-th segment can be calculated as

$$\begin{bmatrix} U_n^+(z) \\ U_n^-(z) \end{bmatrix} = T(z,\mu_{n-1}) \cdots T(z,\mu_1) \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix} U_0, \tag{3}$$

where $U_0 = U_0^+$ is the constant flow at the front end of the tube. The matrices $T(z,\mu_j)$ are given by

$$T(z,\mu_j) = \frac{z^{1/2}}{1-\mu_j} \begin{bmatrix} 1 & \mu_j \\ \mu_j z^{-1} & z^{-1} \end{bmatrix}. \tag{4}$$

$\mu_j = (A_{j+1} - A_j)/(A_{j+1} + A_j)$ denotes the $j-$th reflection coefficient, where A_j is the cross-sectional area of the j-th segment. The transfer function of the tube with N segments is given as:

$$H(z) = \frac{U_0}{U_{N+1}^+(z)}. \tag{5}$$

Please note, that for the calculation of the poles of $H(z)$ the factors $z^{1/2}/(1-\mu_j)$ in Eq. (4) can be neglected, as they don't have any effect on the roots of $U^+_{N+1}(z)$, and that U_0 is contained in the numerator as well as in the denominator in Eq. (5) and thus can be canceled.

It can be shown that in the above setup, where the boundary condition at one end of the tube is lossless and the other end is terminated with an impedance given by a virtual segment with a finite non-zero cross-sectional area, it is possible to determine the reflection coefficients μ_i with an LPC analysis of the speech signal of a vowel (Wakita, 1973). Markel and Gray (1976) present an algorithm to uniquely determine reflection coefficients of a tube model from an all-pole filter model, provided that all poles lie inside the unit circle which means that the reflection coefficients $\mu_j \in (-1, 1)$.

2.2 Branched-tube model

Lim and Lee (1993) proposed a branched-tube model including a nasal tract, which was later upgraded to a "lossy" model (Lim and Lee, 1996). The model consists of three tubes (nasal, oral and pharyngeal tube) with M, N, and L segments each, where the individual tubes are modeled using Eq. (3). To keep notation concise the dependence on z has been dropped, also a slightly different notation than in Lim and Lee (1996) is used:

$$\begin{bmatrix} U_M^{n,+} \\ U_M^{n,-} \end{bmatrix} = T(\mu_{M-1}^n) \cdots T(\mu_1^n) \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix} U_0, \tag{6}$$

$$\begin{bmatrix} U_N^{o,+} \\ U_N^{o,-} \end{bmatrix} = T(\mu_{N-1}^o) \cdots T(\mu_1^o) \begin{bmatrix} 1 \\ \mu_0^o z^{-1} \end{bmatrix} U_0^o, \tag{7}$$

$$\begin{bmatrix} U_{L+1}^{p,+} \\ U_{L+1}^{p,-} \end{bmatrix} = T(\mu_L^p) \cdots T(\mu_0^p) \begin{bmatrix} U_M^{n,+} + U_N^{o,+} \\ U_M^{n,-} + U_N^{o,-} \end{bmatrix}. \tag{8}$$

The upper indices n, o and p in Eqs. (6) to (8) denote the nasal, oral and pharyngeal tracts, respectively. At the glottal end, an additional virtual segment is added to model impedance effects of the glottis (cf. Wakita (1973)). Although Lim and Lee assume in their model that the mouth is closed, they introduce a reflection coefficient $|\mu_0^o| < 1$ at the lips to account for absorption effects in the mouth. μ_0^o thus ensures that the zeros of the transfer function do not lie on the unit circle. The first reflection coefficient for the pharyngeal tract μ_0^p is given by the ratio of the cross-sectional areas of the segments at the branch

$$\mu_0^p = \frac{A_1^p - (A_N^o + A_M^n)}{A_1^p + A_N^o + A_M^n}, \tag{9}$$

which is one of the two parameters describing the coupling of the three tubes. The second coupling parameter σ' is introduced to ensure the continuity of the pressure and the flow at the tube junction which results in the coupling matrix Eq. (13). For nasals, it is assumed that the mouth is closed, and the transfer function of the vocal tract is given by the ratio of the flow at the nostrils and the flow at the glottis:

$$H(z) = \frac{U_0}{U^+_{L+1}(p,z)} = \frac{H_2(z)}{H_1(z)}, \tag{10}$$

where

$$H_1(z) = [1\ \mu^p_L] \cdot T(\mu^p_{L-1}) \cdots T(\mu^p_0) \cdot \mathbf{C}(\sigma', z) \cdot \\ \cdot T(\mu^n_{M-1}) \cdots T(\mu^n_1) \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix}, \tag{11}$$

and

$$H_2(z) = [1\ 1] \cdot T(\mu^o_{N-1}) \cdots T(\mu^o_1) \begin{bmatrix} 1 \\ \mu^o_0 z^{-1} \end{bmatrix}. \tag{12}$$

The coupling matrix $\mathbf{C}(\sigma', z)$ is given by

$$\mathbf{C}(\sigma', z) = \begin{bmatrix} (1+\sigma')U^{o,+}_N + U^{o,-}_N & \sigma' U^{o,+}_N \\ \sigma' U^{o,-}_N & U^{\mathrm{ora},+}_N + (1+\sigma')U^{o,-}_N \end{bmatrix} \tag{13}$$

with $\sigma' = \frac{A^o_N}{A^n_M}$ being the quotient of the area at the oral tract and the nasal tract at the branch.

Compared to one-tube models the estimation of the reflection coefficients of a branched-tube based on an ARMA model of a speech signal is much more complicated. If the number of segments for the pharyngeal, nasal and oral tracts are given by L, M, and N the orders of $H_1(z)$ and $H_2(z)$ are $N + L + M + 1$ and N. To determine the reflection coefficients of the oral tract $\mu^o_i, (i = 0, \ldots, N-1)$, the step-down algorithm presented in Markel and Gray (1976) can be applied to $H_2(z)$. Because of the additional unknown μ^o_0 the oral tube fulfills both boundary conditions for determining the reflections coefficients with the step-down algorithm (just in reverse order). The remaining $L + M$ unknowns corresponding to the μ^p_i $(i = 0, \ldots, L)$ and the μ^n_j $(j = 1, \ldots, M-1)$ for the pharyngeal and nasal tracts and the coupling factor σ' have to be determined from the $N + L + M + 1$ roots of the polynomial $H_1(z)$ using, for example, the method of least squares. It is obvious, that the set of transfer functions provided by an ARMA filter is much bigger than the set of transfer functions generated by the tube model. It is recommended to introduce additional constraints to enhance the stability and efficiency of the least squares solver and to make sure that the least squares fit

results in reasonable cross-sectional areas. For example, in this study all cross-sectional areas are assumed to be inside a fixed interval $[A_{\min}, A_{\max}]$.

As for the one-tube model the factors $z^{1/2}/(1-\mu_j)$ (see Eq. (3)) and the factors U_0 and U_0^o are not relevant for the poles and zeros (i.e. the roots of $H_1(z)$ and $H_2(z)$) and can be neglected in Eq. (11). However, they have to be considered when calculating the energy inside the vocal tract needed for the energy-based sensitivity function.

