ANALELE UNIVERSITĂȚII
"EFTIMIE MURGU" REȘIȚA
ANUL XVIII, NR. 2, 2011, ISSN 1453 - 7397

Christian H. Kasess, Wolfgang Kreuzer

Vocal Tract Modeling: from Signal to Structure

Tube models are very popular for the computational modeling of speech production. For nasals and nasalized vowels a minimum of two tubes is necessary to accurately model the spectral components of the speech signal. Typically, such models are estimated applying a pole-zero model as a first step and then estimate the tubes cross-sectional areas based on this model. Here, we introduce a method that estimates the tube areas without the necessity to estimate a pole-zero filter model. The algorithm is based on a variational Bayesian scheme under Gaussian assumptions. Probabilistic priors are used to enforce smoothness of the tubes. The method is tested on simulated data and results show that under strong smoothness prior the algorithm converges more reliably than an unconstrained method.

Keywords: vocal tract, two tube model, pole-zero model, variational Bayes

1 Introduction

Computational models for speech production and speech analysis have been of research interest since the 1960s [1, 2, 8, 10]. In humans, the vocal tract (VT) is a nonuniform acoustic tube of about 17 cm length. One end is terminated by the vocal chords and the glottis whereas the opposite end is terminated by the lips. At the velum which acts similar to a trap door, the nasal tract, about 12 cm in length, is coupled to the vocal tract [2]. By changing the placement of velum, lips, jaw and tongue the cross section of the VT and thus the resonances of the tract are changed producing different sounds.

Most of today's speech analysis models are based on linear prediction coding (LPC [8]). It is assumed that the vocal tract acts as a linear filter that is driven by an impulse train (opening and closing of the glottis). It is further assumed that for (non nasalized) vowels an all-pole filter with transfer function $H(z) = 1/A(z)$, with $A(z) = \sum_{i=0}^{M} a_i z^{-1}$, $a_0 = 1$, $z = e^{i\omega T_s}$, where T_s is the sampling period, can be

used to model speech production. Given a speech signal, the coefficients a_i of the all-pole filter can be determined by applying the LPC to a given signal. In [10], Wakita showed that the LPC can be directly related to a simple (mechanical) model (see also Section 1.1), where the vocaltract is represented by a single tube and appropriate boundary conditions at lips or glottis are used.

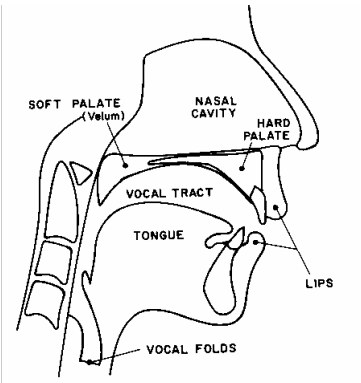

Figure 1. Schematic of the vocal tract. Taken from [8].

During the pronunciation of nasals and nasalized vowels, however, the velum opens and the additional resonances caused by the nasal tract influence the speech signal. The envelope of the speech spectrum has additional sinks (zeros) which cannot be described efficiently using an all-pole filter. In these cases a rational pole-zero filter with transfer function $H(z) = B(z)/A(z)$ is more appropriate for describing the spectrum of the signal. In order to determine the polynomial coefficients of $A(z)$ and $B(z)$ a nonlinear system of equations has to be solved and the construction of a solution algorithm is non trivial (see e.g. [7]). Additionally, whereas most existing pole-zero filter models are based purely on signal processing concepts, only a few are based on acoustical models ([5, 6, 9]), thus lacking a direct link to the mechanics of the vocal tract. Such acoustical models are also based on segmented tubes, however in order to allow for zeroes in the transfer function, at least one branching point is required. Therefore, contrary to the single tube case, there exists no direct relation between the two tube model and the pole-zero model (see Section 1.2). Still, the methods used to estimate two tube models are based on first estimating a pole-zero model and then fitting the vocal tract shape based on this model. Here, the aim is to introduce a all-in-one approach using a variational Bayesian scheme introduced in [3] that utilizes relatively mild assumptions about the vocal tract shape in order to constrain the solution of the non-linear system. Another advantage is the possibility to build hierarchical models where e.g. a second level that introduces an intra person statistic across multiple trials.

