BAYESIAN LEARNING FOR TIME-VARYING LINEAR PREDICTION OF SPEECH

Adrià Casamitjana[*], *Martin Sundin*[*], *Prasanta Ghosh*[†], *Saikat Chatterjee*[*]

[*] ACCESS Linneaus Center, KTH Royal Institute of Technology, Sweden
[†] Department of Electrical Engineering, Indian Institute of Science, India
Emails: adriacd@kth.se, masundi@kth.se, prasantg@ee.iisc.ernet.in, sach@kth.se

ABSTRACT

We develop Bayesian learning algorithms for estimation of time-varying linear prediction (TVLP) coefficients of speech. Estimation of TVLP coefficients is a naturally underdetermined problem. We consider sparsity and subspace based approaches for dealing with the corresponding underdetermined system. Bayesian learning algorithms are developed to achieve better estimation performance. Expectation-maximization (EM) framework is employed to develop the Bayesian learning algorithms where we use a combined prior to model a driving noise (glottal signal) that has both sparse and dense statistical properties. The efficiency of the Bayesian learning algorithms is shown for synthetic signals using spectral distortion measure and formant tracking of real speech signals.

***Index Terms*—** Time-varying linear prediction, sparsity, Bayesian learning, expectation-maximization.

1. INTRODUCTION

Time-varying linear prediction (TVLP) model [1,2] of speech signals is a generalization over the much used linear prediction model [3,4]. In TVLP, each speech signal sample is predicted by a time-varying linear combination of past samples. Here, linear combination coefficients are time-varying unlike the case of standard linear prediction (LP) where the coefficients are fixed. Hence, for each speech signal frame there are more TVLP coefficients to estimate and the associated estimation problem becomes underdetermined. To keep the problem determined (or manageable), a smoothness constraint is typically imposed on time-varying coefficients. For example, time-varying coefficients have dynamics that can be modeled by a linear combination of low-frequency cosine functions, that means by using a fixed subspace in a known cosine basis. Several works considered different basis functions, such as Legendre [5], Fourier [1], discrete prolate spheroidal functions [6], wavelets [7]. For a fixed subspace, a least-squares based method is predominant to estimate the TVLP coefficients. Independently of the basis functions used, the method can be viewed as a subspace estimation method.

In TVLP, a major consideration is modeling the driving noise. For speech signals, driving noise corresponds to a glottal signal. Typically a glottal signal is either assumed to be white noise for unvoiced sounds or periodic pulses at pitch frequency for voiced sounds. Unlike earlier works, we model the glottal signal as an additive combination of sparse (pulses) and dense (white noise) noise. Modeling of the glottal signal by purely sparse noise was recently considered in [8]. The authors in [8] considered an ideal data fit cost, that is, ℓ_0-norm based cost minimization. The practical strategy was to use an iteratively re-weighted least-squares based algorithm (a deterministic solution). In this context, we mention that, for a standard linear prediction scheme, the use of sparse noise is considered in [9]. The work of [9] mainly used convex optimization to minimize a ℓ_1-norm based data fit cost. Further, a Bayesian learning algorithm for a standard linear prediction was recently considered in [10] where a glottal signal was modeled as a combination of dense noise and block-structured sparse noise.

In this paper, we begin with the under-determined problem setup of TVLP and use Bayesian learning for estimation of the TVLP coefficients. We consider a data driven approach (no statistical stationarity is assumed) where the driving noise is modeled by a combination of sparse and dense noise. Bayesian learning methods are derived using expectation-maximization (EM) framework. Through simulations, we have found that the use of sparsity does not lead to a good performance for the undetermined TVLP problem. Therefore, we convert the problem to a determined setup using a fixed subspace. For the determined setup, we show that the Bayesian learning method provides a better estimate of TVLP coefficients vis-a-vis competing methods.

2. TIME-VARYING LINEAR PREDICTION SYSTEM

In TVLP, the n'th speech sample x_n is modeled as

$$x_n = \sum_{p=1}^{P} a_n(p)\, x_{n-p} + q_n \qquad (1)$$

where P is the order of the predictor and $\{a_n(p)\}_{p=1}^{P}$ are the TVLP coefficients at sample n. The term q_n is the driving noise of the generative process model (1) and it is assumed to model the glottal signal. In our case, we assume that q_n has two additive parts: the sparse noise e_n to purely voiced sounds and the dense noise w_n for purely unvoiced sounds. Both allow us to represent any kind of speech sound. Let us consider that an N-point sequence of the signal x_n is represented by $\mathbf{x}$ vector. Following (1) with $q_n = e_n + w_n$, we write

$$\mathbf{x} = \bar{\mathbf{X}}\bar{\mathbf{a}} + \mathbf{e} + \mathbf{w} \in \mathbb{R}^{N\times 1}, \tag{2}$$

where $\mathbf{x} = [x_1, x_2, \dots x_N]$, $\bar{\mathbf{a}} = [\mathbf{a}_1^\top, \mathbf{a}_2^\top, \dots, \mathbf{a}_N^\top]^\top \in \mathbb{R}^{PN\times 1}$ is the vector of TVLP coefficients where $\mathbf{a}_i$ represents the vocal shape at each time instant, $\mathbf{e}$ is the sparse noise vector and $\mathbf{w}$ is the dense noise vector, and finally $\bar{\mathbf{X}} \in \mathbb{R}^{N\times PN}$ is the data matrix as below

$$\bar{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1^\top & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_2^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{x}_3^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{x}_N^\top \end{bmatrix}. \tag{3}$$