3 SENSITIVITY FUNCTIONS

3.1 Energy-based sensitivity

In Carre (2004); Fant (1970), and Story (2006) acoustic sensitivity functions were defined as the scaled difference between the kinetic (KE) and the potential (PE) energy in the j-th segment at the frequency of the i-th formant

$$S_{ij}^{\text{Energy}} = \frac{\text{KE}_i(j) - \text{PE}_i(j)}{\sum_{k=1}^{N_{\text{areas}}} \text{KE}_i(k) + \text{PE}_i(k)}. \tag{14}$$

For the j-th segment the kinetic and the potential energy are given as (cf. Story, 2006):

$$\text{KE}_i(j) = \frac{1}{2}\frac{\rho\ell}{A_j}|U_j(z_i)|^2 = \frac{1}{2}\frac{\rho\ell}{A_j}|U_j^+(z_i) - U_j^-(z_i)|^2 \tag{15}$$

and

$$\text{PE}_i(j) = \frac{1}{2}\frac{A_j\ell}{\rho c^2}|P_j(z_i)|^2 = \frac{1}{2}\frac{\rho\ell}{A_j}|U_j^+(z_i) + U_j^-(z_i)|^2, \tag{16}$$

where z_i is given by the frequency of the i-th formant, ℓ denotes the length of the segments, and c and ρ denote the speed of sound and the density of air.

By construction, the flow in each segment of the oral and the nasal tract is dependent on the constant flows U_0 and U_0^o at the nostrils and at the lips, respectively. Thus, to compare the energy in the oral and nasal tracts, U_0^o has to be expressed as a function of the flow at the nostrils U_0. By continuity of the pressure at the branch of the tube model

$$p_M^n = p_N^o = \rho c\frac{U_M^{n,+} + U_M^{n,-}}{A_M^n} = \rho c\frac{U_N^{o,+} + U_N^{o,-}}{A_N^o} \tag{17}$$

and with Eq. (3) it follows that

$$U_0^o = \frac{A_N^o}{A_M^n} \cdot \frac{[1\ 1] \cdot \prod_{j=1}^{M-1} T(\mu_{M-j}^n) \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix}}{[1\ 1] \cdot \prod_{j=1}^{N-1} T(\mu_{N-j}^o) \begin{bmatrix} 1 \\ \mu_0^o z^{-1} \end{bmatrix}} U_0. \tag{18}$$

It is noted that U_0 has no influence on the energy-based sensitivity function. It is contained in the denominator and the numerator in Eq. (14) and can therefore be canceled.

3.2 Jacobian-based sensitivity

When using the tube model, the poles of the transfer function $H(z)$ (thus the roots of $H_1(z)$) can be used to determine the frequencies of the formants $F_i = \arg(z_i)F_s/(2\pi)$, where z_i is the i-th root of $H_1(z)$ and F_s denotes the sample frequency. In the same manner the frequencies of the anti-formants can be deduced by the roots of $H_2(z)$.

As the coefficients of the polynomials are functions of the cross-sectional areas $\mathbf{A} = (A_1, \ldots, A_{N+M+L+1})$, the roots z_i can be seen as functions of the cross-sectional areas. It is therefore natural to describe the "sensitivity" of the poles (and zeros) using the Jacobian (or an approximation of the Jacobian) of $z_i(\mathbf{A})$ (for brevity only the Jacobian with respect to the poles is presented here, for the zeros the calculation is analogous)

$$J_{ij} = \frac{\partial z_i(\mathbf{A})}{\partial A_j} \approx \frac{z_i(\mathbf{A} + \Delta A \mathbf{e}_j) - z_i(\mathbf{A})}{\Delta A}, \tag{19}$$

where $\mathbf{e}_j$ is the j-th unit vector and $z_i(\mathbf{A})$ is the i-th pole calculated by the model (i.e. the i-th root of the polynomial $H_2(z)$ in Eq. (11)). In Story (2006) the sensitivity function is calculated with respect to the formant frequency and it is by construction dimensionless. In a similar way a Jacobian-based sensitivity function can be defined by using the polar angle $\arg(z_i)$ instead of z_i. If the sensitivity function is scaled with a factor $A_j/\arg(z_i)$ it becomes dimensionless and is given by

$$\arg(J_{ij}) \frac{A_j}{\arg(z_i(\mathbf{A}))}, \tag{20}$$

which can be approximated by $\mathrm{imag}(S_{ij}^{\mathrm{Jac}})$ where

$$S_{ij}^{\mathrm{Jac}} = \frac{\log\left(z_i(\mathbf{A} + \Delta A \mathbf{e}_j)/z_i(\mathbf{A})\right)}{\Delta A} \cdot \frac{A_j}{\arg z_i(\mathbf{A})}. \tag{21}$$

The S_{ij}^{Jac} can be collected as entries of the complex valued $n \times m$ matrix $\mathbf{S}^{\text{Jac}}, i = 1, \ldots, m; j = 1, \ldots, n$, where n is given by the number of poles and zeros and m is given by the number of unknown cross-sectional areas. The definition of a sensitivity using the logarithm has an additional advantage. While the imaginary part of Eq. (21) describes a sensitivity with respect to the formant frequency $F_i = \arg(z_i)F_s/2\pi$, the real part of $\mathbf{S}^{\text{Jac}}$ describes a sensitivity with respect to the bandwidth $B_i = -F_s/\pi \log(|z_i|)$ (cf. Markel and Gray, 1976, p. 167).

In order to be able to compare two sets of roots (for example during the computation of the Jacobian the difference of $z_i(\mathbf{A})$ and $z_i(\mathbf{A} + \Delta A_j)$ needs to be calculated) it is necessary to order them in a fixed way. In this study the ordering is as follows: The first entries of vector $\mathbf{z} = (z_1, \cdots, z_n)$ consist of the roots with an imaginary part bigger zero ordered by their polar angle. These entries are followed by the real valued roots ordered by their value from lowest to highest. In case that one set S1 contains more real valued roots than the other set S2, which in turn contains more roots with non-zero imaginary part than S1, two real valued roots of S1 have to be matched to one complex valued root pair of S2. For the matching process the two subsets R1 and R2 of S1 and S2 that contain only the real valued roots are investigated first. The two elements of R1 with the biggest distance $d(x, R2) = \inf\{(||x - y||, y \in R2)\}$ to R2 are chosen for the real valued roots of S1 to be matched. This approach is repeated for the subsets of roots with non-zero imaginary parts to find the complex valued root-pair from S2 used for the matching.

3.3 Comparison

In this section the energy-based and the Jacobian-based sensitivity functions for a given cross-sectional area configuration for a nasal /m/ based on the examples given in Story (1995) are compared. The cross-sectional areas (as functions of the distance from the glottis) for the branched-tube model used in this example are given in Fig. 1. The model consists of $L = 32, M = 40$ and $N = 32$ segments for pharyngeal, nasal and oral tract, respectively. Similar to Story (1995) it is assumed that the combined lengths of pharyngeal and oral tracts is 16 cm, thus each segment has a length of 0.25 cm. Above the x-axis the cross-sectional areas for the pharyngeal and nasal tracts are given, below the x-axis the area function of the oral tract is depicted. The area of each segment is given by the value at the midpoint of the segment (see Story, 1995, Figs. 5-15a,b and 5-17b). For the segments of the nasal tract the mean value of right and left nasal branch given in Story (1995, Fig 5-15b) is used.

