Aalto University
School of Electrical Engineering

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

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

[1] *Aalto University, Finland* [2] *University of Tokyo, Japan*

[3] *Nippon Telegraph and Telephone Corporation, Japan*

[4] *National Institute of Informatics, Japan*

Introduction

Composite autoregressive system
- System structure
- Signal model

Majorisation-minimisation
- Generalized gamma prior
- Auxiliary function
- Update rules

Experiments
- Source prior
- Convergence rate
- Glottal inverse filtering
- Test signals
- Evaluation

Conclusion

Introduction

- Glottal inverse filtering (GIF) is useful, for example in excitation modeling parametric speech synthesis [Raitio et al., 2011]
- Recent text-to-speech synthesis quality improvements using quasi-closed phase (QCP) inverse filtering [Juvela et al., 2016]

Issues:

- QCP requires accurate pitch-marks, which are difficult to estimate with breathy voices or noisy speech
- Frame-by-frame analysis does not take advantage of the relatively stationary voice production process

Introduction

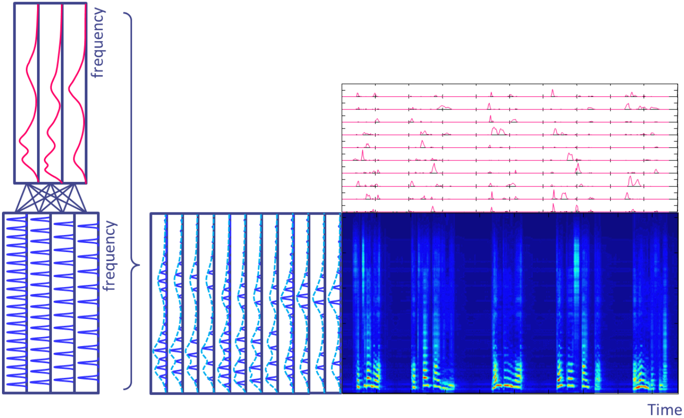
- Composite autoregressive (CAR) system [Kameoka and Kashino, 2009] provides a robust statistical model for source-filter estimation
- Model is optimised in time-frequency domain, which allow taking advantage of inter-frame dependencies
- Current expectation-maximisation (EM) based algorithm is somewhat slow

This paper:

- Derive faster optimisation method for the CAR system, similarly to NMF multiplicative updates
- Develop a GIF method based on CAR

Composite autoregressive system

- Spectrum is modeled as a weighted sum of source and filter pair combinations



Signal model

- Each component in frame n has the distribution $\boldsymbol{X}_n^{i,j} \sim \mathcal{N}_{\mathbb{C}}(0, \boldsymbol{\Lambda}_n^{i,j})$, where $\boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}(\lambda_{1,n}^{i,j}, \cdots, \lambda_{K,n}^{i,j})$
- The observed complex spectrogram $\boldsymbol{Y}_n$ is given by sum of $\boldsymbol{X}_n^{i,j}$

$$\boldsymbol{Y}_n = \sum_{i,j} \boldsymbol{X}_n^{i,j} \sim \mathcal{N}(0, \boldsymbol{\Phi}_n), \quad (1)$$

$$\boldsymbol{\Phi}_n = \sum_{i,j} \boldsymbol{\Lambda}_n^{i,j} = \mathrm{diag}(\phi_{1,n}, \ldots, \phi_{K,n}) \quad (2)$$

Composite autoregressive system

- Model component $\lambda_{k,n}^{i,j}$ for frame n and spectrum bin k
- Source components F_k^i, in total I templates
- All-pole filter components $H_k^j = 1/|A^j(e^{\mathrm{j}2\pi k/K})|^2$, in total J
- Model spectrogram component $\phi_{k,n} = \sum_{i,j} \lambda_{k,n}^{i,j}$

$$\lambda_{k,n}^{i,j} = \frac{U_n^{i,j} F_k^i}{\left|A^j(e^{\mathrm{j}2\pi k/K})\right|^2} = U_n^{i,j} F_k^i H_k^j \quad (3)$$

$$A^j(z) = 1 - \alpha_1^j z^{-1} - \cdots - \alpha_P^j z^{-P} \quad (4)$$

Composite autoregressive system

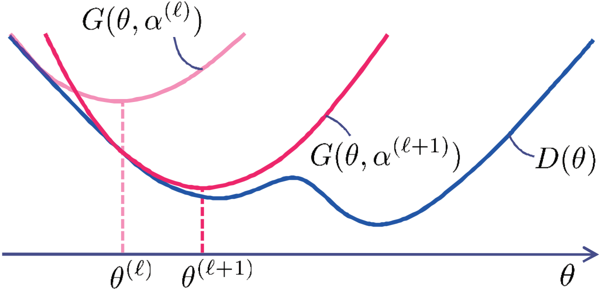
- Maximising the likelihood of $Y_{k,n}$ with respect to $\phi_{k,n}$ amounts to minimising the Itakura-Saito divergence D_{IS}

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

- We already know how to do this for NMF (good description in [Kameoka, 2016])
- MM gives the multiplicative NMF update rules

Majorisation-minimisation

- Minimizing objective function $D(\theta)$ directly is difficult
- Construct upper bound (auxiliary) function $G\left(\theta, \alpha^{(\ell)}\right)$ that is easy to minimize
- Alternating between setting a new $G\left(\theta, \alpha^{(\ell)}\right)$ (majorization) and updating θ (minimization) is guaranteed to decrease $D(\theta)$



Generalized gamma prior

What do we want from the prior?

