ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Estimation of the vocal tract shape of nasals using a Bayesian scheme

Christian H. Kasess[1], *Wolfgang Kreuzer*[1], *Ewald Enzinger*[1,2], *Nadja Kerschhofer-Puhalo*[1]

[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[2]School of Elec. Eng. & Telecom., Univ. of New South Wales, Australia

{christian.kasess, wolfgang.kreuzer, ewald.enzinger, nadja.kerschhofer}@oeaw.ac.at

10.21437/Interspeech.2012-219

Abstract

For nasal stops and nasalized vowels, one-tube models offer only an inadequate representation. To model the spectral components of nasal speech signals, a minimum of two connected tubes is necessary. Typically, the estimation of branched-tube area functions is based on a pole-zero model. The present paper introduces a variational Bayesian scheme under Gaussian assumptions to estimate the tube areas directly from the log-spectrum of the speech signal. Probabilistic priors are used to enforce smoothness of the tubes. The method is tested on recorded tokens of /m/ from several speakers using different prior variances. Results show that mild smoothness assumptions yield the best results in terms of model error and marginal likelihood. Furthermore, while yielding comparable fits, the estimated reflection coefficients from the Bayesian scheme show less intra-subject variability between tokens than an unregularized non-linear solver.

Index Terms: vocal tract, estimation, nasal stops, Bayesian statistics

1. Introduction

Computational models for speech production and analysis have been of major research interest since the 1960s [1, 2, 3]. The most common model for speech coding is linear predictive coding (LPC, [2]). For LPC it is assumed that the vocal tract (VT) acts as a linear filter and that (non-nasalized) vowels can be modeled using an all-pole filter. For a given speech signal, the coefficients of the all-pole filter can be determined by applying the LPC to a given signal. Using appropriate boundary conditions at the glottis and the lips, the all-pole model can be directly related to a simple mechanical model where the VT is represented by a single tube [3].

During the production of nasal stops (e.g. /m/ or /n/) and nasalized vowels, however, the velum is lowered and the additional resonances caused by the nasal tract influence the speech signal. The envelopes of nasal spectra show additional sinks (zeros) and thus a pole-zero filter is a more efficient description of nasal signals than an all-pole filter. A number of algorithms has been proposed to solve a non-linear system of equations for the numerator and denominator polynomials (see e.g. [4]). Linking this pole-zero representation to an acoustical model requires the use of a branched-tube model where the nasal tract is added as an additional segmented tube. Estimating such an acoustical model may provide a link to the physiology of the speakers' VT. Unfortunately, the pole-zero representation has more degrees of freedom than the branched-tube model excluding the possibility of an exact mapping. A few methods have been suggested to estimate the VT area function based on a pole-zero model [5, 6].

In this study, the aim is to introduce an approach that estimates the VT model directly from the log spectral envelope without estimating a pole-zero model explicitly. A variational Bayesian scheme is applied that is based on the Laplace approximation [7], modified by using the unscented transform [8] for integration. This scheme utilizes relatively mild assumptions about the VT shape in order to constrain the solution of the non-linear system.

2. Methods

2.1. Two Tube Model

In their model for nasal stops, Lim and Lee [5] consider an acoustic tube model consisting of three parts: (1) a pharyngeal tract (L segments) between the glottis and the velum (nasal-oral branching point), (2) a nasal tract (M segments) open at the nostrils, and (3) a closed (non-radiating) oral tract (N segments). Each tract is modeled by a segmented tube. 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. These polynomials are related to the area function of the VT via the reflection coefficients (RCs)

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

where A_m is the cross-sectional area of the m-th segment starting at the nostrils (or lips for the oral part). The numerator polynomial $B(z)$ is of degree N and dependent on the oral RCs $\tilde{\mu}_0, \ldots, \tilde{\mu}_{N-1}$. The denominator polynomial $A(z)$ of degree $L + M + N$ is dependent on the oral RCs, the pharyngeal RCs $\mu_M, \ldots, \mu_{M+L}$, the nasal RCs $\mu_1, \ldots, \mu_{M-1}$ and the relation between the cross-

sectional areas of oral and nasal coupling sections at the velum $\nu = \tilde{A}_{N-1}/(A_{M-1} + \tilde{A}_{N-1})$ [5].