At the glottal end and at the lips additional virtual segments have been added

to model lossy boundary conditions. For the numerical calculations the reflection coefficient at the lips was set to $\mu_0^o = 0.877$, which is equivalent to a virtual segment at the end of the lips with an area of 0.01 cm^2. The reflection coefficients at the nostrils and at the glottis were set to -1.0 and -0.74 (equivalent to a virtual segment of size $A_g = 0.1$ cm^2), respectively . For the Jacobian-based functions Eq. (21), $\Delta A = 1 \cdot 10^{-6}$ cm^2 is assumed.

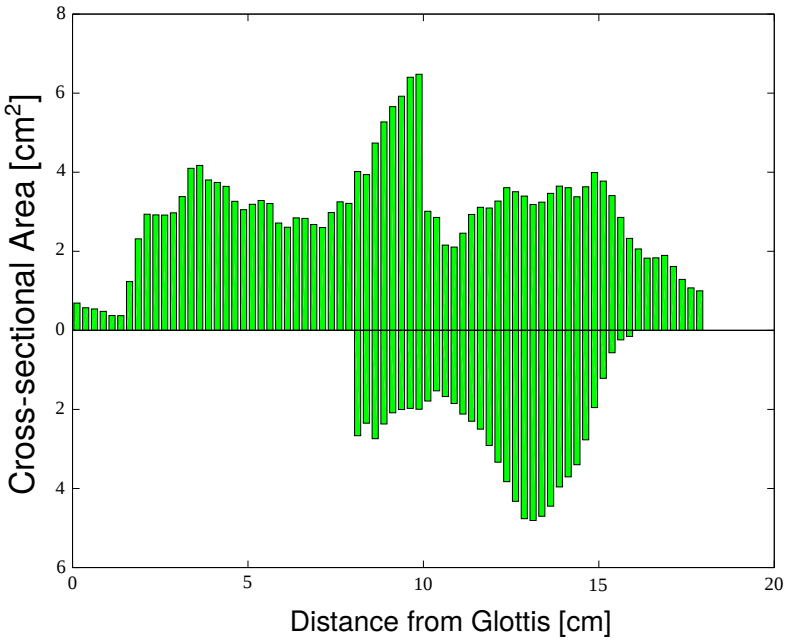

Figure 1: Vocal tract cross-sectional areas for a nasal /m/ as a function of the distance from the glottis used for the comparison of the sensitivity functions. The areas of the tube segments for pharyngeal and nasal tracts are depicted above the x axis, the areas for the oral tract below (Color online).

In Fig. 2 the energy-based (line) and the Jacobian-based (dashed lines) sensitivity functions with respect to the formant frequency (Eqs. (14) and (20)) for a branched-tube with the cross-sectional areas coefficients based on the measurements from Story (1995) (see Fig. 1) are compared. As the energy-based sensitivity function is only given with respect to the frequency of the pole/zero, just the imaginary part of $\mathbf{S}^{\mathrm{Jac}}$ is used in the comparison. The sensitivity functions for the first two poles are depicted in the upper row of the figure, the sensitivity functions for the zeros in the lower row. The sensitivity functions for the pole frequencies are given by three lines as function of the distance from the glottis for the pharyngeal, nasal and oral segments, respectively. For the sensitivity of the zeros it is sufficient to only consider the sensitivity with regard to changes of the oral tract, since the

zeros are only dependent on the reflection coefficients μ_i^o (see Eq. (12)).

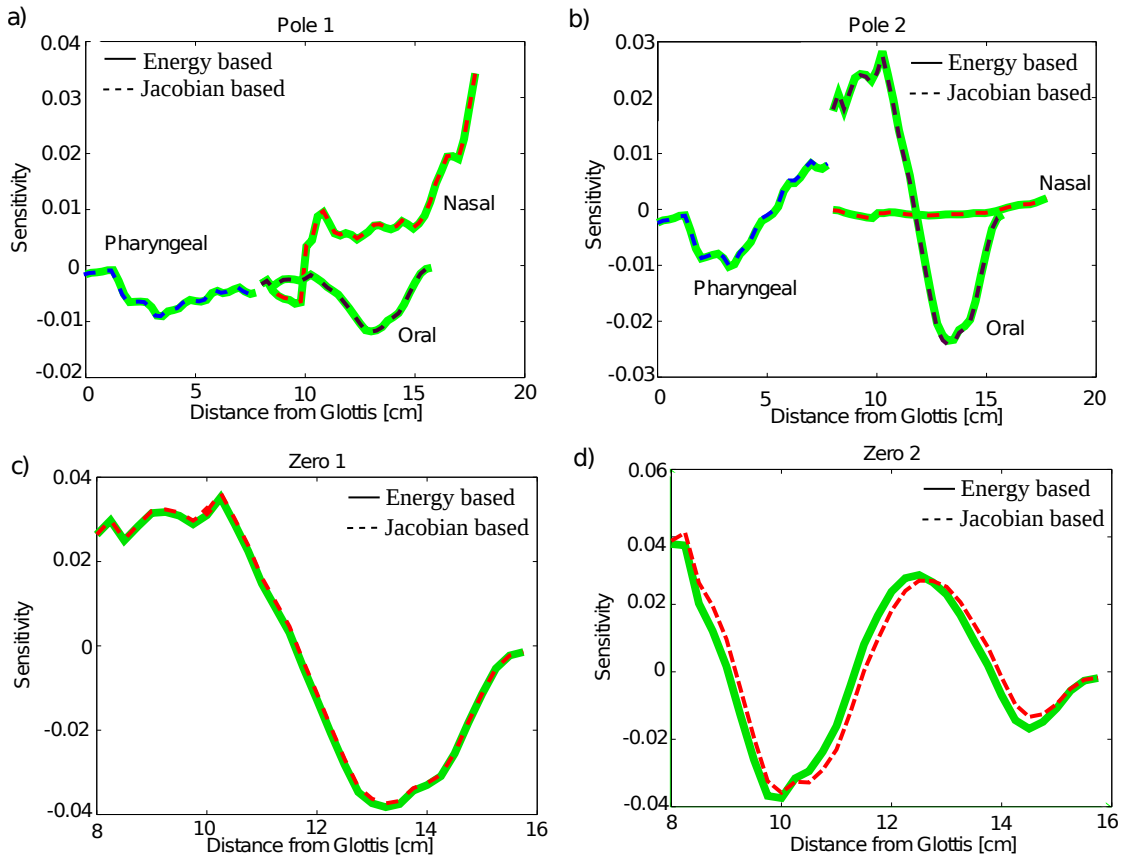

Figure 2: Comparison of the energy-based (lines) and the Jacobian-based sensitivity functions (dashed lines) with respect to the frequency of the first two poles (upper row) and the first two zeros (lower row). In subfigures a) and b) the sensitivity functions are given for the pharyngeal, nasal and oral tracts, for the sensitivity of the zeros (subfigures c and d) only the oral tract has to be considered. Only the imaginary part of the $\mathbf{S}^{\text{Jac}}$ is used in the comparison (Color online).

