INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Majorisation-minimisation based optimisation of the composite autoregressive system with application to glottal inverse filtering

Lauri Juvela[1,2], *Hirokazu Kameoka*[2,3], *Manu Airaksinen*[1], *Junichi Yamagishi*[4], *Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2]University of Tokyo, Japan
[3]Nippon Telegraph and Telephone Corporation, Japan
[4]National Institute of Informatics, Japan

{lauri.juvela,manu.airaksinen,paavo.alku}@aalto.fi
kameoka.hirokazu@lab.ntt.co.jp, jyamagis@nii.ac.jp

Abstract

The composite autoregressive system can be used to estimate a speech source-filter decomposition in a rigorous manner, thus having potential use in glottal inverse filtering. By introducing a suitable prior, spectral tilt can be introduced into the source component estimation to better correspond to human voice production. However, the current expectation-maximisation based composite autoregressive model optimisation leaves room for improvement in terms of speed. Inspired by majorisation-minimisation techniques used for nonnegative matrix factorisation, this work derives new update rules for the model, resulting in faster convergence compared to the original approach. Additionally, we present a new glottal inverse filtering method based on the composite autoregressive system and compare it with inverse filtering methods currently used in glottal excitation modelling for parametric speech synthesis. These initial results show that the proposed method performs comparatively well, sometimes outperforming the reference methods.
Index Terms: Composite autoregressive system, majorisation-minimisation, glottal inverse filtering

1. Introduction

The source-filter model for speech production states that a speech signal can be decomposed to a glottal source excitation and a vocal tract filter. However, the solution for the decomposition problem is not unique and evaluation of different solutions is difficult since glottal volume velocity ground truth is not available without intrusive measurements. Various glottal inverse filtering (GIF) techniques use some specific knowledge from the speech production theory to aid in the estimation of the source-filter decomposition. Techniques such as the iterative adaptive inverse filtering (IAIF) [1] model the spectral tilt present in the voice source with a low order all-pole filter and estimate it in cascade with the vocal tract filter. Other popular type of techniques, such as closed phase covariance analysis [2] and quasi-closed phase (QCP) [3] analysis, use weighted linear prediction to focus on the glottal closed phased of the speech cycle, attenuating the source contribution in the spectral estimate.

Glottal inverse filtering has applications in excitation modelling for parametric speech synthesis, as originally proposed in [4] using IAIF. Recently, the overall quality and flexibility for varying phonation types in synthesis has been improved by modelling the glottal excitation waveforms with Deep Neural Networks [5, 6]. Moreover, use of the more advanced QCP method resulted in significant improvement for female voice in [7]. However, the pre-requisite for all such modelling is reliable glottal inverse filtering, and since QCP requires accurate glottal closure instant (GCI) estimates, its performance is likely to degrade with a breathy voice or noisy speech. A review on other GIF methods and glottal excitation parameterisations used for method evaluation is given in [8].

The Composite Autoregressive system (CAR) [9] provides a convenient statistical approach for estimating the source-filter decomposition from speech by modelling the signal as a linear combination of various source-filter pair templates that are estimated from the data. However, the current expectation-maximisation optimisation method is time-consuming and can be improved upon. Optimising the CAR model is equivalent to minimising the reconstruction error of the autoregressive source-filter model, i.e., the Itakura-Saito divergence [10]. This, with the structure of the CAR system, connects the model to non-negative matrix factorisation (NMF) with I-S divergence [11], further motivating us to apply a majorisation-minimisation based approach similar to I-S NMF presented in [12, 13].

Additionally, this paper proposes a glottal inverse filtering technique based on the CAR system. The conventional frame-by-frame analysis sometimes suffers from noisy data or the sparse harmonics in high-pitched voices. As the vocal tract remains relatively stationary over several frames, we hypothesise that optimisation over a longer time period gives benefits over the conventional approach. To add flexibility to the CAR optimisation process, we introduce a generalised Gamma distribution prior into the model, which induces sparsity in the representation and enables adding frequency dependent prior information for the source model. We set the source spectral priors using the Liljencrants-Fant (LF) [14] glottal model in order to embed spectral tilt in the estimated source templates. The resulting inverse filtering method is evaluated by using synthetic speech data and calculating various objective error measures based on glottal flow parameterisations [8]. These measures are then compared with those of the established GIF methods IAIF and QCP currently used for in glottal vocoding.