However, in general, no exact mapping from the $M + L + 2N$ polynomial coefficients to the $M + L + N + 1$ tube model parameters exists. Hence, estimation of the area function of a two tube model is not straight-forward.

Previous works suggested to determine first the pole-zero transfer function and then make use of the fact that the numerator polynomial $B(z)$ can be mapped exactly to the oral RCs $\tilde{\mu}_i$ using a step down algorithm [2, 5, 6]. Nasal and pharyngeal parameters are then either estimated using the polynomial coefficients of the denominator [5] or the residual signal [6]. Both methods assume that the zeros are modeled accurately by whatever pole-zero estimation method is used.

Here, we suggest a different approach that estimates all coefficients simultaneously, thus avoiding the use of a separate pole-zero estimation algorithm. This is highly non-trivial due to the complex relation between RCs and polynomial coefficients and the restrictions that apply to the RCs which must lie between -1 and 1. Hence, a Bayesian algorithm is used that includes probabilistic prior assumptions on the VT, in this case on smoothness. The estimation scheme introduced here is based on a general variational Bayesian scheme under Gaussian assumptions [7] and will be described below.

2.2. Estimation

The estimation scheme models the logarithm of the transfer function $H(z)$ based on the log of the spectral envelope $G(\omega)$ of the recorded signal [4]. The generative model for the log-envelope can be written as

$$y = \ln G(\omega) = f(\theta, \omega) + \epsilon(\omega). \tag{2}$$

The function $f(\theta, \omega)$ incorporates the non-linear transformation from the RCs to the log transfer function as well as a non-linear mapping from the i-th parameter θ_i to the i-th RC μ_i using a sigmoidal function (specifically the Gaussian error function) ensuring that the RCs are restricted to the open interval $(-1, 1)$. The nasal-oral coupling parameter ν is restricted to the interval $(0, 1)$. A scaling factor for the transfer function is also added. This parameter has to be positive, which is achieved by a log transformation. Therefore, the parameter vector θ is of dimension $M + N + L + 2$.

The measurement error ϵ is assumed to be normally distributed with $\mathcal{N}(0, \Sigma(\lambda))$ 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

$$p(y|\theta, \lambda, m) = \mathcal{N}(y|f(\theta), \Sigma), \tag{3}$$

where $\Sigma(\lambda)$ is now written as Σ for simplicity and m denotes the model assumptions, e.g. prior settings and VT structure. The priors for θ and λ are also Gaussian, i.e.

$$p(\theta) = \mathcal{N}\left(\theta|\eta_\theta, \Pi_\theta^{-1}\right) \quad \text{and} \quad p(\lambda) = \mathcal{N}\left(\lambda|\eta_\lambda, \Pi_\lambda^{-1}\right) \tag{4}$$

where m was dropped for simplicity. Π_θ and Π_λ are the respective precision matrices.

2.3. Variational Bayes

As the relation between spectral envelope and the model paramemters is non-linear even under normality assumptions no closed form solution exists and an approximation scheme has to be applied. In variational Bayes, the posterior distribution is partitioned into multiple independent sets of distributions. Here, the parameters are partitioned into the VT parameters θ and the error model parameter(s) λ. Hence, the variational distribution is given as $q(\theta, \lambda) = q(\theta)q(\lambda)$ with $q(\theta) = \mathcal{N}(\theta|\mu_\theta, \Sigma_\theta)$ and $q(\lambda) = \mathcal{N}(\lambda|\mu_\lambda, \Sigma_\lambda)$ due to the normality assumption. The posterior distribution $q(\theta)$ ($q(\lambda)$) is derived by integrating the log joint probability $\ln p(y, \theta, \lambda)$ with respect to $q(\lambda)$ ($q(\theta)$). Since both posteriors are assumed to be normal, the distribution is derived by finding the maximum and the 2nd-order derivative (approximated based on first order partial derivatives) and applying the Laplace approximation. Contrary to the original scheme [7], the integration is carried out applying the unscented transform [8], using an approximation up to 5th order. The parameter updates are carried out after each other until both updates converge twice in a row. Finally, the marginal likelihood $p(y|m)$ is calculated, again using the unscented transform. The algorithm was implemented in R [9].