For the sensitivity with respect to the poles only small differences can be observed between the two types of sensitivity functions (see Figs. 2a,b), for the sensitivity with respect to the zeros, a slight shift of the two curves can be observed, especially for the second zero (Figs. 2c,d).

4 ITERATION

In Story (2006) the sensitivity function for a one-tube model was used to tune cross-sectional areas for some target formant changes. This approach shall now be extended to the three-tube model. In the approach presented in this study the complex valued Jacobian-based sensitivity function $\mathbf{S}^{\mathrm{Jac}}$ is used because it has the advantage that it can be constructed for complex valued poles and zeros. Thus, the definition of the error and the sensitivity functions includes not only the frequency (imaginary part of Eq. (21)) but also the bandwidth (represented by the real part of Eq. (21)) of the (anti)formant.

By construction, the zeros in the model of Lim and Lee are dependent only on the oral tract. Lim and Lee (1996) use a step-down algorithm (Markel and Gray, 1976, chapter 5) to determine the reflection coefficients for the oral tract, and then use an iterative least square minimization of the difference in the polynomial coefficients to determine the coefficients for the nasal and pharyngeal tracts. In this study a slightly different approach is used. The error is defined with respect to the differences in the poles and zeros and the step-down algorithm is only used to get a first estimate for the oral tract. The areas of the oral tract are still unknowns to be determined in the following iteration steps. This has the advantage that the approximation error for arbitrary poles and zeros can be balanced between poles and zeros.

For the iteration, a (damped) Gauss-Newton scheme with very simple step size control is used. As the sensitivity is defined with respect to relative changes, the updated cross-sectional area vector $\mathbf{A}^{[n+1]}$ is given by:

$$\mathbf{A}^{[n+1]} = \mathbf{A}^{[n]} \left(1 + h_0 \mathbf{M}^{\dagger} \mathbf{r}^{[n]}\right), \tag{22}$$

where $\mathbf{M}^{\dagger}$ denotes the pseudo-inverse of

$$\mathbf{M} = \begin{bmatrix} \mathrm{Imag}(\mathbf{S}^{\mathrm{Jac}}) \\ \mathrm{Real}(\mathbf{S}^{\mathrm{Jac}}) \end{bmatrix} \tag{23}$$

and $\mathbf{r}^{[n]}$ is given by the imaginary and real parts of the logarithm of the relative difference vector of estimated and target poles and zeros after the n-th step. h_0 defines the length of the update step. An iteration step with step size h_0 is accepted if $||\mathbf{r}^{[n+1]}|| \leq \alpha ||\mathbf{r}^{[n]}||$ and the updated cross-sectional areas are still within allowed bounds ($A_i \in [0.1, 20]$), otherwise $h_0 \leftarrow h_0/2$ and the step is repeated. The control parameter α may be set bigger than 1.0 to avoid convergence to a local minimum.

Additionally, a simple procedure to adapt the step size is chosen. In each step the residual is calculated for step sizes h_0, $h_0/2$ and $2h_0$ and the step size with the smallest norm of the residual is chosen for the next iteration step.

As the matrices involved are small, the additional computational effort of the Gauss-Newton approach is negligible compared with a steepest descent approach (cf. Story (2006))

$$\mathbf{A}^{[n+1]} = \mathbf{A}^{[n]} + h_0 \mathbf{M}^T \mathbf{r}^{[n]}, \tag{24}$$

It is known, that if a good starting value for the iteration is available the convergence of the Gauss-Newton method is faster than the steepest descent method , thus the additional effort is justified. The iteration is stopped if either

- the norm of the residual $\mathbf{r}^{[n]}$ is below a given error tolerance or
- the norm of the gradient $\mathbf{M}^T \mathbf{r}^{[n]}$ is below a given error tolerance or
- the step size h_0 drops below a given tolerance.

5 NUMERICAL EXPERIMENTS

In all numerical experiments the iterative process described in Section 4 is used to tune the cross-sectional areas of the VT for given target poles and zeros. The first experiment uses a setup where the target poles and zeros are in the subgroup of all poles and zeros that can be generated by the branched-tube model. The performance of the tuning algorithm for two different sensitivity functions is compared. First, the sensitivity is only based on the frequency, thus only the imaginary part of $\mathbf{S}^{\text{Jac}}$ is used. For the second sensitivity function the real and the imaginary part of $\mathbf{S}^{\text{Jac}}$ is used. Additionally, a comparison of the convergence behavior of the Gauss-Newton method compared with the steepest descent method is made. Also the case where the target poles and zeros are given by an ARMA model of the log-spectral envelope for a recorded nasal /m/ of a male speaker will be discussed.

In all numerical experiments it is assumed that the pharyngeal, nasal and oral tracts consist of 8, 10 and 8 segments, respectively. This choice of segment numbers is a compromise between the assumption of plane wave propagation inside the tubes on the one hand and the motivation to create a challenging numerical setup for the iteration procedure on the other.

For the first experiment the target poles and zeros are determined using the branched-tube model with the data for a nasal /m/ given in Story (1995). To that end, the x-axis in Figs. 5-15a,b and 5-17b in Story (1995) was subdivided into 10 and 16 intervals, respectively, and the cross-sectional area of each target segment was taken as the value of the functions depicted in Figs. 5-15 and 5-17 at the midpoint of the corresponding interval. For the cross-sectional areas of the nasal tube, the mean of the left and right nasal branches is used. The length of pharyngeal plus oral tract is assumed to be 16.0 cm, thus the length of each

segment is 1.0 cm. Assuming the speed of sound to be $c = 340$ m/s this choice of segment numbers corresponds to a sampling frequency of $F_s = 17$ kHz.

For the second case where the target poles and zeros are determined using an ARMA model on a recorded nasal of a male speaker, a setup often found in literature is used. For the tube model it is assumed that the pharyngeal and the oral tract combined have a length of 17.5 cm for an average male and that $c = 350$ m/s. This setting corresponds to a sample frequency of $F_s = 16$ kHz.

The choice of different vocal tract parameters may seem odd at a first glance, it was mainly done for "cosmetic" reasons. By construction of the model the sampling frequency is related to the segment length ℓ by $F_s = c/(2\ell)$ and it was assumed that the number of segments is the same for all calculations. In the first case, the total length of the vocal tract was given by the data and c was chosen to get an integer for F_s. In the second case, however, the sampling frequency was given by the sound sample, and it was decided to use a vocal tract length often used in literature for a male speaker. This choice corresponds to $c = 350$ m/s.

The iteration process is started with a "neutral" configuration for the nasal and the pharyngeal tracts, i.e. all areas are set to 1.0 cm^2, except for the virtual segment at the end of the glottis. At the nostrils the reflection coefficient is set to -1. For the oral tract the starting values are determined using the step-down algorithm.

Please note that the model is based only on reflection coefficients, thus it is possible to rescale all cross-sectional areas with the same constant factor. In this study the cross-sectional area of the nasal tract segment closest to the nostrils is set to 1.0 cm^2, and all the other cross-sectional areas are given relative to this segment. For the least square error all poles and zeros in the upper half plane including the real line (thus all (anti)formants up to $F_s/2$ Hz) were used.

5.1 Given area function

In this section the target poles and zeros are given by the rational function that is provided by the model for pre-defined cross-sectional areas based on Figs. 5-15a,b and 5-17b in Story (1995). This setup has the advantage that the error can be measured directly between the cross-sectional areas of the tuned and the target model. In Section 5.1.1 the starting values for the cross-sectional areas for the oral tube were calcluated using the step-down algorithm.