1.1 One Tube Model

The simplest and most commonly used model for the vocal tract can be constructed by representing the vocal tract as a straight segmented tube ([1, 10]). It is assumed that the transverse dimension of each section of the tube is small enough compared with the wavelength so that the wave propagating through the tube can be modelled by a plane wave. The second assumption is that heat loss and loss due to viscosity can be neglected. Using these assumptions, the wave motion inside the m-th segment can be described by the 1D-wave equation

$$\frac{\partial \phi_m(t,x)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 \phi_m(t,x)}{\partial t^2} = 0, \tag{1}$$

where c is the speed of sound. In each segment m with length Δl its solution is given as a forward and backward-going wave with volume velocity components $U_m^+(z)$ and $U_m^-(z)$ with $z = \exp(\mathrm{i}\omega 2\Delta l/c)$. Using the continuity conditions of pressure and flow at the boundary between two segments, it can be shown that the velocity flow components in two adjacent segments are connected by [10]

$$\begin{pmatrix} U_{m+1}^+(z) \\ U_{m+1}^-(z) \end{pmatrix} = \frac{z^{1/2}}{1-\mu_m}\begin{pmatrix} 1 & \mu_m \\ \mu_m z^{-1} & z^{-1} \end{pmatrix}\begin{pmatrix} U_m^+(z) \\ U_m^-(z) \end{pmatrix}, \tag{2}$$

where the reflection coefficient μ_m is given by the relation of the cross-section areas A_m of segment m and segment $m+1$:

$$\mu_m = \frac{A_{m+1} - A_m}{A_{m+1} + A_m}. \tag{3}$$

The transfer function of the vocal tract is then defined by the relation of the flow at the lips and the flow at the glottis. In [10], Wakita showed that the reflection coefficients can be calculated using the LPC-algorithm if the number of segments M are related to the sampling frequency F_s by $F_s = Mc/(2l)$ where l is the overall length of the vocal tract.

1.2. Two Tube Model

The drawback of the one tube model is that the contribution of the nasal tract on the speech signal is neglected. For vowels this poses, in general, no problem, but if nasals like /m/ or /n/ are to be considered, it is necessary to include an additional tract. In their model, Lim and Lee [5, 6] consider an acoustic tube consisting of two tracts. The model itself consists of three parts (see Fig. 2): A pharynx part between glottis and the velum (nasal-oral branch) consisting of L segments, an open nasal tract (M segments) and the oral tract (N segments) which is closed

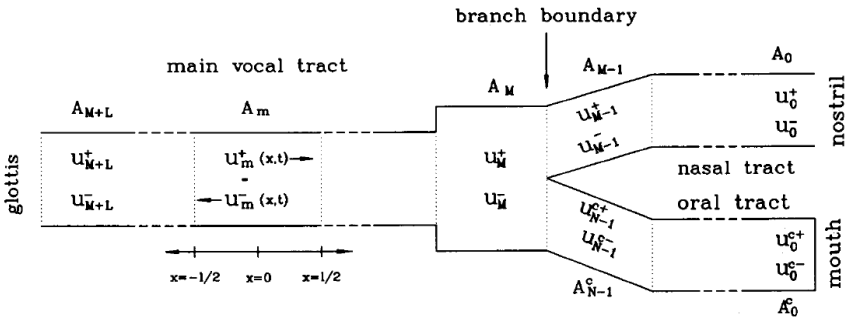

Figure 2. Generalized vocal tract model. Taken from [5].