2.4. Vocal tract priors

Informative priors for the RCs would require probabilistic information about the VT shape. As those quantities are not well known in general, we use a very straightforward approach by requiring a certain smoothness of the VT (a similar example is obtaining the area function using LPC when both glottal and lip losses are estimated [10]). Using Gaussian priors centered around zero, solutions with smaller RCs and hence smoother VTs are preferred. A smaller prior variance implies stronger regularization. The prior for the nasal-oral coupling coefficient ν is also centered around zero resulting in equal nasal and oral coupling areas due to the non-linear mapping.

2.5. Noise priors and assumptions

The error covariance Σ has an impact on the posterior VT parameter distribution and the marginal likelihood. As it is unknown and in general varies across utterances and speakers, the scheme described above models the inverse of the error covariance matrix Σ^{-1} as a non-linear function of a set of parameters λ_i which are to be estimated.

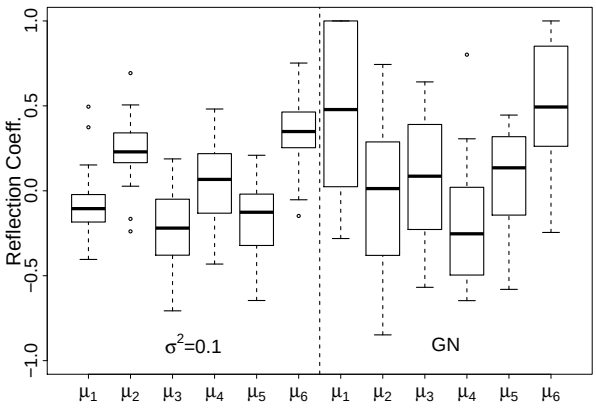

Figure 1: Nasal RCs using a prior variance of 0.1 (left) and unconstrained estimation (right) for speaker 1.

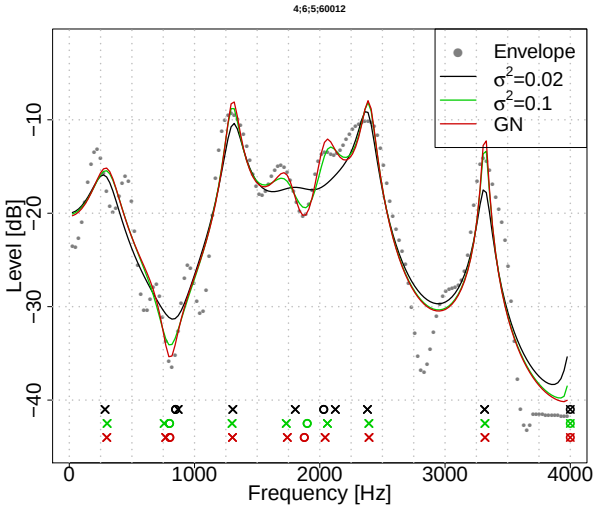

Figure 2: Envelope (gray dots) for /eme/ by speaker 1 as well as the corresponding estimated spectra and poles/zeros based on prior variances of 0.02 (black), 0.1 (green) and unregularized estimation (red).

Here, we chose the simplest parameterization possible, resulting in a diagonal precision matrix $\Sigma^{-1} = \exp(\lambda) I_n$ with I_n being the unit matrix of the dimension of the number of samples n. The prior variance for λ was set to 10^5 which essentially implies a flat prior.