5.1.1 Frequency and Bandwidth

In Section 3 it was pointed out that the Jacobian-based sensitivity has the additional advantage that the bandwidths as well as the frequency of the poles and zeros can be considered for the tuning. This advantage shall be illustrated in the following example. In Fig. 3 the results of the tuning process with a sensitivity function only defined with respect to the polar angle are depicted. In Fig. 3a transfer functions of the tuned and the target model are compared, in Fig. 3b the poles and zeros of the tuned and the estimated model in the z-domain are depicted. The norm of the residual and the cross-sectional areas including the virtual segments at the lips and at the glottis are given in Figs. 3c and 3d. In Fig. 4 results of the tuning process are given when the sensitivity function and the residual are defined using both the real *and* the imaginary parts of $\mathbf{S}^{\mathrm{Jac}}$ and $\mathbf{r}^{[n]}$ (see Eqs. (22) and (23)).

Although both iterations converge very quickly and the polar angles of the poles and zeros of target and estimated functions are equal up to $\varepsilon = 1 \cdot 10^{-8}$, there is a visible difference between the transfer functions in the case where the bandwidth is not included in the the definition of the sensitivity. When, for example, poles and zeros are close in frequency, different bandwidths (especially when the pole or zero is close to the unit circle) can change the characteristics of the transfer function completely. In Fig. 3a the target transfer function has a small but visible peak around 1300 Hz that is not visible in the tuned function. In that case the target pole is very close to the unit circle (the pole-zero pair in Fig. 3b at approximately 20°) and compared to the tuned function the balance between the zero and the pole is different, thus the peak is more pronounced. A similar effect can be observed around 3700 Hz. The 5th target pole is very close to the unit circle and the difference in bandwidth between estimated and target pole is relatively large. This differences in the absolute values of the poles and zeros around 3700 Hz to 4000 Hz shifts the visibility of the peak from the 5th pole to the 6th pole. In Fig. 3d the difference in cross-sectional areas is also clearly visible. As the model is based on calculating the reflection coefficients and not the areas themselves, target and estimated cross-sectional areas are scaled to have 1.0 cm^2 at the segment closest to the nostrils. Fig. 3 shows there are at least two different possible sets of cross-sectional areas that produce the same pole/zero frequencies, thus it is not possible to uniquely determine reflection coefficients based on the formant frequency alone. If the bandwidth is included in the definition of the sensitivity function and the error, the iteration converges after 8 iterations and no visible differences in the tuned and target cross-sectional areas are given (see Fig. 4). Thus, the full Jacobian including real and imaginary part is taken in all following numerical experiments.

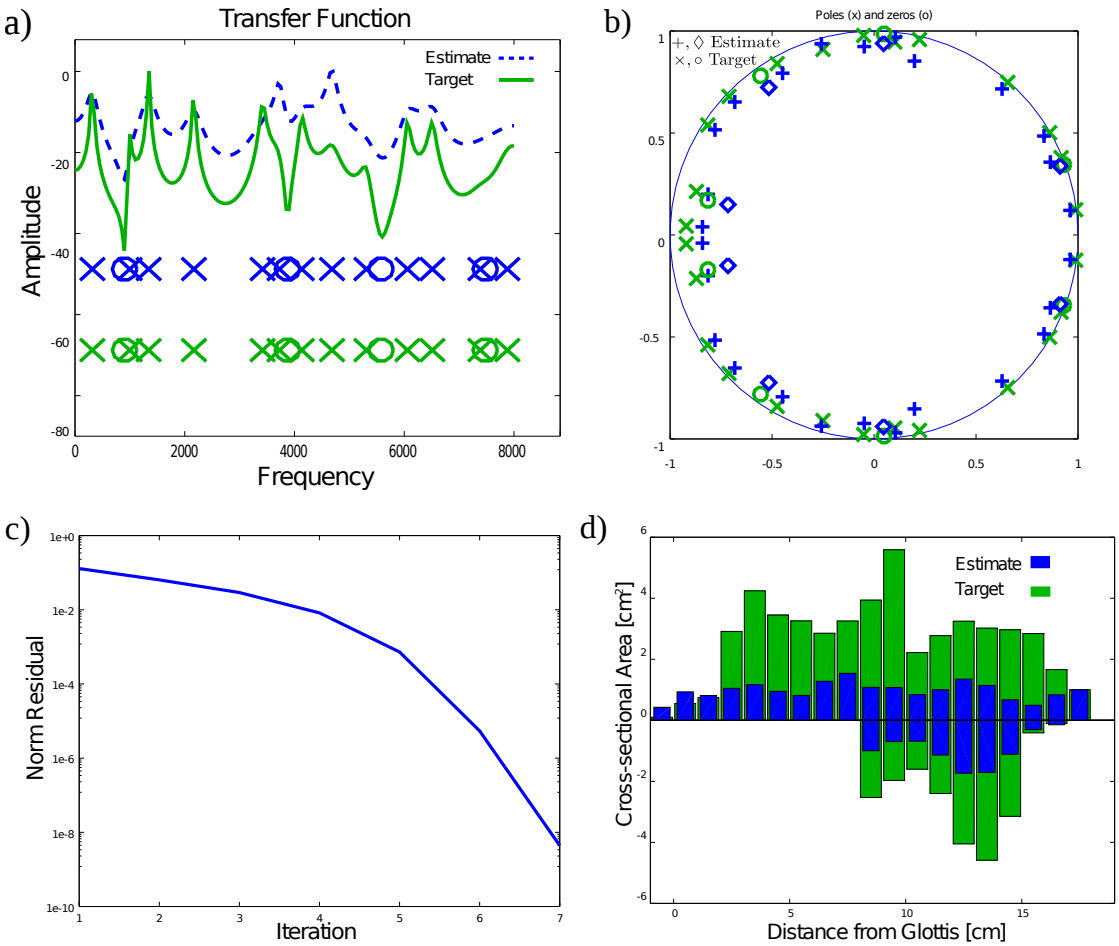

Figure 3: a) Estimated and target transfer functions plus the frequencies of the poles 'x' and zeros 'o', b) position of poles $(\times, +)$ and zeros $(\circ, \diamond)$ inside the unit circle, c) norm of the residual and d) difference in cross-sectional areas of the tube segments including the virtual segments at the lips and at the glottis. For all plots the sensitivity is defined with respect to the frequency only (Color online).