at the lips. As in Section 1.1 the tracts are modeled using a segmented tube, and the same assumptions for the tube as in Section 1.1 are used. Again, using continuity conditions between the segments and at the coupling of the three branches, a rational transfer function $H(z) = B(z)/A(z)$ can be derived, where

$$A(z) = \begin{pmatrix} 1 & \mu_{M+L} \end{pmatrix} \begin{pmatrix} 1 & \mu_{M+L-1} \\ \mu_{M+L-1} z^{-1} & z^{-1} \end{pmatrix} \cdots \begin{pmatrix} 1 & \mu_M \\ \mu_M z^{-1} & z^{-1} \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{pmatrix} \begin{pmatrix} 1 & \mu_{M-1} \\ \mu_{M-1} z^{-1} & z^{-1} \end{pmatrix} \cdots \begin{pmatrix} 1 & \mu_1 \\ \mu_1 z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1} \end{pmatrix},$$

$$B(z) = C_{N-1}^{+}(z) + C_{N-1}^{-}(z),$$

with $C_{N-1}^{\pm}(z)$ given by the scaled flow at the back end of the oral tract

$$\begin{pmatrix} C_{N-1}^{+}(z) \\ C_{N-1}^{-}(z) \end{pmatrix} = \begin{pmatrix} 1 & \tilde{\mu}_{N-1} \\ \tilde{\mu}_{N-1} z^{-1} & z^{-1} \end{pmatrix} \cdots \begin{pmatrix} 1 & \tilde{\mu}_1 \\ \tilde{\mu}_1 z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mu}_0 z^{-1} \end{pmatrix} \qquad (4)$$

and

$$P(z) = C_{N-1}^{+} + (1-\sigma) C_{N-1}^{-}, \quad Q(z) = +\sigma C_{N-1}^{+},$$

$$R(z) = -\sigma C_{N-1}^{-}, \text{ and } S(z) = (1-\sigma) C_{N-1}^{+} + C_{N-1}^{-}.$$

Here, $\sigma = \tilde{A}_{N-1} / \left(A_{M-1} + \tilde{A}_{N-1} \right)$ defines the relation of cross section areas of oral and nasal tract areas at the velum. The $\tilde{\mu}_1, \ldots, \tilde{\mu}_{N-1}$ are the reflection coefficients for the oral part, $\mu_1 \ldots, \mu_{M-1}$ are the reflection coefficients for the nasal part and $\mu_M \ldots, \mu_{M+L}$ the reflection coefficients for the pharynx part. Lim and Lee added an additional damping term $\tilde{\mu}_0$ at the lips to include losses into the model [6].

From the above equations for $B(z)$ and $A(z)$ it is clear, that, unlike in the case

of the single tube model, no one to one mapping between the $M+L+2N$ polynomial coefficients and the $M+L+N+1$ reflection coefficients exists. Hence, estimation of the area function of a two tube model is not straight forward. Thus far, two different approaches have been suggested in the literature [6, 9]. The general scheme of both approaches is similar, determining first the polynomial coefficientes of the pole-zero or ARMA (autoregressive moving average) transfer function and using this information to solve for the tube model parameters. Both methods make use of the fact, that the numerator polynomial $B(z)$ can be directly mapped to the oral reflection coefficients $\tilde{\mu}_i$ using a step down algorithm [6, 8], similar to the one tube model. Based on these values, the nasal and pharyngal parameters are either estimated by minimizing the error with respect to the polynomial coefficients of the denominator [6] or by applying an inverse filter algorithm to the signal after dividing it by the numerator transfer function [9].
Both these methods give precedence to the numerator polynomial and thus assume, that the zeroes are modeled accurately by whatever ARMA estimation is used. Here, we suggest a different approach that estimates all coefficients simultaneously, thus avoiding this unequal weighting. This is highly non-trivial due to the complex relation between reflection coefficients and polynomial coefficients and the restrictions that apply to the reflection coefficients which must lie between -1 and 1. Hence, a Bayesian algorithm is used that includes probabilistic prior assumptions about the vocal trace, in this case about the smoothness. The estimation scheme introduce here is based on a general variational Bayesian scheme under Gaussian assumptions introduced in [3] and will be described below.