- Should induce sparsity in the activations (approximately only one source-filter pair is active at a time)
- Prior mean should encourage spectral tilt to the source templates
- Use generalized gamma prior (with parameters η, d, p), shown here only for the activations U

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

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

Generalized gamma prior

- U^p term is problematic, construct upper bound by constraining $p < 1$, making U^p concave
- Then U^p is bounded by its tangent at $U = V$, where V is the auxiliary variable

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

Auxiliary function

- Construct upper bound by using Jensen's inequality and tangent inequality

$$
\begin{aligned}
D_{\mathrm{IS}}(\boldsymbol{Y}, \boldsymbol{\Phi}) &= \sum_{k,n} \left(\frac{Y_{k,n}}{\phi_{k,n}} + \log(\phi_{k,n}) \right) - \sum_{n,i,j} \log(p(U_n^{i,j})) \\
G_{\mathrm{IS}} &= \sum_{k,n} \left[\sum_{i,j} \frac{Y_{k,n} (\xi_{k,n}^{i,j})^2}{F_k^i U_n^{i,j} H_k^j} + \sum_{i,j} \frac{F_k^i U_n^{i,j} H_k^j}{\alpha_{k,n}} \right] \\
&\quad - \sum_{n,i,j} \left[(d-1) \log(U_n^{i,j}) \right. \\
&\quad \left. - \frac{1}{\eta} p (V_n^{i,j})^{p-1} (U_n^{i,j} - V_n^{i,j}) + \frac{1}{\eta} (V_n^{i,j})^p \right]
\end{aligned}
\tag{9}
$$

Auxiliary function

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

$$\xi_{k,n}^{i,j} = \frac{U_n^{i,j} F_k^i H_k^j}{\phi_{k,n}} \tag{10}$$

$$\alpha_{k,n} = \phi_{k,n} \tag{11}$$

$$V_n^{i,j} = U_n^{i,j} \tag{12}$$

Update rules

- Differentiating G_{IS} w.r.t. $U_n^{i,j}$ and substituting the aux. vars. gives

$$U_n^{i,j} \leftarrow \frac{b_U + \sqrt{b_U^2 + 4a_U c_U}}{2a_U} \tag{13}$$

$$a_U = \sum_k \frac{F_k^i H_k^j}{\phi_{k,n}} + \frac{p}{\eta}(U_n^{i,j})^{p-1} \tag{14}$$

$$b_U = K(d-1) \tag{15}$$

$$c_U = \sum_k \frac{Y_{k,n} F_k^i H_k^j (U_n^{i,j})^2}{\phi_{k,n}^2} \tag{16}$$

- Constraining $d \geq 1$ guarantees positivity

Update rules

- Similar procedure for F_k^i gives

$$F_k^i \leftarrow \frac{b_H + \sqrt{b_H^2 + 4a_H c_H}}{2a_H} \quad (17)$$

$$a_H = \sum_{n,j} \frac{U_n^{i,j} H_k^j}{\phi_{k,n}} + \frac{p}{\eta_{k,i}} (F_k^i)^{p-1} \quad (18)$$

$$b_H = NJ(d-1) \quad (19)$$

$$c_H = \sum_{n,j} \frac{Y_{k,n} U_n^{i,j} H_k^j (F_k^i)^2}{\phi_{k,n}^2} \quad (20)$$

Update rules

- Uniform prior simplifies update to

$$F_k^i \leftarrow F_k^i \sqrt{\frac{\sum_{n,j} \frac{Y_{k,n} U_n^{i,j} H_k^j}{\phi_{k,n}^2}}{\sum_{n,j} \frac{U_n^{i,j} H_k^j}{\phi_{k,n}}}} \tag{21}$$

