Joint Bayesian Estimation of Time-Varying LP Parameters and Excitation for Speech

Srikanth Raj Chetupalli and T. V. Sreenivas,
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore.

1. Introduction

Joint estimation of time-varying linear prediction (TVLP) filter coefficients and the excitation signal parameters for long duration speech segments

Salient Features:

- random Gaussian prior for the prediction coefficients promotes sparse filters.
- Student's-t excitation model: random Gaussian excitation with time-dependent Gamma distributed precision. Learning parameters of Gamma prior can adapt to different excitation distributions.
- Maximum likelihood parameter estimation: iterative Expectation Maximization (EM) algorithm.

2. Time varying linear prediction

- Speech $x[n]$ is modeled as the output of a time-varying auto-regressive system of order p, excited by $e[n]$

$$x[n] = \sum_{k=1}^{p} a_k[n]x[n-k] + e[n], \; n \in [0 \quad N-1].$$

- Signal modeling $\implies$ estimate $\{a_k[n]\}$ under some assumptions about $e[n]$.
- Under-determined: Np parameters and N observations.
- Solution: Parametric model for $a_k[n]$

$$a_k[n] = \sum_{j=1}^{q} a_{kj}u_j[n],$$

  $\{u_j[n]\}$ is a known basis set.
  Examples: DCT, Fourier basis, Power series etc.

- In vector form:

$$x[n] = \mathbf{x}_n^T\mathbf{a} + e[n] \; \forall \; n, \text{ and } \mathbf{x} = \mathbf{X}\mathbf{a} + \mathbf{e}.$$

- *Speech excitation is sparse for voiced sounds, Gaussian like for unvoiced sounds.*
- *For long duration segments, the excitation distribution is non-Gaussian.*
- *ℓ_2 or ℓ_1 minimization are not optimal for non-Gaussian excitation.*

References

1. M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech," Signal Processing, vol. 5, no. 3, pp. 267-285, 1983.
2. P. Ha and S. Ann, "Robust time-varying parametric modelling of voiced speech," Signal Processing, vol. 42, no. 3, pp. 311-317, 1995.
3. D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 5, pp. 1644-1657, July 2012.

3. Proposed signal model

- Excitation signal is independent Gaussian distributed, with time dependent variance

$$p(\mathbf{x}|\mathbf{a},\boldsymbol{\Gamma}) = \prod_{n=0}^{N-1}\sqrt{\frac{\gamma_n}{2\pi}}\exp\left[-\frac{\gamma_n}{2}(x[n]-\mathbf{x}_n^T\mathbf{a})^2\right].$$

- Precision γ_n is Gamma distributed,

$$p(\boldsymbol{\Gamma}|\alpha,\beta) = \frac{\beta^{\alpha N}}{\Gamma(\alpha)^N}\left(\prod_{n=0}^{N-1}\gamma_n^{\alpha-1}\right)\exp\left(-\beta\sum_{n=0}^{N-1}\gamma_n\right).$$

  Marginal distribution for $x[n]$ is Student's-t.

- Prediction filter $\mathbf{a}$ is Gaussian distributed,

$$p(\mathbf{a}|\boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{1/2}\exp\left(-\frac{1}{2}\mathbf{a}^T\boldsymbol{\Lambda}\mathbf{a}\right); \quad \boldsymbol{\Lambda} = diag(\lambda_i).$$

  Independent Gaussian model promotes sparse predictor.

- Joint distribution

$$p(\mathbf{x},\mathbf{a},\boldsymbol{\Gamma}|\boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{a},\boldsymbol{\Gamma})p(\mathbf{a}|\boldsymbol{\Lambda})p(\boldsymbol{\Gamma}|\alpha,\beta).$$