2 Ansatz

Similar to [7], the estimation scheme models the log of the transfer function $H(z)$ based on the log of the spectral envelope $G(\omega)$ of the recorded signal. The generative model for the log transfer function can be written as

$$\ln G(\omega) = y = f(\theta, \omega) + \varepsilon(\omega). \tag{5}$$

The function $f(\theta, \omega)$ incorporates the non-linear transformation from the reflection coefficients to the log transfer function (see Section 1.2) as well as a non-linear mapping from the i-th parameter θ_i to the i-th reflection coefficient μ_i defined as $\mu_i = \operatorname{erf}(\theta_i / \sqrt{2})$, with erf being the Gauss error function. The sigmoidal shape of this function ensures, that the reflection coefficients are restricted to the open interval $(-1,1)$. The nasal-oral coupling parameter σ is restricted to the interval $(0,1)$ by using the scaled and shifted error function. One additional parameter is added that models a scaling factor for the transfer function.

This parameter has to be positive, which is achieved by a log transformation. Therefore, the parameter vector is of dimension $M+N+L+2$.

The measurement error ε is assumed to be normally distributed with $\mathrm{N}(0,\Sigma(\lambda))$ with λ parameterizing the error covariance. The details of this parameterization will be given below. The normality assumption about the error yields a Gaussian likelihood function

$$p(y \mid \theta,\lambda) = \mathrm{N}(f(\theta),\Sigma), \tag{7}$$

where $\Sigma(\lambda)$ is now written as Σ for simplicity. The priors for θ and λ are also Gaussian i.e.

$$p(\theta) = \mathrm{N}(\eta_\theta,\Pi_\theta^{-1}) \quad \text{and} \quad p(\lambda) = \mathrm{N}(\eta_\lambda,\Pi_\lambda^{-1}). \tag{8}$$

Π_θ and Π_λ are the respective precision matrices (i.e. inverse covariance matrices). The parameters are estimated based on a variational Bayesian scheme derived in [3]. The variational distribution is given as $q(\theta,\lambda) = q(\theta)q(\lambda)$ with $q(\theta) = \mathrm{N}(\mu_\theta,\Sigma_\theta)$ and $q(\lambda) = \mathrm{N}(\mu_\lambda,\Sigma_\lambda)$ due to the normality assumption. The function to be maximized is the marginal likelihood (sometimes also called the lower bound estimate for the model evidence) and is given as

$$\begin{aligned} \mathrm{F} \;=\; & -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln\left|\Sigma^{-1}\right| - \frac{n}{2}\ln 2\pi - \\ & -\frac{1}{2}\varepsilon_\theta^T\Pi_\theta\varepsilon_\theta + \frac{1}{2}\ln\left|\Pi_\theta\right| + \frac{1}{2}\ln\left|\Sigma_\theta\right| - \\ & -\frac{1}{2}\varepsilon_\lambda^T\Pi_\lambda\varepsilon_\lambda + \frac{1}{2}\ln\left|\Pi_\lambda\right| + \frac{1}{2}\ln\left|\Sigma_\lambda\right|, \end{aligned} \tag{9}$$

with $\varepsilon_\theta = \mu_\theta - \eta_\theta$ and $\varepsilon_\lambda = \mu_\lambda - \eta_\lambda$. The scheme that maximizes this quantity performs two steps alternately. First, the vocal tract parameters are updated using the following set of equations:

$$\begin{aligned} \Sigma_\theta^{-1} &= J^T\Sigma^{-1}J + \Pi_\theta, \\ \left(\frac{\partial I}{\partial\theta}\right)_k &= -J^T\Sigma^{-1}\varepsilon - \Pi_\theta\varepsilon_\theta + \frac{1}{2}\mathrm{tr}\left(\Sigma_\lambda A^{(k)}\right), \\ \left(\frac{\partial^2 I}{\partial\theta^2}\right)_{k,l} &= -\left(\Sigma_\theta^{-1}\right)_{kl} + \frac{1}{2}\mathrm{tr}\left(\Sigma_\lambda B^{(k,l)}\right), \\ \Delta\mu_\theta &= -\left(\frac{\partial^2 I}{\partial\theta^2}\right)^{-1}\left(\frac{\partial I}{\partial\theta}\right), \end{aligned} \tag{10}$$