The paper is structured as follows: in the theoretical section 2, we overview the composite autoregressive system, deriving new update rules (section 2.1) and a glottal inverse filtering method (section 2.2), and finally explaining the role of priors (section 2.3). In the experimental section, we define our test signals (section 3.1), examine the convergence rate of the proposed optimisation algorithm (section 3.2) and compare the proposed GIF method with established methods (section 3.3).

http://dx.doi.org/10.21437/Interspeech.2016-735

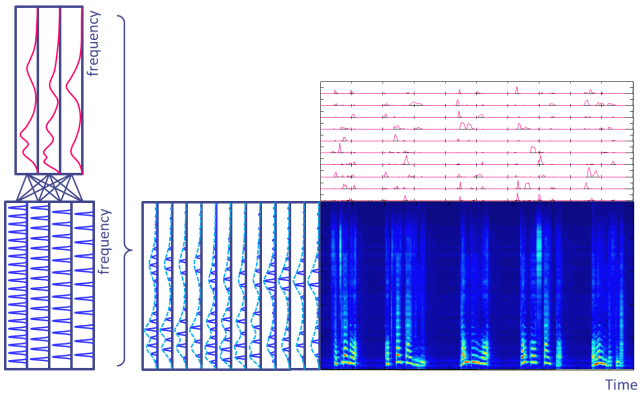

Figure 1: Figure illustrating the concept of the composite autoregressive system: signal spectrogram is modelled as a linear combination of $I \times J$ source-filter pairs compounded from I source and J filter templates.

2. Composite autoregressive system

The composite autoregressive system models the signal as a weighted sum of source–filter pair spectra compounded from I source and J filter templates. Each of these components at time instant n has the distribution $\boldsymbol{X}_n^{i,j} \sim \mathcal{N}_{\mathbb{C}}(0, \boldsymbol{\Lambda}_n^{i,j})$, where $\boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}(\lambda_{1,n}^{i,j}, \cdots, \lambda_{K,n}^{i,j})$ and

$$\lambda_{k,n}^{i,j} = \frac{U_n^{i,j} F_k^i}{|A^j(e^{\mathrm{j}2\pi k/K})|^2} = U_n^{i,j} F_k^i H_k^j, \tag{1}$$

$$A^j(z) = 1 - a_1^j z^{-1} - \cdots - a_P^j z^{-P}, \tag{2}$$

where F_k^i corresponds to the vocal source spectrum at k:th frequency bin, $H_k^j = 1/|A^j(z)|^2$ is an all-pole filter, and $U_n^{i,j}$ is a non-negative weighting coefficient. The observed complex spectrogram $\boldsymbol{Y}_n$ is then given by the sum of these components

$$\boldsymbol{Y}_n = \sum_{i,j} \boldsymbol{X}_n^{i,j} \sim \mathcal{N}(0, \boldsymbol{\Phi}_n), \tag{3}$$

$$\boldsymbol{\Phi}_n = \sum_{i,j} \boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}(\phi_{1,n}, \ldots, \phi_{K,n}), \tag{4}$$

where $\phi_{k,n} = \sum_{i,j} \lambda_{k,n}^{i,j}$. The model structure is illustrated in Fig. 1. The original EM-based optimisation algorithm is presented in [9], while this paper introduces a new majorisation-minimisation algorithm presented in section 2.1. By using either optimisation approach, maximising the probability density function of Y_n with respect to $\phi_{k,n}$ amounts to minimising the Itakura-Saito divergence, which is given up to constant by

$$D_{\mathrm{IS}}(\boldsymbol{Y}, \boldsymbol{\Phi}) = \sum_{k,n} \left(\frac{Y_{k,n}}{\phi_{k,n}} + \log(\phi_{k,n}) \right). \tag{5}$$

2.1. Auxiliary function approach

In this section we derive a majorisation-minimisation optimisation algorithm for the composite autoregressive system using the auxiliary function approach. This is done by minimising the Itakura-Saito divergence between the signal and model spectrograms. The procedure is similar to deriving the multiplicative update rules for NMF as done in [12, 13]. In addition, we apply a generalised Gamma prior to the activations and source spectral components to better control the behaviour of these parts of the model. Under certain restrictions this results in closed form update rules satisfying the non-negativity requirement for activations $U_n^{i,j}$ and source spectral templates F_k^i. Minimising the I-S divergence D_{IS} directly is difficult, but local optima can be found iteratively using the auxiliary function approach: first construct an upper bound auxiliary function G_{IS}, where the majorising function is equal to D_{IS} at exactly one point and can be minimised in closed form. After this, minimise the auxiliary function and update the model parameters, which is guaranteed to decrease the original objective function D_{IS} for the next iteration.