5.1.2 Gauss-Newton vs. Steepest Descent

It is a well known fact that if a good starting value for the iteration is given, the Gauss-Newton converges much faster than the steepest descent method. In Fig. 5 the norm of the residual in each iteration step for the Gauss-Newton method and the steepest descent method is compared. As starting values for the Gauss-Newton

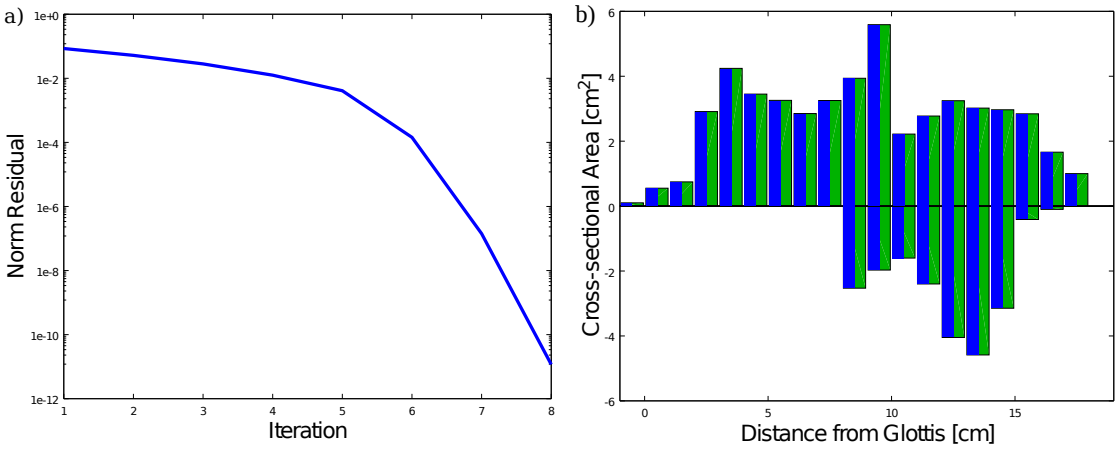

Figure 4: a) Norm of the residual and b) difference in cross-sectional areas after 8 iteration steps with the Gauss-Newton method. The sensitivity is defined with respect to the frequency and the bandwidth, the step-down algorithm was used to get the starting values for the oral cross-sectional areas (Color online).

method the neutral configuration was also chosen for the oral tract, thus the step-down algorithm was *not* used to determine the first estimate for the oral tract. For the steepest descent method the iteration was started with and without using the step-down method. As expected, the iteration converges slower for the steepest descent method than for the Gauss-Newton. After 40 iterations there are still some differences in the transfer functions for the tuned and the estimated models which also has an effect on the tuned cross-sectional areas (Fig. 6b). The scaling of the transfer functions in Fig. 6a was chosen such that the highest peak is set to 0 dB.

Fig. 5 also illustrates the well known fact, that the steepest descent approach is relatively stable with respect to different starting values of the iteration, whereas a good starting value is essential for a fast convergence of the Gauss-Newton method. If the Gauss-Newton method is used without starting values for the oral tract based on the step-down algorithm the method converges very slowly in the beginning until the estimated solution gets close to the target solution. If the step-down algorithm is used, the iteration is stopped after 8 steps (see Fig. 4a) because the norm of the residual is below 10^{-8}.

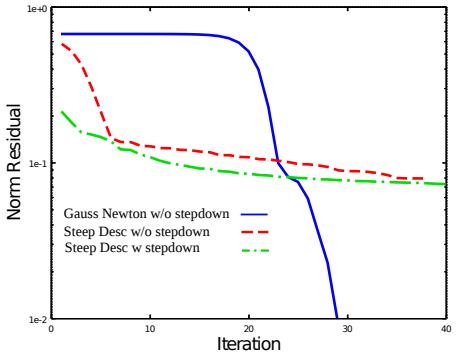

Figure 5: Norm of the residual for the Gauss-Newton method without step-down estimation of the starting value (line) and for the steepest descent method with (dash dotted) and without (dashed) step-down algorithm for the first 40 iterations (Color online).

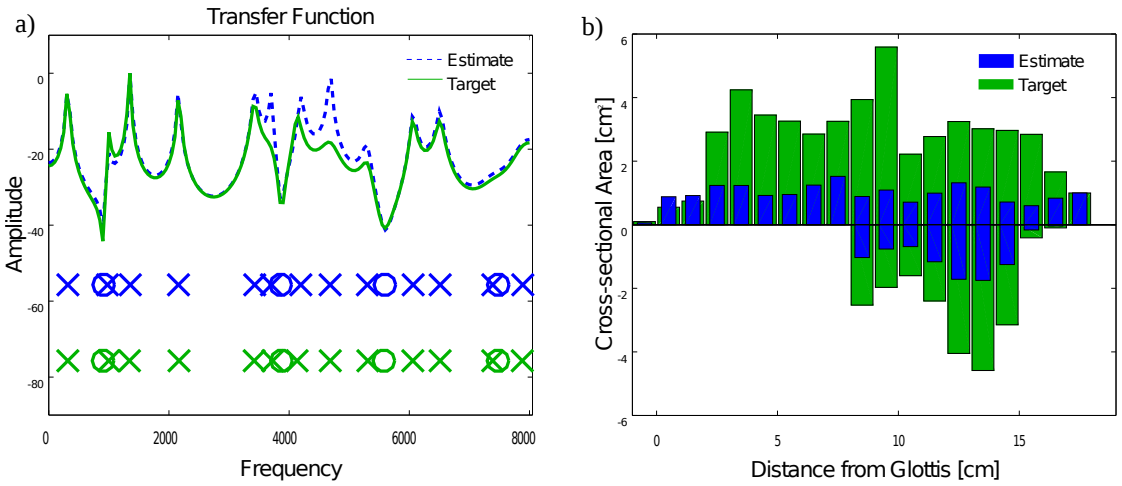

Figure 6: Difference between target and estimated transfer function and size of the cross-sectional areas after 40 iterations using the steepest descent method (Color online).

5.2 Practical Aspects

It is obvious that apart from the assumption about the length of the different tracts, the tuning of the cross-sectional areas for a three-tube-model from a given signal is dependent on a good estimation of the transfer function of the envelope of the spectrum by an ARMA model. In the literature several methods to estimate the transfer function of an ARMA model can be found (see Marelli and Balazs

(2010) for a small overview) and different methods may result in different pole-zero configurations. As an example, two different pole-zero models applied to the envelope of the spectrum for a nasal /m/ of a male speaker sampled at 16 kHz are compared in Fig. 7. It can be observed that especially in the lower frequency region, the position of the poles and zeros are different. The recursive weighted linear least squares (WLLS) algorithm resulted in two real valued poles, whereas the model presented in Marelli and Balazs (2010) had an additional conjugate complex pole at around ± 750 Hz. It is only natural to ask which of the two models provide appropriate poles and zeros.

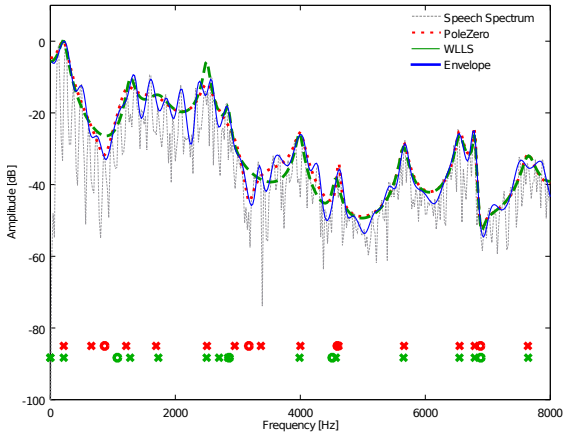

Figure 7: Spectrum, envelope and pole-zero estimations using two models given in Marelli and Balazs (2010). The 'x' and 'o' denote the frequencies of the poles and zeros of the transfer functions. In the upper line the poles and zeros for the model of Marelli and Balazs (2010) are given, in the lower line, the poles and zeros for the WLLS algorithm. (Color Online)

Most ARMA estimation methods have the drawback that they do not consider the background of voice production, i.e. the mechanics of the vocal tract. A combination with a tube-model may help to give additional insight by introducing constraints based on physical properties of the tube model. This shall be motivated by the following numerical experiment.