- Total log-likelihood

$$\begin{aligned}\log[p(\mathbf{x},\mathbf{a},\boldsymbol{\Gamma}|\boldsymbol{\theta})] &\propto (\alpha-1)\sum_{n=0}^{N-1}\log(\gamma_n) - \beta\sum_{n=0}^{N-1}\gamma_n \\ &+ N\alpha\log(\beta) - N\log(\Gamma(\alpha)) + \frac{1}{2}\log(|\boldsymbol{\Lambda}|) - \frac{1}{2}\mathbf{a}^T\boldsymbol{\Lambda}\mathbf{a} \\ &+ \frac{1}{2}\sum_{n=0}^{N-1}\left[\log(\gamma_n) - \frac{1}{2}\gamma_n(x[n]-\mathbf{x}_n^T\mathbf{a})^2\right].\end{aligned}$$

4. Maximum likelihood estimation

Log-Likelihood: $\log p(\mathbf{x}) \geq \mathcal{L}(q,\boldsymbol{\theta})$

$$\mathcal{L}(q,\boldsymbol{\theta}) \triangleq \mathbb{E}_{(\boldsymbol{\Gamma},\mathbf{a})}\left[\log p(\mathbf{x},\boldsymbol{\Gamma},\mathbf{a}|\boldsymbol{\theta})\right] - \mathbb{E}_{(\boldsymbol{\Gamma},\mathbf{a})}\left[\log q(\boldsymbol{\Gamma},\mathbf{a})\right],$$

for any $q(\boldsymbol{\Gamma},\mathbf{a})$ defined over the joint support of $\{\boldsymbol{\Gamma},\mathbf{a}\}$, and $\boldsymbol{\theta} = \{\alpha,\beta,\boldsymbol{\Lambda}\}$.

Maximize using EM-like approach:

- E-step: Maximize $\mathcal{L}(q,\boldsymbol{\theta})$ w.r.t $q(\boldsymbol{\Gamma},\mathbf{a})$ for fixed $\boldsymbol{\theta}$.
  - $q(\boldsymbol{\Gamma},\mathbf{a}) = p(\boldsymbol{\Gamma},\mathbf{a}|\mathbf{x})$ achieves objective, but not tractable.
  - Assume factorization $q(\boldsymbol{\Gamma},\mathbf{a}) = q(\boldsymbol{\Gamma})q(\mathbf{a})$, and perform coordinate ascent: Mean field variational inference.
  - Closed form expressions for $q(.)$: conjugate priors.
- M-step: Maximize $\mathcal{L}(q,\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ for fixed $q(\boldsymbol{\Gamma},\mathbf{a})$.

$$\mathcal{L}(q,\boldsymbol{\theta}) \leq \mathcal{L}(q^*,\boldsymbol{\theta}) \leq \mathcal{L}(q^*,\boldsymbol{\theta}^*)$$

5. Algorithm steps

$$\mathcal{L}(q,\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\Gamma}}\mathbb{E}_{\mathbf{a}}\left[\log p(\mathbf{x},\boldsymbol{\Gamma},\mathbf{a}|\boldsymbol{\theta})\right] - \mathbb{E}_{\boldsymbol{\Gamma}}\left[\log q(\boldsymbol{\Gamma})\right]\mathbb{E}_{\mathbf{a}}\left[\log q(\mathbf{a})\right]$$

Algorithm:

Inputs: $\mathbf{x}$, $\mathbf{X}$.
Initialize $i = 0$, $\boldsymbol{\theta}^0 = \{1, 0.001, 0.01\mathbf{I}\}$.

while not converged **do**

E-step:
Maximize w.r.t $\mathbf{a}$:

$$\log q^{(i)}(\mathbf{a}) \propto \mathbb{E}_{\boldsymbol{\Gamma}}\left[\log p(\mathbf{x},\boldsymbol{\Gamma},\mathbf{a}|\boldsymbol{\theta}^{(i-1)})\right]$$

$q^{(i)}(\mathbf{a})$ is Gaussian: $\mathcal{N}(\tilde{\mathbf{a}},\tilde{\boldsymbol{\Lambda}})$.
Maximize w.r.t γ_n:

$$\log q^{(i)}(\gamma_n) \propto \mathbb{E}_{\mathbf{a}}\left[\log p(\mathbf{x},\boldsymbol{\Gamma},\mathbf{a}|\boldsymbol{\theta}^{(i-1)})\right]$$

$q^{(i)}(\gamma_n)$ is Gamma distributed: $\Gamma(\alpha^i + \frac{1}{2}, \beta^i + \frac{1}{2}\mathbb{E}\{(x[n]-\mathbf{x}_n^T\mathbf{a})^2\})$.