with I being the variational Energy, a quantity related to F (for details see [3]). After a single update, the error covariance parameter is updated, such that the marginal likelihood is maximized:

$$
\begin{aligned}
\Sigma^{-1}_{\lambda,i,j} &= \frac{1}{2} tr\left(P^{(i,j)}\left(\varepsilon\varepsilon^T - \Sigma\right) + P^{(i)}\Sigma P^{(j)}\Sigma\right) + \Pi_{\lambda,i,j}, \\
\left(\frac{\partial I}{\partial \lambda}\right)_i &= \frac{1}{2} tr\left(P^{(i)}\left(\varepsilon\varepsilon^T - \Sigma + J\Sigma_\theta J^T\right)\right) + \Pi_{\lambda,i,j}\varepsilon_\lambda, \\
\left(\frac{\partial^2 I}{\partial \lambda^2}\right)_{i,j} &= -\Sigma^{-1}_{\lambda,i,j} - \frac{1}{2}\mathrm{tr}\left(\Sigma_\theta J^T P^{(i,j)} J\right), \\
\Delta\mu_\lambda &= -\left(\frac{\partial^2 I}{\partial \lambda^2}\right)^{-1}\left(\frac{\partial I}{\partial \lambda}\right).
\end{aligned}
\tag{11}
$$

The matrices $P^{(i)}$, $P^{(i,j)}$, $A^{(k)}$, and $B^{(k,l)}$ are defined as

$$
P^{(i)} = \frac{\partial \Sigma^{-1}}{\partial \lambda_i} \quad P^{(i,j)} = \frac{\partial^2 \Sigma^{-1}}{\partial \lambda_i \partial \lambda_j} \quad A^{(k)}_{i,j} = J^T_{,k} P^{(i,j)} \varepsilon \quad B^{(k,l)}_{i,j} = J^T_{,k} P^{(i,j)} J_{,l}. \tag{12}
$$

$J_{,l}$ denotes the l-th colum of the Jacobi matrix $J = \partial f / \partial \mu$. If, after performing both steps, F has not increased, the step size of the first step is halved and the whole process is repeated. It is important to note that here second order derivatives of the function f are ignored.

Vocal tract priors

Informative priors for the reflection coefficients would require the knowledge of area function distributions for different phonemes and also probabilistic information of nasal tract areas. As those quantities are not well known in general, we use a very straight forward approach by requiring a certain smoothness of the vocal tract (see e.g. [4] for the area function of the LPC when both, glottal and lip losses are estimated). Such a constraint can be implemented using priors that are centered around zero, thereby preferring solutions with small reflection coefficients and thus a smooth vocal tract. The nasal-oral coupling coefficients σ is also centered around zero, resulting in equal nasal and oral coupling areas because of the non-linear transformation.

Noise priors and assumptions

In theory, the scheme described above allows to model the precision matrix Σ^{-1} as a non-linear function of a set of parameters λ_i. Here, we chose the simplest parameterization possible, resulting in a diagonal precision matrix given by a single parameter λ. Therefore,

$$
\Sigma^{-1} = \exp \lambda I_n \tag{13}
$$

with I_n being the unit matrix of the dimension of the number os samples n. The prior variance for λ was set to 10^5 which essentially implies a flat prior.

2.1 Evaluation

For the evaluation of the method we used two sets of simulated data based on the parameters given in [6]. Based on an assumed sampling frequency of 10kHz, $L = 4$, $M = 6$, and $N = 5$, thus resulting in a total of 15 reflection coefficients plus σ and the scaling factor.