Here $\mathbf{x}_n = [x_{n-1}, x_{n-2}, \dots x_{n-P}]^\top$, and $\mathbf{a}_n = [a_n(1), a_n(2), \dots, a_n(P)]^\top$. Let $\mathbf{A} \in \mathbb{R}^{P\times N}$ be formed by the column vectors $\mathbf{a}_n$ as $\mathbf{A} = [\mathbf{a}_1\, \mathbf{a}_2 \dots \mathbf{a}_N]$, and hence $\text{vec}(\mathbf{A}) = \bar{\mathbf{a}}$. Denoting the k'th row of $\mathbf{A}$ by $\mathbf{A}(k,:)$ as $\mathbf{A}(k,:) = [a_1(k)\, a_2(k) \dots a_N(k)]$, we assume that the components of $\mathbf{A}(k,:)$ are slowly varying over n and hence highly correlated. The correlation can be exploited by a decorrelating orthonormal transform $\mathbf{T}_1^{-1} \in \mathbb{R}^{N\times N}$ in the transform domain $\mathbf{T}_1^{-1}\mathbf{A}^\top$. Further the correlation between the components of rows of $\mathbf{T}_1^{-1}\mathbf{A}^\top$ can be exploited by another orthonormal transform $\mathbf{T}_2^{-1} \in \mathbb{R}^{P\times P}$, and hence the full transform can be realized as $\mathbf{T}_2^{-1}[\mathbf{T}_1^{-1}\mathbf{A}^\top]^\top$. .Using the relation $\text{vec}(\mathbf{T}_2[\mathbf{T}_1^{-1}\mathbf{A}^\top]^\top) = \text{vec}(\mathbf{T}_2^{-1}\mathbf{A}[\mathbf{T}_1^{-1}]^\top) = [\mathbf{T}_1^{-1} \otimes \mathbf{T}_2^{-1}]\text{vec}(\mathbf{A}) = [\mathbf{T}_1 \otimes \mathbf{T}_2]^{-1}\bar{\mathbf{a}} = \mathbf{d}$, we can write (1) as

$$\mathbf{x} = \bar{\mathbf{X}}\mathbf{T}\mathbf{d} + \mathbf{e} + \mathbf{w}, \tag{4}$$

where $\mathbf{T} \triangleq [\mathbf{T}_1 \otimes \mathbf{T}_2]$ and $\otimes$ denotes Kronecker product. In this paper, we use the standard discrete cosine transform (DCT) II for $\mathbf{T}_1^{-1}$ and identity matrix for $\mathbf{T}_2$. Hence $\mathbf{T}$ is known. Note that $\mathbf{x}$ is N-dimensional and $\mathbf{d}$ is PN dimensional. Therefore, (4) is underdetermined by a factor of P. Denoting an estimate of $\mathbf{d}$ by $\hat{\mathbf{d}}$, we find the estimate of $\bar{\mathbf{a}}$ by $\mathbf{T}\hat{\mathbf{d}}$. To estimate $\mathbf{d}$ we can use two approaches: (1) sparsity assumption on $\mathbf{d}$, and (2) restrict to the first L coefficients of $\mathbf{d}$. The first approach is motivated by standard sparse representations and compressed sensing. The second approach fixes in advance the subspace where most of the energy is concentrated and thus, deals with a determined system (subspace based estimation). Denoting the first L columns of $\mathbf{T}$ by $\mathbf{T}_{(L)}$, and correspondingly first L coefficients of $\mathbf{d}$ by $\mathbf{d}_{(L)}$, the determined setup is

$$\mathbf{x} = \bar{\mathbf{X}}\mathbf{T}_{(L)}\mathbf{d}_{(L)} + \mathbf{e} + \mathbf{w} + \mathbf{n}, \tag{5}$$

where $\mathbf{n}$ is the noise due to truncation ($\mathbf{d}$ to $\mathbf{d}_{(L)}$). We use $P \le L \le PN$. For TVLP, we must need $L > P$. Note that, as $\mathbf{T}_1$ and $\mathbf{T}_2$ are orthonormal matrices, the case $L = P$ corresponds to a standard linear prediction, that means $\forall n,\ a_n(p) = a(p)$.

3. BAYESIAN LEARNING

Using EM framework, we consider Bayesian learning for the two approaches in the following subsections.