M-step: α^{i+1} is a solution to,

$$\log\alpha - \psi(\alpha) = \log\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\{\gamma_n\}\right) - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\{\log(\gamma_n)\},$$

$$\frac{1}{\beta^{i+1}} = \frac{1}{N\alpha^{i+1}}\sum_{n=1}^{N}\mathbb{E}\{\gamma_n\}, \; 1/\lambda_k^{i+1} = \left[\mathbb{E}\{\mathbf{a}\mathbf{a}^T\}\right]_{kk}$$

$i \leftarrow i+1$
Output: $\tilde{\mathbf{a}}$.

6. Evaluation (synthetic signals)

- Synthetic signals are generated using the TVLP model: LP order P=10, DCT basis order $q = 7$.
- Coefficient trajectories derived from 256 ms segments taken from 10 TIMIT sentences (142 segments).
- Excitation signal is generated as

$$e[n] = q[n] + w[n]$$

  $q[n]$ is a periodic impulse train 250 Hz; $w[n]$ is zero mean white Gaussian noise of variance σ_w^2.

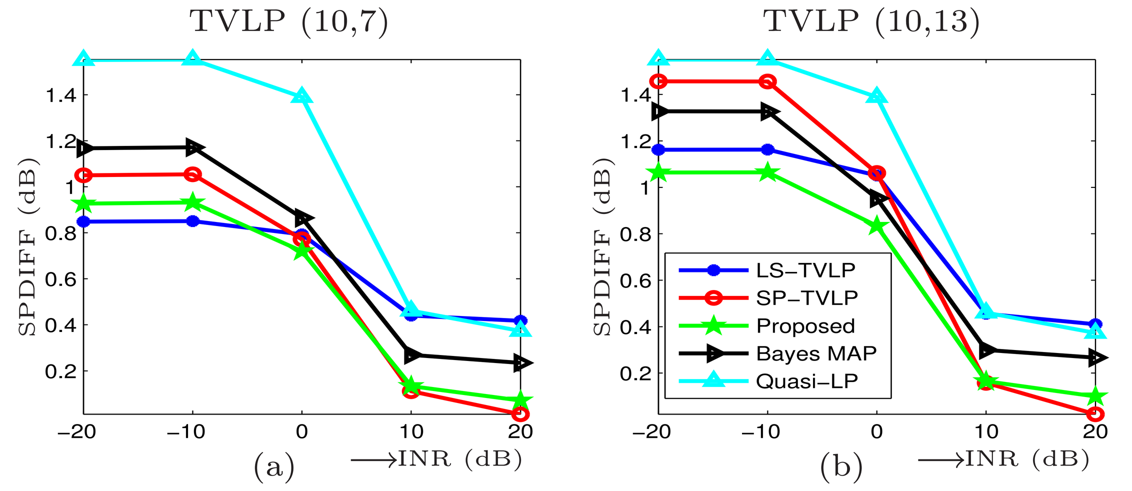
- σ_w^2 is chosen such that Impulse to Noise Ratio (INR) defined below has a specific value,

$$\text{INR } (dB) = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1}q^2[n]/\sigma_w^2\right)$$

  small INR $\implies$ Gaussian excitation;
  high INR $\implies$ Sparse excitation.

- Performance measured using average spectral difference (SPDIFF) measure.