Firstly, the transfer function was calculated from given reflection coefficients and used as input data for the model. Six different prior precisions values ($diag(\Pi_\theta^{-1}) = 0.05, 0.1, 0.2, 0.5, 1,$ and 2) were tested with a lower variance (higher precision or tighter priors) implying stronger a-priori assumptions. The prior for the scaling factor was set to 100. Since there is no noise in the data, the algorithm should ideally converge to the true values. In order to investigate the effect of different prior settings, the estimation was repeated for 200 different initial parameter settings. The starting conditions were varied by adding a Gaussian noise with zero mean and standard deviation 0.1 and 1 to the neutral starting position (i.e. all reflection coefficients are zero and σ equals 0.5). The 200 different starting conditions were fixed across conditions. The results were evaluated by looking at the area functions and the pole and zero positions as well as the deviation from the envelope over the different initial conditions. This analysis was done for two different assumed speech segment lengths of 20ms (resulting in 100 data points) and 40ms (200 data points) In addition we also used a standard Gauss-Newton (GN) scheme minimizing the sum of squares of the error without any prior assumptions in order to evaluate the reliability of the convergence depending on the initial condition. The function nlm implemented in R was used.

Secondly, we also generated a speech signal based on the pole-zero coefficients given in [6] using a impulse train of 100 Hz, sampling frequency of 10kHz and a speech segment length of 40ms. We applied the method used in [7] to calculate the spectral envelope that acts as the input to the VB algorithm. The model was estimated and compared to the true transfer function and parameters.

3 Results

Looking at the results starting from different initial conditions (Fig. 3 and Table 1) it is clear that for tighter priors the algorithm converged more reliably to the global minimum. For the highest precision, the algorithm converges to the true solution most of the time in the case of a not so strong deviation from the neutral setting. The less stringent the prior assumption, the less reliable the algorithm converges towards the global optimum. For larger variation of the initial conditions the percentages are less but the general tendency stays the same. As a reference, the Gauss-Newton scheme works relatively well for the low variation, but extremely poor for the high variation. Therefore, the results indicate that the use of smoothness priors results in the dependence on the starting value.

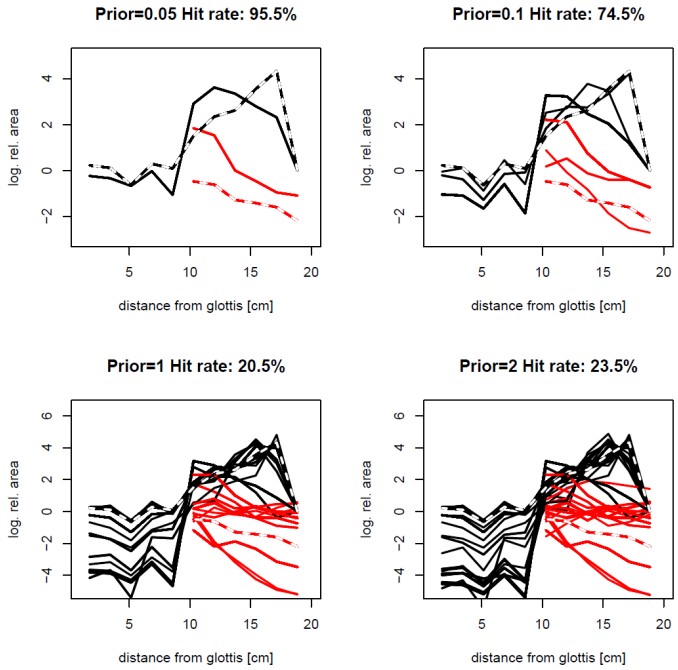

Figure 3. Illustration of the reconstructed logarithmic area functions for different prior variances. Black lines show nasal and pharyngal tract whereas red lines mark the oral tract. The dashed lines show the true values. On top the prior variance and percentage of convergences to the true values are given.

What can also be seen is that small deviations from the transfer function can yield large deviations from the area function. Looking at the pole-zero plot in Fig. 4 it is clear, that the major resonances are captured quite reliably, even for higher prior variances. The maximum error that occurs for the low initial variation from the neutral position is about 0.4 dB except in one case for the highest prior variance where the algorithm shows a poor fit. For large pertubation, the algorithm shows poor convergence for the highest three prior variances (a maximum of 12 times out of 200). The reason for the small changes in the transfer function is that, in this particular example, there are almost matched pole-zero pairs, whose position does not strongly affect the transfer function, but has a strong effect on the reflection coefficient and therefore the area function.
Using the synthetic speech signal sample, the results show, that the algorithm yields a good fit for all prior settings with the error being smaller for less tight priors (Fig. 5). The error reported is the mean absolute deviation in dB relative to the envelope.