For both estimated transfer functions from the example above, a tube model was tuned to match the poles and zeros. It was assumed that pharyngeal, nasal, and oral tubes were represented by 8, 10, and 8 segments, respectively, and the vocal tract itself was assumed to be 17.5 cm long. The segment closest to the nostrils was assumed to have a cross-sectional area of 1 cm and for the speed of sound was set to $c = 350$ m/s. At the lip end of the oral tract and at the glottal end of the pharyngeal tract virtual segments were added to model boundary

conditions. While the algorithm converged for the poles and zeros provided by the model of Marelli and Balazs (2010), no convergence could be achieved in case of the WLLS-model. This was mainly due to the fact that some of the cross-sectional areas reached the lower or upper bound for the allowed range of [0.1,20]. In Fig. 8a the cross-sectional areas tuned to the poles and zeros calculated by the model of Marelli and Balazs (2010) (dark boxes) and the WLLS algorithm (bright boxes) are compared, in Fig. 8b the poles (marked by x,+,□ and ◇) and zeros (marked by circles) of the target and the tuned functions are depicted. In order to achieve the small error between the estimated and the target poles and zeros for the WLLS algorithm, the upper limit for the allowed cross-sectional areas was raised to 50 cm^2. It can observed that in the case of the WLLS algorithm the

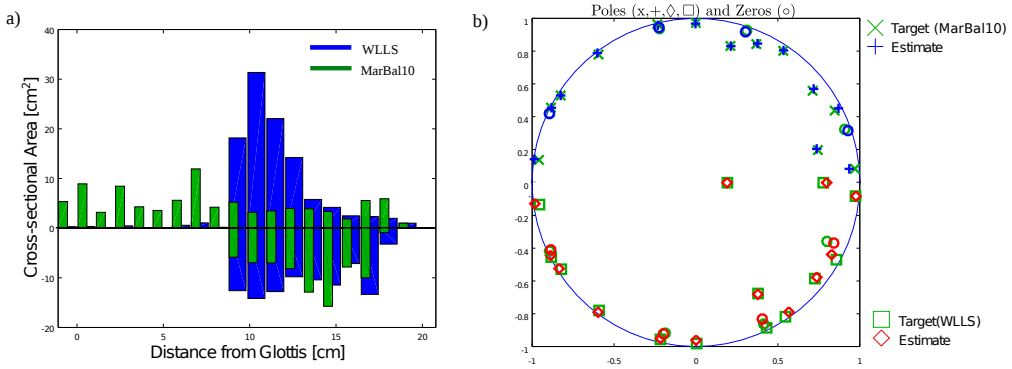

Figure 8: a) Tuned cross-sectional areas derived from poles and zeros calculated using the WLLS method (dark boxes) and the model presented in Marelli and Balazs (2010) (bright boxes). b) Poles and zeros for the method of Marelli and Balazs (2010) and the tuned tube-model (upper half circle) and for the WLLS method and the tuned tube-model (lower half circle). The upper limit for the cross-sectional areas was set to 50, the lower limit was kept at 10^{-1}. (Color Online).

reflection coefficient μ_0^p after the branch is very close to -1, resulting in unrealistic cross-sectional areas.

6 DISCUSSION

In this paper the concept of sensitivity functions for a tube model was expanded to a branched three-tube vocal tract model. The sensitivity functions were used to tune the parameters of a branched-tube model (i.e. the cross-sectional areas) from on a given rational transfer function using the steepest descent and the Gauss-Newton method.

Relating to previous work (Carre, 2004; Fant, 1970), and (Story, 2006) an energy-based sensitivity function for the frequencies of formants and anti-formants (poles and zeros) was derived. This function was shown to be essentially equivalent to a Jacobian-based approach, if the sensitivity is only defined with respect to the formant frequency. Despite this equivalence, the Jacobian-based approach has the additional advantage that it can be defined with respect to changes in frequency as well as bandwidth. In numerical experiments based on simulated data it was shown that the tuning of the area function solely based on formant and anti-formant frequencies does not yield the correct transfer function despite a near perfect fit in the frequencies. Including the bandwidth of the poles and zeros led to the correct results. Thus, the Jacobian-based approach using both real and complex part is better suited for the task of vocal tract tuning.

Contrary to Lim and Lee (1996) the cost function of the least squares approximation was not defined with respect to the polynomial coefficients of the rational transfer function, but with respect to the difference in the poles and zeros of the target and the tuned model in the z-domain. The additional effort for the numerical calculation of the roots of the two polynomials $H_1(z)$ and $H_2(z)$ is negligible, and the definition of the cost function via the poles and zeros has the advantage that constant factors in the transfer functions (for example the factors $z^{1/2}/(1-\mu)$ in Eq. (4)) can be neglected as they do not play a role for the calculation of the roots. Also, the sensitivity is directly based on the cross-sectional areas and not on the reflection coefficients, which makes it easier to directly impose constraints on the areas.

As for the choice of the iterative scheme the comparison of Gauss-Newton and steepest decent yielded what was to be expected namely a much faster convergence for the Gauss-Newton scheme but also a high dependence on the starting condition of the Gauss-Newton scheme, whereas the steepest descent method is quite robust with respect to starting values. Using a step-down algorithm for getting a good starting value for the cross-sectional areas of the oral tract provides satisfactory starting conditions for the Gauss-Newton scheme. Clearly, there are many other schemes and variants in the non-linear optimization literature that may yield improvements in convergence such as trust region approaches (Byrd *et al.*, 1987). As this topic was, however, not the main goal of this work only these two basic approaches were considered.

Although first results look quite promising, several points have to be addressed in future research. First, the proper choice of constraints for the cross-sectional areas is important and it was shown that fully unconstrained tuning may lead to unrealistic area functions. In this work, it is assumed that the cross-sectional areas are inside a predefined interval and that the area at the nostrils is set to 1. In related work on nasal estimation using a probabilistic scheme (Kasess

et al., 2012) it was shown that enforcing spatial smoothness by using constraints on the reflection coefficients (i.e. the coefficients should be close to zero) can reduce the within-subject variance of the estimates. Alternatively, by including tracking over time it could be enforced that the cross-sectional area between two time steps may only vary slightly, thus making sure that the changes over time are smooth.

Also, it is not clear yet, if two different sets of cross-sectional areas can produce the same transfer function. However, with the introduction of the additional reflection coefficient $|\mu_0^o| \neq 1$, the uniqueness of the cross-sectional areas of the *oral* tube is guaranteed. Additionally, when looking at the highest and the lowest coefficient of the polynomial $H_1(z) = \mu_0^0 \mu_L^p (1+\sigma')z^{-(L+M+N+1)} + \cdots + (\sigma'+1)$ (see Eq. (11) the boundary conditions at the tube ends are fixed. Nevertheless, the uniqueness problem will be a topic for future research. Additionally, by construction the three-tube model can only generate transfer functions that have certain properties, thus when tuning the tube model to arbitrary transfer functions, a unique (in the least square sense) three-tube configuration cannot be expected anymore without assuming additional constraints on possible cross-sectional areas.