2.6. Preemphasis

The effects of the glottal pulse were modeled with up to three real poles [6]. A single pole estimation using the Burg-algorithm was repeated up to three times unless the pole was on the negative real axis. Then the procedure was terminated earlier. The poles correct for the spectral tilt caused by the glottal pulse and nostril radiation.

2.7. Evaluation

For the evaluation of the method, we used 6 utterances of /m/ in 5 different vowel contexts. 3 untrained male speakers repeated the pseudoword 'ramadama' (remedeme, rimidimi, romodomo, rumudumu) 3 times. It was embedded within a German sentence to control for prosody. Tokens were manually segmented and a central segment of 40 ms duration was extracted for each token. 4 tokens were discarded as they were shorter than 40 ms. Speech samples were downsampled to a frequency of 8 kHz. A Hanning window was applied and envelopes were extracted using a peak detection and interpolation [4]. VT area functions were estimated using 4 different prior variances for the VT parameters ($\sigma^2 = 0.02, 0.05, 0.1$, and 1 with a diagonal matrix $\left(\Pi_\theta^{-1}\right)_{ij} = \sigma^2 \delta_{ij}$). VT parameters were set to $L = 4$, $M = 6$, and $N = 5$, thus resulting in a total of 15 RCs plus ν and the scaling factor. We compared the root mean square (RMS) and the log marginal likelihood of the different prior settings for the speech samples. In addition, we also used an unregularized standard Gauss-Newton (GN) scheme minimizing the sum of squares of the error.

3. Results

Table 1: *Mean RMS of the errors in dB for different subjects and prior settings.*

Subject	GN	prior variance			
		0.02	0.05	0.1	1
1	2.28	2.60	2.39	2.29	2.28
2	2.88	3.33	2.96	2.85	2.97
3	3.68	4.52	3.83	3.71	3.66
All	2.914	3.436	3.026	2.911	2.941

Comparing the different prior settings, $\sigma^2 = 0.1$ yielded the smallest overall error (Table 1) and also the highest marginal likelihood (not shown here). The table also shows that the unregularized GN yielded comparable results for the modeling error. However, looking at the distribution of the RCs, the Bayesian scheme shows much less variation as illustrated in Fig. 1 for subject 1 regarding the 6 nasal RCs. Interestingly, the Bayesian estimation and GN do not necessarily converge to a comparable solution. Still, spectral estimates are often very similar (Fig. 2). The strongly constrained estimation (tight priors) does follow a pattern similar to the estimation with the more loose priors. However, specific spectral features are not captured, probably due to the higher regularization. From Fig. 2 it is also clear that there is a number of spectral dips that cannot be modeled due to the limited number of oral coefficients and the fact that the paranasal cavities (sinuses) are not included in the present model. Fig. 3 shows the areas derived from the RC estimates for subjects 1 and 2 with $\sigma^2 = 0.1$. The area next to the