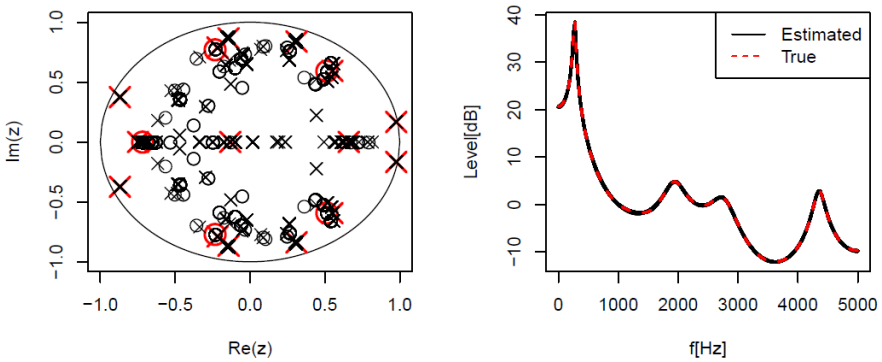

Figure 4. Shown are the poles (black x) and zeroes (black o) and the transfer functions (right panel) for a prior variance of 1. Red symbols (left panel) and red dotted line (right panel) mark the true values.

Table 1: Percentages of convergence to the true values as a function of the prior variance. Also reported are the values for the Gauss-Newton method.

Variation	Sample	Prior variance						GN
		0.05	0.1	0.2	0.5	1	2	
low	short	100	96.5	76	36.5	14.5	16.5	53
low	long	95.5	74.5	42	18	20.5	23.5	52
high	short	87	45.5	27.5	18.5	23	23.5	1.5
high	long	42.5	26.5	23.5	22	24	19	5

However, looking at the poles and zeroes of the original transfer function the tightest and most restrictive prior shows the zeroes to be closer to their true values than for the loose priors (Fig. 6). The estimation of the poles is comparable for all different settings but the area function shows stronger deviations for less tight priors.

4 Discussion

Two tube models are important models for representing nasals but also nasalized vowels. Here, we introduced a variational Bayesian approach in order to estimate the vocal tract area functions for a two tube model of nasal consonants. In contrast to existing procedures, the algorithm estimates all (oral, nasal, and pharyngal) reflection coefficients simultaneously and does not require a separate pole-zero estimation and is hence not dependent on which algorithm is used for the pole-zero estimation. The algorithm fits the log spectral envelope using zero mean Gaussian priors for the reflection coefficients, thereby preferring smooth vocal tract shapes.

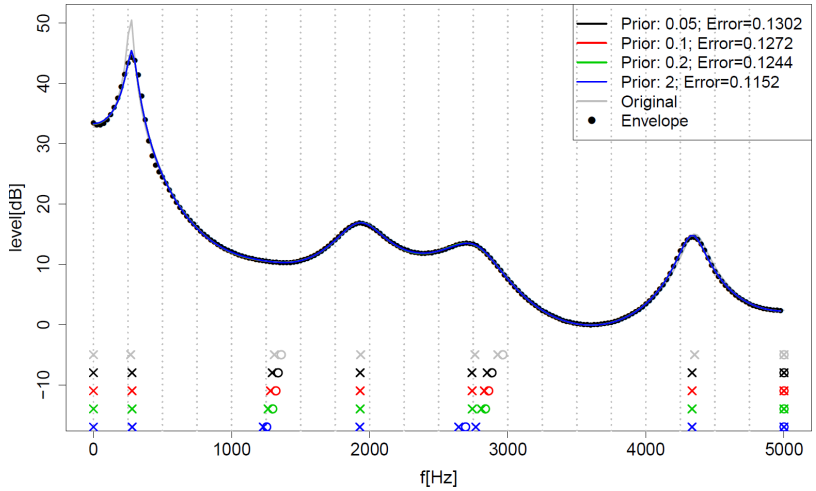

Figure 5. Shown are the estimated transfer functions for different prior variances as well as the true transfer function and the extracted envelope. Furthermore, the positions of the poles and zeroes are shown below the transfer functions.