The generalised Gamma distribution (with parameters η, d, p) [15] provides a flexible way to apply various desirable properties to the model, as described in section 2.3. Here we show the derivation for the activations $U_n^{i,j}$ and note that the source spectrum templates F_k^i can be derived in a similar manner. The generalised Gamma prior for U is given by

$$p(U) \propto U^{d-1} \exp\left(-\frac{U^p}{\eta}\right), \tag{6}$$

$$\log(p(U)) = (d-1)\log(U) - \frac{U^p}{\eta} + \text{const.} \tag{7}$$

Since U^p is concave at $p < 1$, an upper bound is given by the tangent at point V, where V is an auxiliary variable.

$$U^p \leq pV^{p-1}(U - V) + V^p + \text{const.} \tag{8}$$

Then the prior log-probability is bounded up to constant by

$$\log(p(U)) \leq (d-1)\log(U) - \frac{1}{\eta} pV^{p-1}(U - V) + \frac{1}{\eta} V^p \tag{9}$$

Disregarding the constant terms, the posterior upper bound function is given by subtracting the prior upper bound from the likelihood-based G_{IS} presented in [12, 13]:

$$\begin{aligned} G_{\mathrm{IS}} = &\sum_{k,n} \left[\sum_{i,j} \frac{Y_{k,n}(\xi_{k,n}^{i,j})^2}{F_k^i U_n^{i,j} H_k^j} + \sum_{i,j} \frac{F_k^i U_n^{i,j} H_k^j}{\alpha_{k,n}} \right] \\ &- \sum_{n,i,j} \Bigg[(d-1)\log(U_n^{i,j}) \\ &- \frac{1}{\eta} p (V_n^{i,j})^{p-1} (U_n^{i,j} - V_n^{i,j}) + \frac{1}{\eta} (V_n^{i,j})^p \Bigg], \end{aligned} \tag{10}$$

where $(\alpha_{k,n}, \xi_{k,n}^{i,j}, V_n^{i,j})$ are the auxiliary variables and the equality for the upper bound holds only when

$$\xi_{k,n}^{i,j} = \frac{U_n^{i,j} F_k^i H_k^j}{\phi_{k,n}}, \tag{11}$$

$$\alpha_{k,n} = \phi_{k,n}\,, \tag{12}$$

$$V_n^{i,j} = U_n^{i,j}\,. \tag{13}$$

Differentiating this with respect to $U_n^{i,j}$ yields a full second degree equation, from which we choose the solution

$$U_n^{i,j} \leftarrow \frac{b_U + \sqrt{b_U^2 + 4 a_U c_U}}{2 a_U}, \tag{14}$$

$$a_U = \sum_k \frac{F_k^i H_k^j}{\phi_{k,n}} + \frac{p}{\eta} (U_n^{i,j})^{p-1}, \tag{15}$$

$$b_U = K(d-1), \tag{16}$$

$$c_U = \sum_k \frac{Y_{k,n} F_k^i H_k^j (U_n^{i,j})^2}{\phi_{k,n}^2}, \tag{17}$$

where positivity is guaranteed by constraining $d \geq 1$. Similarly, we obtain the update rule for F_k^i

$$F_k^i \leftarrow \frac{b_H + \sqrt{b_H^2 + 4a_H c_H}}{2a_H}, \tag{18}$$

$$a_H = \sum_{n,j} \frac{U_n^{i,j} H_k^j}{\phi_{k,n}} + \frac{p}{\eta_{k,i}} (F_k^i)^{p-1}, \tag{19}$$

$$b_H = NJ(d-1), \tag{20}$$

$$c_H = \sum_{n,j} \frac{Y_{k,n} U_n^{i,j} H_k^j (F_k^i)^2}{\phi_{k,n}^2}, \tag{21}$$

where $\eta_{k,i}$ is frequency dependent to accommodate for the LF prior. Here also p and d are chosen differently from $U_n^{i,j}$ prior as the priors have different desirable properties. Using a uniform prior simplifies the update rule to one resembling the multiplicative NMF update rule [12, 13] with IS-divergence:

$$F_k^i \leftarrow F_k^i \sqrt{\frac{\sum_{n,j} \frac{Y_{k,n} U_n^{i,j} H_k^j}{\phi_{k,n}^2}}{\sum_{n,j} \frac{U_n^{i,j} H_k^j}{\phi_{k,n}}}}. \tag{22}$$