Apart from setting proper constraints for the optimization process, modifications of the three-tube model are needed for practical situations. One drawback of using only three connected tubes as in Lim and Lee (1996) is that the zeros of the calculated transfer function are only dependent on the oral tract. Paranasal cavities, however, may also play a role in the acoustics of nasals, introducing further pole-zero pairs in the transfer function (for a detailed investigation on this topic using one dimensional models see e.g. Pruthi *et al.* (1997)). Tuning a more complex model, however, poses a problem as it is not clear how given zeros can be assigned to either sinus cavity or oral tract (Pruthi *et al.*, 1997) introducing some ambiguity in the modeling of actual data. A potential approach to address this issue is to introduce some physiologically motivated constraints on the paranasal cavities as was done in Kasess and Kreuzer (2013) where the maxillary sinus was approximated as a Helmholtz resonator (Dang and Honda, 1996) with different a-priori assumptions on resonance frequency, bandwidth, and coupling to the nasal cavity. As expected, the tuned results were highly dependent on the assumptions, illustrating the importance of proper constraints. This problem becomes even more pressing when the model is expanded to include a right and a left nasal tract, as the paranasal resonances would not appear explicitly as zeros in the spectrum (Pruthi *et al.*, 1997). An additional problem when expanding the model will be the choice of good starting values for the iteration because the zeros are now dependent on oral *and* nasal tracts.

It is obvious that the tube model is highly dependent on several parameters, which are, in general, not trivial to determine. First, a good estimation of physical

properties like length of the vocal tract or speed of sound inside the vocal tract need to be available. By construction the length of the segments ℓ and the sampling frequency are related by $F_s = c/(2\ell)$. If the sampling frequency is given by the speech signal different choices for c or ℓ therefore will have shift positions of the poles and zeros in the z-plane, and the tube-model will be fitted to a different set of poles and zeros. Second, the determination of the set of target poles and zeros of the transfer function is essential for a good estimation of the vocal tract parameters. In the literature several signal processing tools for estimating these sets can be found (for an overview see e.g. (Marelli and Balazs, 2010)) and different methods may lead to different poles and zeros. Alternatively estimating the area function based on the signal directly (cf. Kasess *et al.* (2012); Kasess and Kreuzer (2013)), e.g. based on the spectral envelope, would minimize the dependence of the vocal tract tuning on the ARMA model used for determining the target poles and zeros, however, a good model for the spectral envelope will still be necessary.

Acknowledgments

The authors would like to thank Michel Ta for determining the cross-sectional areas used in Fig. 1 from the graphs published in Story (1995) and the first studies with respect to the sensitivity of the three-tube model.

References

Badin, P., Makarov, I. S., and Sorokin, V. N. (**2005**). "Algorithm for calculating the cross-section areas of the vocal tract", Acoust. Phys. **51**, 52–58.

Byrd, R. H., Schnabel, R. B., and Shultz., G. A. (**1987**). "A trust region algorithm for nonlinearly constrained optimization", SIAM J. Numer. Anal. **24**, 1152–1170.

Carre, R. (**2004**). "From an acoustic tube to speech production", Speech Commun. **42**, 227–240.

Deng, H., Ward, R. K., Beddoes, M. P., and Hodgson, M. (**2005**). "Effects of glottal and lip boundary conditions on vocal-tract area function estimates from speech signals", *ICASSP '05*, 589–592.

Fant, G. (**1970**). *The Acoustic Theory of Speech Production, with calculation based on X-ray studies of Russion articulations, 2nd edition* (Mouton, Den Hague), 15–90.

Flanagan, J. I. (**1972**). "Voices of Men and Machines", J. Acoust. Soc. Am. **51**, 1375–1387.

Heinz, J. M. (**1967**). "Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues", Technical Report 8, KTH - Dept. for Speech, Music and Hearing.

Dang, J. and Honda, K. (**1996**). "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation", J. Acoust. Soc. Am. **100**, 3374–3383.

Kasess, C. H. and Kreuzer, W. (**2013**). "Estimation of multiple-branch vocal tract models: The influence of prior assumptions", in *Interspeech 2013*, 1663–1667 (Lyon, France).

Kasess, C. H., Kreuzer, W., Enzinger, E., and Kerschhofer-Puhalo, N. (**2012**). "Estimation of the vocal tract shape of nasals using a Bayesian scheme", in *Proceedings of Interspeech 2012*, 698–701.

Kelly, J. L and Lochbaum, C. C., (**1962**). "Speech Synthesis", in *Proc. 4th Int. Congr. Acoustics, Copenhagen*, 1–4.

Lim, I.-T. and Lee, B. (**1993**). "Lossless pole-zero modeling of speech signals", IEEE Trans. Speech Audio Processing **1**, 269–276.

Lim, I.-T. and Lee, B. (**1996**). "Lossy pole-zero modeling for speech signals", IEEE Trans. Speech Audio Processing **4**, 81–88.

Makarov, I. S. and Sorokin, V. N. (**2004**). "Resonances of a Branched Vocal Tract with Compliant Walls", Acoust. Phys. **50**, 323–330.

Marelli, D. and Balazs, P. (**2010**). "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis", IEEE Trans. Audio Speech Lang. Process. **18**, 237–248.

Markel, J. and Gray, A.H.J. (**1976**). *Linear Prediction of Speech* (Springer, Berlin), 95–98.

Mathur, S., Story, B. H., and Rodríguez, J. J. (**2006**). "Vocal-Tract Modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays", IEEE Audio Speech Lang. Process. **14**, 1754–1762.

Matsuzaki, H., Serrurier, A., Badin, P., and Motoki, K. (**2014**). "One-dimensional and three-dimensional propagation analyses of acoustic characteristics of Japanese and French vowel /a/ with nasal coupling", Acoust. Sci. & Tech. **35**, 35–41.

Mrayati, M., Carre, R., and Guerin, B. (**1988**). "Distinctive regions and modes: A new theory of speech production", Speech Communication **7**, 257–286.

Pruthi, T., Espy-Wilson, C. Y., and Story, B. H. (**1997**). "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data", J. Acoust. Soc. Am. **121**, 3858–3873.

Radolf, V. (**2007**). "Comparison of optimization methods for human vocal tract resonance properties tuning", Appl. Comput. Mech. **1**, 613–620.

Schnell, K. and Lacroix, A. (**2000**). "Parameter estimation for branched tube systems", in *KONVENS*, 127–130.

Schroeter, J. and Sondhi, M. M. (**1994**). "Techniques for estimating vocal-tract shapes from the speech signal", IEEE Speech Audio Process. **2**, 133–150.

Story, B. H. (**1995**). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract", Ph.D. thesis, University of Iowa, 192–207.

Story, B. H. (**2006**). "Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions", J. Acoust. Soc. Am. **119**, 715–718.

Story, B. H., Titze, I. R., and Hoffman, E. A. (**1996**). "Vocal tract area functions from magnetic resonance imaging", J. Acoust. Soc. Am. **100**, 537–554.

Wakita, H. (**1973**). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. Audio Electroacoust. **AU-21**, 417–427.