7. SPDIFF vs INR

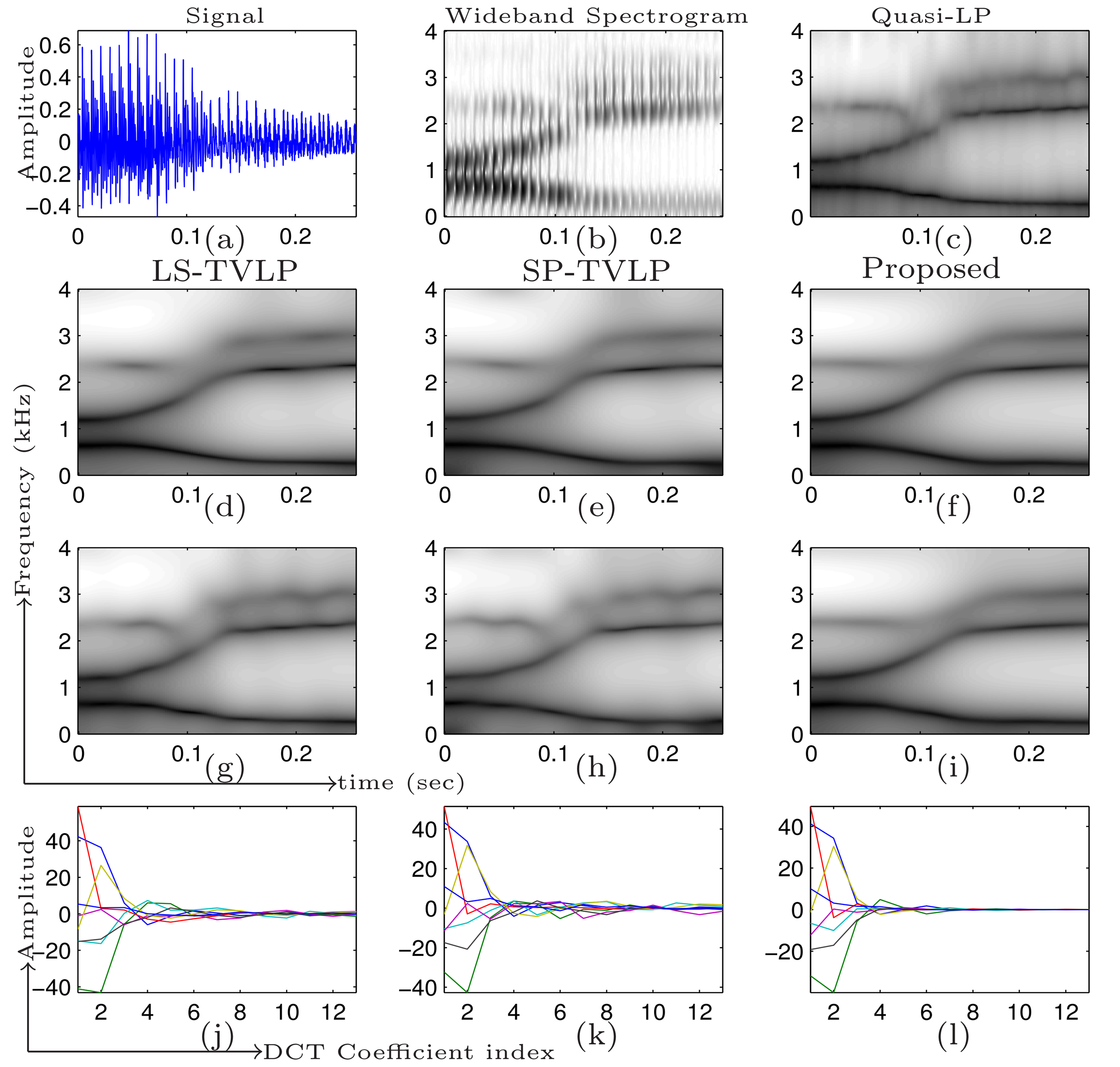


Average SPDIFF measure for TVLP analysis.

- Quasi stationary analysis has the highest SPDIFF.
- Sparse excitation model based methods perform better for high INR, and Gaussian excitation model based methods perform better for small INR.
- Ground truth model order $(10, 7)$
  - For Gaussian like excitation (INR= -20 dB), LS TVLP performs better.
  - For sparse excitation (INR= -20 dB), SP TVLP performs better.
  - Intermediate values for INR, Bayesian TVLP performs better.
- Over estimated model order $(10, 13)$
  - Bayes TVLP performs better for all values of INR.

** *Proposed Bayesian TVLP approach performs better for different excitation signal types.*

8. Evaluation (Speech Signal)



(d-f) TVLP model of order $(8, 7)$, (g-l) TVLP model of order $(8, 13)$.