Finally for the filter templates, we follow a similar procedure and enforce the all-pole constraint by solving the filter coefficients from the normal equations

$$\begin{bmatrix} r_0^j & r_1^j & \dots & r_{P-1}^j \\ r_1^j & r_0^j & & r_{P-2}^j \\ \vdots & & \ddots & \vdots \\ r_{P-1}^j & r_{P-2}^j & \dots & r_0^j \end{bmatrix} \begin{bmatrix} a_1^j \\ a_2^j \\ \vdots \\ a_P^j \end{bmatrix} = \begin{bmatrix} r_1^j \\ r_2^j \\ \vdots \\ r_P^j \end{bmatrix}, \tag{23}$$

where

$$r_{1,\dots,P}^j = \mathrm{DFT}^{-1} \left\{ H_k^j \sqrt{\frac{\sum_{n,i} \frac{Y_{k,n} U_n^{i,j} F_k^i}{\phi_{k,n}^2}}{\sum_{n,i} \frac{U_n^{i,j} F_k^i}{\phi_{k,n}}}} \right\}. \tag{24}$$

Despite the filter convergence is not directly guaranteed with this update rule, we observe convergence in practice at a faster rate than with the original EM algorithm (see Fig. 3).

2.2. Inverse filtering based on the CAR model

The expected value of individual model components is given by

$$\mathbb{E}\left[\hat{Y}_{k,n}^{i,j}\right] = Y_{k,n} \frac{\lambda_{k,n}^{i,j}}{\phi_{k,n}} = Y_{k,n} \frac{U_n^{i,j} F_k^i H_k^j}{\phi_{k,n}}, \tag{25}$$

where $Y_{k,n} \in \mathbb{C}$. The source components $\hat{S}_{k,n}^{i,j}$ can be estimated by removing the all-pole filter contribution H_k^j and the source estimate is the obtained by summing the source components:

$$\hat{S}_{k,n} = \sum_{i,j} \hat{S}_{k,n}^{i,j} = \sum_{i,j} Y_{k,n} \frac{U_n^{i,j} F_k^i}{\phi_{k,n}}. \tag{26}$$

The reference methods operate in a framework where the source estimate is obtained in time domain by inverting the estimated vocal tract filter. In order to conform to this, the obtained source spectrogram is cancelled from the signal spectrogram to obtain vocal tract spectrum estimates at each frame, and finally all-pole filters are fit onto these before inverse filtering. Alternatively the complex valued source spectrogram can be converted directly to time domain with inverse short-time Fourier transform.

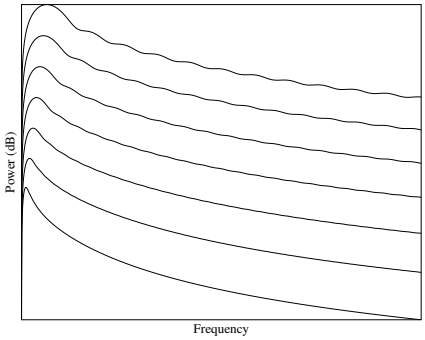

Figure 2: Examples of LF pulse spectra used for the priors of F_k^i. Using different LF priors for the individual source components allows varying degrees of spectral tilt.

2.3. Setting the prior parameters

The generalised Gamma parameters can be chosen to achieve different goals for the activations and source templates. For the activations $U_n^{i,j}$, the main requirement for the prior is to induce sparsity. The power parameter p can be set anywhere between zero and one, with smaller values providing more sparsity and $p = 1$ corresponding to the inverse Gamma prior used in [9]. In this work we used $p = 0.7$.

For the source templates F_k^i, the desired properties for the prior are slightly different. First, to encourage spectral tilt, we set η_k^i, which is proportional to the mean of the distribution, to correspond to a LF spectral prior in the update equations (18) and (19). The spectra were created by sampling the LF parameter space evenly and using the expression for LF spectral envelope given in [16]. Examples of such spectra are shown in Fig. 2. Second, to avoid potential numerical problems occurring with very small values of F_k^i, we set d slightly above one, since the d-dependent term in Eq. (10) effectively acts as a logarithmic barrier function, which drives the error towards infinity as F_k^i tends to zero. This type of barrier function has been used explicitly for positive matrix factorisation in [17].