Based on the simulated transfer function, the algorithm was shown to be more robust against varying initial conditions compared to an unconstrained non-linear solver, particularly for tighter priors. As expected, tighter priors consistently lead to less dependence on the starting condition. Furthermore, using a noise free synthesized speech signal, the algorithm showed good results with respect to the estimated envelope for all prior settings starting from a neutral position.

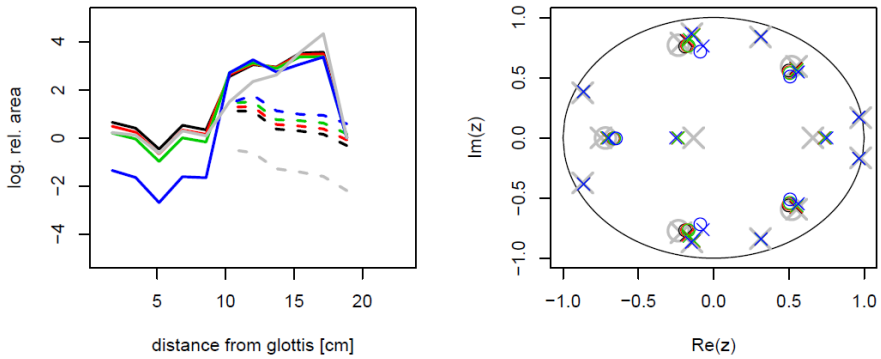

Figure 6. The left panel provides the estimated area functions for the oral part (dashed lines) and the pharyngal and nasal part (solid lines). The colors are coded as in Fig. 5. The right panel shows the poles and zeroes in the z-plane.

There are several issues that still need to be addressed. First, the effect of priors is, as is well known, sample size dependent. It is therefore unclear, what determines a suitable prior variance. Another important point is the statistical model for the error. Currently, the error is assumed to be Gaussian as well as indepently and identically distributed. Clearly, these are two strong assumptions, especially the independence, as adjacent points in a smoothed spectral envelope cannot be assumed to be independent. A possibility is to model the error in terms of an AR(n) process assuming a serial correlation across frequencies. This, however, requires a careful parameterization as the AR coefficientes underly certain stability constraints.

References

[1] G. Fant. *Acoustic theory of speech production, with calculation based on X-ray studies of Russian articulations.*. Mouton De Gruyter, 1060.

[2] J. Flanagan. *Speech analysis, synthesis, and perception*. Springer, Berlin, 1972.

[3] K. Friston and J. Mattout and N. Trujillo-Barreto and J. Ashburner and W. Penny. *Variational free energy and the Laplace approximation. Neuroimage,* 34:220--234, 2006.

[4] K. Kalgaonkar and M. Clements. *Vocal Tract and Area Function Estimation with both Lip and Glottal Losses*. INTERSPEECH, :550--553, 2007.

[5] I.-T. Lim and B.G. Lee. *Lossless Pole-Zero Modeling of Speech Signals.* IEEE Trans. Speech Audio Processing, 1(3):269--276, 1993.

[6] I.-T. Lim and B.G. Lee. *Lossy Pole-Zero Modeling for Speech Signals.* IEEE Trans. Speech Audio Processing, 4(2), 1996.

[7] D. Marelli and P. Balazs. *On Pole-Zero Model Estimation Methods Minimizing a Logarithmic Criterion for Speech Analysis*. IEEE, IEEE Trans. Audio Speech Lang. Process., 18(2):237--248, 2010.

[8] J.D. Markel and A.H. Gray, Jr. *Linear Prediction of Speech*. Springer, Berlin, 1976.

[9] K. Schnell. *Rohrmodelle des Sprechtraktes. Analyse, Parameter-schätzung und Syntheseexperimente*. PhD thesis, Universität Frankfurt, 2000.

[10] H. Wakita. *Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms*. IEEE Trans. on Aud. and Electroacoustics, AU-21(5):417--427, 1972.