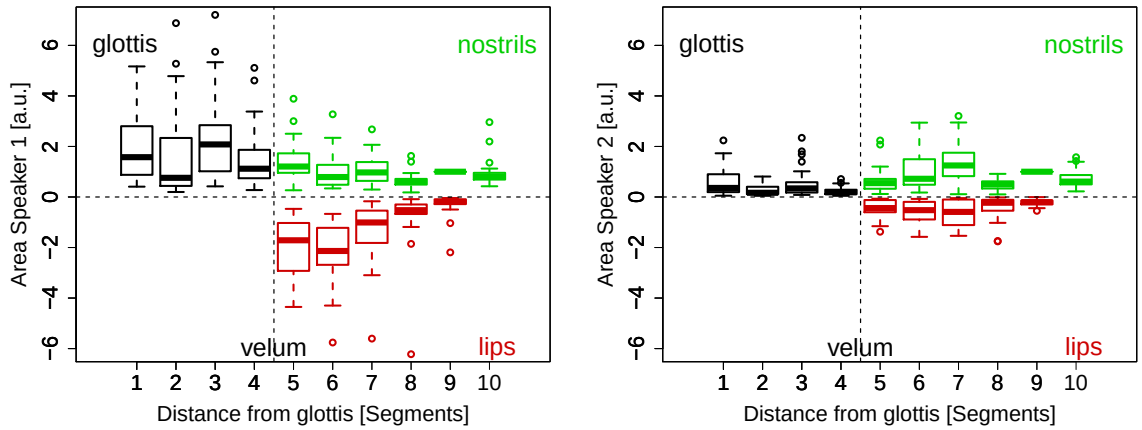

Figure 3: Distributions of area functions of 2 speakers for pharyngeal (black),nasal (green) and oral (red, negative) cavities.

nostrils was normalized to a value of 1. The scaling of the graph was chosen for better visibility. Thus, a few outliers that mainly affect the estimates of the oral cavity (2 of 30 utterances) and the pharyngeal cavity (3 of 30 utterances) could not be entirely displayed. The higher variance of these two cavities might be caused by the different vowel contexts, as the variance within a context appears to be smaller for many contexts. The nasal estimates show less variance and vary more strongly towards the velum which is to be expected.

4. Discussion

Branched-tube models are important for representing nasal stops as well as nasalized vowels. Here, we introduced a variational Bayesian approach based on the Laplace approximation and the unscented transform in order to estimate the VT area functions for a tube model of nasal stops. The aim was to estimate all (oral, nasal, and pharyngeal) RCs simultaneously and directly from the signal to avoid the dependence on the algorithm for pole-zero estimation. The algorithm fits the model to the log spectral envelope using zero mean Gaussian priors for the RCs, thereby preferring smooth VT shapes. The prior variance defines the strength of this regularization.

Application to recorded speech data yields in general good spectral fits. However, not all potential zeros can be fitted due to the limited complexity of the model. Low prior variances (strong regularization) result in poorer fits. Although, the use of an unregularized scheme that minimizes only the mean square error results in equally good fits, the variance of the RCs estimates is considerably larger across tokens. Concerning the selection of a proper prior, the marginal likelihood would be an indicator to identify the best model.

There are, however, still a number of open questions and issues. It is clear, that the use of this simple branched tube model is only partially adequate for modeling nasals, since paranasal cavities are not included. As the addition of such side branches increases the number of parameters, adequate priors may turn out to be vital in the estimation process. Other important issues concerning the VT model are the use of frequency-dependent glottal and nasal terminations as well as a realistic glottis model, thus increasing physical plausibility. A further improvement would be the use of priors deduced from anatomical measurements, e.g. through imaging methods.

To conclude, the probabilistic Bayesian approach shows promise for the estimation of area functions based on complex models of the vocal tract.

5. References

[1] G. Fant, *Acoustic theory of speech production, with calculation based on X-ray studies of Russian articulations.*, 2nd ed. The Hague: Mouton, 1970.

[2] J. Markel and A. Gray, Jr., *Linear Prediction of Speech*. Berlin: Springer, 1976.

[3] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-21, no. 5, pp. 417–427, 1972.

[4] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 237–248, 2010.

[5] I.-T. Lim and B. Lee, "Lossy pole-zero modeling for speech signals," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 2, 1996.

[6] K. Schnell, "Rohrmodelle des Sprechtraktes. Analyse, Parameterschätzung und Syntheseexperimente," Ph.D. dissertation, Universität Frankfurt, 2000.

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

[8] S. Julier and J. Uhlmann, "New extension of the Kalman filter to nonlinear systems," *Proceedings of the 1997 SPIE Conference on Signal Processing, Sensor Fusion, and Target Recognition*, vol. 3068, pp. 182–193, 1997.

[9] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[10] K. Kalgaonkar and M. Clements, "Vocal tract and area function estimation with both lip and glottal losses," *INTERSPEECH*, pp. 550–553, 2007.