3. Experiments

3.1. Test signals

For evaluating glottal inverse filtering methods, a test signal with the ground truth for source and filter should be available, effectively forcing the use of synthetic data. Instead of using the conventional fixed filters and fundamental frequency (f_0) for test data, we attempt to create more realistic test signals by using simple analysis-synthesis framework: first extract the LPC spectral envelope and f_0 from real continuous speech, then use the pitch information to create an LF model based excitation signal, and finally combine these to synthesise the test signals. To achieve this, we used a modified version of the GlottHMM vocoder [4], where the GIF was replaced with plain LPC analysis, and the glottal excitation pulses were created using the anti-aliased LF model presented in [18]. This excitation signal was considered as the ground truth for the vocal source in the GIF evaluation. The speech database used for the experiment consists of Finnish sustained vowel utterances in neutral, breathy, and pressed phonation styles, totalling 9.5 minutes in 792 utterances. In analysis-synthesis, the LF parameters were chosen from three sets of parameters in accordance with the phonation style, using parameters from the software presented in [18]. A noise component was added to the voiced excitation in based on

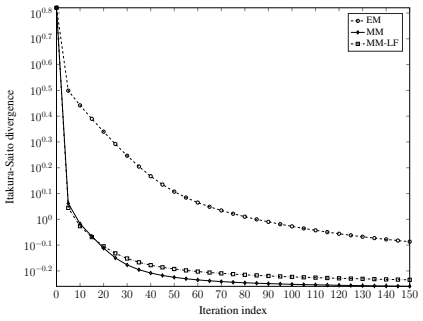

Figure 3: *The average Itakura-Saito divergence as a function of iteration index illustrates the convergence rates of the different optimisation algorithms. Majorisation-minimisation based methods (MM) and (MM-LF) tend to converge faster than the baseline expectation maximisation based method (EM).*

the harmonic-to-noise ratio similarly to [4].

3.2. Comparison of optimisation methods

The convergence rates of three CAR system optimisation methods were examined briefly: first, the original EM-based method (EM) [9], second, a MM-based method with uniform source prior (MM) using Eq. (22), and third, a MM-based method with LF based source prior (MM-LF). Since optimising the CAR system amounts to minimising the I-S divergence, we examine the divergence as a function of the iteration index to compare the convergence rates. For this experiment, we used the sustained vowel data described in section 3.1. The time-normalised average I-S divergence curves are presented in Fig. 3. Both MM and MM-LF converge faster than the original EM-based method, with MM reaching slightly lower error at convergence.

3.3. Glottal inverse filtering

The performance of the proposed GIF method was compared with the established methods IAIF [1] and QCP [3]. The objective evaluation of the GIF was carried out by comparing the errors in the following glottal flow parameterisations: the mean squared error (MSE) between the ground truth and estimated glottal flow derivatives; H1H2 [19], which measures the difference between the first and second harmonics in dB; the Harmonic Richness Factor (HRF) [20], which characterises the amount spectral tilt; the Normalised Amplitude Quotient (NAQ) [21], which describes the relative length of the glottal closing phase; and Quasi-Open Quotient (QOQ) [22], measuring the relative length of the glottal open phase. Glottal inverse filtering was performed for the test sets described in section 3.1 and the waveforms were normalised by first scaling the glottal flow derivative to match ground truth energy and then shifting the integrated glottal flow minimum to zero. After this, the error measures were evaluated in voiced frames.

The reference methods were used with their default settings for 16 kHz sample rate, and SEDREAMS [23] GCI detection was used for QCP. The CAR optimisation was run for 150 iterations for both proposed inverse filtering methods: CAR-MM without source prior, and CAR-MM-LF with LF-based prior. Table 1 presents the average errors, grouped by the phonation style classes: neutral (top), breathy (middle), and pressed (bottom). The presented error measures are mean absolute error for MSE and H1H2 (dB), and mean relative absolute error for HRF, NAQ and QOQ, with smaller error measures signifying better performance. The results show that the proposed methods both perform reasonably well for neutral and pressed phonation, and tend to outperform the reference methods for breathy phonation.