- This closely resembles the I-S multiplicative NMF updates

Update rules

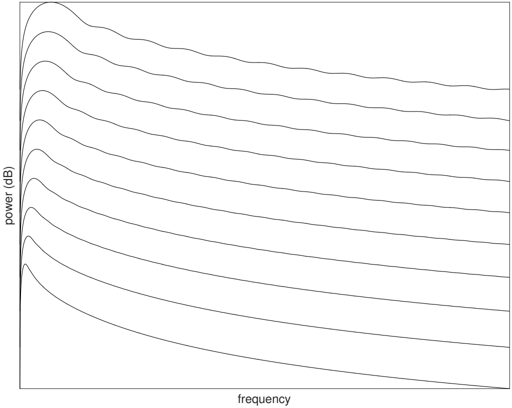
- The all-pole filter coefficients are solved from the normal equations resulting from a similar update in spectral domain

$$\begin{bmatrix} r_0^j & r_1^j & \cdots & r_{P-1}^j \\ r_1^j & r_0^j & & r_{P-2}^j \\ \vdots & & \ddots & \vdots \\ r_{P-1}^j & r_{P-2}^j & \cdots & r_0^j \end{bmatrix} \begin{bmatrix} a_1^j \\ a_2^j \\ \vdots \\ a_P^j \end{bmatrix} = \begin{bmatrix} r_1^j \\ r_2^j \\ \vdots \\ r_P^j \end{bmatrix} \tag{22}$$

$$r_{1,\ldots,P}^j = \mathrm{DFT}^{-1} \left\{ H_k^j \sqrt{\frac{\sum_{n,i} \frac{Y_{k,n} U_n^{i,j} F_k^i}{\phi_{k,n}^2}}{\sum_{n,i} \frac{U_n^{i,j} F_k^i}{\phi_{k,n}}}} \right\} \tag{23}$$

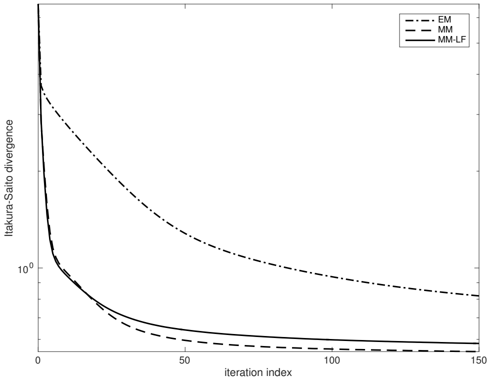
Source prior

- Set an LF glottal model based prior for the source templates
- Use the η_k^i, parameter which is proportional to distribution mean



Convergence rate

- Plot mean I-S divergence versus iteration index
- MM-based methods converge faster than the original EM-based method



Glottal inverse filtering

- The expected value of individual model component contribution to $Y_{n,k}$ is

$$\mathbb{E}\left[\hat{Y}^{i,j}_{k,n}\right] = Y_{k,n}\frac{\lambda^{i,j}_{k,n}}{\phi_{k,n}} = Y_{k,n}\frac{U^{i,j}_{n}F^{i}_{k}H^{j}_{k}}{\phi_{k,n}} \tag{24}$$

- Source estimate is obtained by removing the filter component and summing over components

$$\hat{S}_{k,n} = \sum_{i,j}\hat{S}^{i,j}_{k,n} = \sum_{i,j}Y_{k,n}\frac{U^{i,j}_{n}F^{i}_{k}}{\phi_{k,n}} \tag{25}$$

Test signals

- Synthetic signals with known ground truth are required to evaluate glottal inverse filtering methods
- Use a corpus of sustained Finnish vowels with labelled neutral, breathy and pressed phonation
- Estimate LPC envelope and f_0 from speech, synthesize with LF pulses modified by harmonic-to-noise ratio (HNR)

Style	Speaker 1	Speaker 2
Neutral	▶ orig.	▶ orig.
	▶ syn.	▶ syn.
Breathy	▶ orig.	▶ orig.
	▶ syn.	▶ syn.
Pressed	▶ orig.	▶ orig.
	▶ syn.	▶ syn.

Evaluation

- Compare with iterative adaptive inverse filtering (IAIF) [Alku, 1992] and QCP methods (both currently used in glottal vocoding)
- Use error measures derived from glottal parameterisations:
  - Mean squared error (MSE)
  - First harmonic to second harmonic difference in dB (H1H2)
  - Harmonic Richness Factor (HRF)
  - Normalised Amplitude Quotient (NAQ)
  - Quasi-Open Quotient (QOQ)
- Error measures grouped by phonation, lower score is better
- Proposed method CAR-MM without source prior and CAR-MM-LF with LF-based source prior

Evaluation

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

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

Evaluation

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

Conclusion

- Composite autoregressive system provides a convenient NMF like source-filter modelling framework
- Derived MM optimisation algorithm for CAR system converges faster than the original EM-based algorithm
- Proposed glottal inverse filtering method performs reasonably well for neutral and pressed phonation and outperforms reference methods with breathy phonation

References

[Alku, 1992] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2–3):109–118. Eurospeech '91.

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

[Kameoka, 2016] Kameoka, H. (2016). Non-negative matrix factorization and its variants for audio signal processing. In Sakata, T., editor, *Applied Matrix and Tensor Variate Data Analysis*, chapter 2, pages 23–51. Springer Japan.

[Kameoka and Kashino, 2009] Kameoka, H. and Kashino, K. (2009). Composite autoregressive system for sparse source-filter representation of speech. In *Proc. of ISCAS*, pages 2477–2480.

[Raitio et al., 2011] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. (2011). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):153–165.