Table 1: *Objective error measures for GIF: MSE and H1H1 are given as average absolute errors and the rest as average relative errors. Smaller error measures indicate better performance. I source templates and J filter templates were used for CAR, and a total number of N voiced frames contribute to the error measure.*

Neutral phonation ($I = 5, J = 3, N = 26593$)

	MSE	H1H2	HRF	NAQ	QOQ
IAIF	**6.31e-04**	2.05	0.36	**0.13**	**0.18**
QCP	7.92e-04	2.03	0.79	0.14	0.28
CAR-MM	8.35e-04	1.76	**0.28**	0.23	0.24
CAR-MM-LF	8.18e-04	**1.74**	0.49	0.16	0.26

Breathy phonation ($I = 5, J = 3, N = 32281$)

	MSE	H1H2	HRF	NAQ	QOQ
IAIF	4.95e-04	4.91	0.39	**0.07**	0.11
QCP	9.69e-04	**2.44**	0.71	0.17	0.24
CAR-MM	**4.82e-04**	3.36	**0.37**	**0.07**	**0.10**
CAR-MM-LF	5.36e-04	4.06	0.43	**0.07**	0.12

Pressed phonation ($I = 5, J = 3, N = 26774$)

	MSE	H1H2	HRF	NAQ	QOQ
IAIF	9.27e-04	1.75	0.72	**0.14**	**0.20**
QCP	8.51e-04	2.03	1.23	0.20	0.30
CAR-MM	8.26e-04	**1.68**	**0.47**	0.18	0.24
CAR-MM-LF	**8.18e-04**	1.74	0.49	0.16	0.26

4. Discussion and Conclusion

This paper applied the composite autoregressive system to glottal inverse filtering analysis of speech production. The optimisation process was improved with an auxiliary function -based method for the composite autoregressive system, leading to a faster convergence rate. Generalised gamma prior was added to the optimisation to add sparsity to the component activations and spectral tilt to the source spectral templates.

The proposed glottal inverse filtering technique based on the CAR system was studied experimentally with synthetic data derived from natural sustained vowel utterances using the LF model for glottal excitation. The method was compared with existing inverse filtering techniques IAIF and QCP using objective error measures based on glottal parameterisations. For breathy phonation style, the proposed method tends to outperform the reference methods, while also performing reasonably well for other test data. These initial experiments show promise for application in glottal vocoding, but the proposed method is not as such applicable to data with variable linguistic content. Instead of examining the full utterance at once, the proposed method should be modified to work on relatively stationary segments of the signal and compound these results afterwards.

Future work includes tuning the model parameters and finding the optimal amount of source and filter templates to be used for inverse filtering. The proposed inverse filtering method should also be refined by decoupling the all-pole filter gains from the activations. Additionally, the source prior used for CAR-MM-LF in this work is rather soft and did not produce much difference to CAR-MM. Setting sharp harmonic source priors instead to enforce source spectral tilt should be studied.

5. Acknowledgements

This work was supported by the Academy of Finland (proj. no. 256961 and 284671), and the European Union TEAM-MUNDUS scholarship (TEAM1400081).

6. References

[1] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992, Eurospeech '91.

[2] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[3] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[4] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[5] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. of EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 2290–2294.

[6] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. of Interspeech*, Singapore, Sep. 2014, pp. 1969–1973.

[7] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *"Proc. of ICASSP"*, Mar. 2016, pp. 5120–5124.

[8] P. Alku, "Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications. (invited article)," *Sadhana – Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 623–650, 2011.

[9] H. Kameoka and K. Kashino, "Composite autoregressive system for sparse source-filter representation of speech," in *Proc. of ISCAS*, May 2009, pp. 2477–2480.

[10] H. Kameoka, T. Yoshioka, M. Hamamura, J. Le Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 245–253.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[12] H. Kameoka, M. Goto, and S. Sagayama, "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes," *IPSJ Technical Report*, vol. 2006-MUS-66, no. 90, pp. 77–84, Aug. 2006, in Japanese.

[13] H. Kameoka, "Non-negative matrix factorization and its variants for audio signal processing," in *Applied Matrix and Tensor Variate Data Analysis*, T. Sakata, Ed. Springer Japan, 2016, ch. 2, pp. 23–51.

[14] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[15] E. W. Stacy, "A generalization of the gamma distribution," *The Annals of Mathematical Statistics*, pp. 1187–1192, 1962.

[16] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *Proc. of ICASSP*, vol. 2, Apr. 1997, pp. 1295–1298 vol.2.

[17] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[18] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation," in *Proc. of APSIPA ASC*, Dec. 2015, pp. 520–529.

[19] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–15, 1995.

[20] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[21] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[22] T. Hacki, "Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie," *Folia Phoniatrica et Logopaedica*, vol. 41, no. 1, pp. 43–48, 1989.

[23] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals." in *Proc. of Interspeech*, 2009, pp. 2891–2894.