3.1. Underdetermined setup

Here we deal with (4) where both $\mathbf{d}$ and $\mathbf{e}$ are assumed to be sparse, and $\mathbf{w}$ is dense. To reduce the number of parameters to be estimated as well as exploiting the structure of the tranform matrix $\mathbf{T}$, we use block sparsity in $\mathbf{d}$. Let $\mathbf{d}$ comprise K-dimensional sub-vectors $\mathbf{d}_i$ such that $\frac{PN}{K}$ is an integer. For Bayesian learning, we use the following Gaussian prior to promote block sparsity in $\mathbf{d}$

$$\begin{aligned} \mathbf{d} \sim \prod_{i=1}^{\frac{PN}{K}} \mathcal{N}(0, \gamma_i^{-1}\mathbf{I}) = \mathcal{N}(0, \boldsymbol{\Gamma}^{-1}), \\ \boldsymbol{\Gamma} = \text{diag}((\boldsymbol{\gamma}^\top \otimes \mathbf{1}_K)^\top), \end{aligned} \tag{6}$$

where $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \dots, \gamma_{\frac{PN}{K}}]^\top$ and $\mathbf{1}_K$ denotes a constant vector of ones of size $K \times 1$. For $K = 1$, the prior induces sparsity in a usual sense (fully unstructured), and its use can be found in several earlier works [11, 12] including our work [13]. Then, motivated by our recent result in [14], we use a combined model prior for the joint noise as

$$\mathbf{e} + \mathbf{w} \sim \prod_{i=1}^{N} \mathcal{N}(0, \beta_i^{-1}) = \mathcal{N}(0, \mathbf{B}^{-1}),\ \ \mathbf{B} = \text{diag}(\boldsymbol{\beta}), \tag{7}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_N]^\top$. The precisions are $\{\gamma_i\}$ and $\{\beta_i\}$ that have Gamma distribution as hyper-priors

$$p(\gamma_i) = \text{Gamma}(\gamma_i | a+1, b),\ \ p(\beta_i) = \text{Gamma}(\beta_i | c+1, d),$$

where $\text{Gamma}(\gamma_i | a+1, b) \propto \gamma_i^a \exp(-b\gamma_i)$. The hyperparameters are $\{a, b, c, d\}$. We find the maximum-a-posteriori (MAP) estimate of $\mathbf{d}$ by maximization of $p(\mathbf{d}|\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\beta})$, as follows

$$\begin{aligned} \hat{\mathbf{d}} &= \boldsymbol{\Sigma}(\bar{\mathbf{X}}\mathbf{T})^\top\mathbf{B}\mathbf{x}, \\ \boldsymbol{\Sigma} &= (\boldsymbol{\Gamma} + (\bar{\mathbf{X}}\mathbf{T})^\top\mathbf{B}\bar{\mathbf{X}}\mathbf{T})^{-1}. \end{aligned} \tag{8}$$

The precisions are updated by using the EM algorithm. Let $\boldsymbol{\theta} \triangleq \{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ denote the parameters that are updated in each iteration by maximizing the cost (EM help function in MAP estimation)

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') + \ln p(\boldsymbol{\theta}), \tag{9}$$

where $\boldsymbol{\theta}'$ are the parameter values from the previous iteration. The function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is defined as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathcal{E}_{\mathbf{d}|\mathbf{x},\boldsymbol{\theta}'}[\ln p(\mathbf{x}, \mathbf{d}|\boldsymbol{\theta})],$$

where $\mathcal{E}$ denotes the expectation operator. The maximization of (9) leads to following update equations over iterations

$$\begin{aligned} \gamma_i^{new} &= \frac{K+2a}{\sum_{j=(i-1)K+1}^{iK}(\boldsymbol{\Sigma}'+\hat{\mathbf{d}}\hat{\mathbf{d}}^\top)_{jj}+2b}, \\ \beta_i^{new} &= \frac{1+2c}{(\mathbf{x}-\bar{\mathbf{X}}\mathbf{T}\hat{\mathbf{d}})_i^2+(\bar{\mathbf{X}}\mathbf{T}\boldsymbol{\Sigma}'(\bar{\mathbf{X}}\mathbf{T})^\top)_{ii}+2d}, \end{aligned}$$

where $\boldsymbol{\Sigma}' = (\boldsymbol{\Gamma}' + (\bar{\mathbf{X}}\mathbf{T})^\top\mathbf{B}'\bar{\mathbf{X}}\mathbf{T})^{-1}$. The derivation of the update equations is shown in Section 6. Here we seek sparsity promoting solutions, and hence the requirement is that most of the precisions of the prior distribution will turn out to be high. In order to satisfy the requirement, we use non-informative hyper-priors (flat distribution) for the precisions $\boldsymbol{\gamma}, \boldsymbol{\beta}$ by fixing their parameters to small values: e.g. $a = b = c = d \sim 10^{-3}$.

3.2. Determined setup

Here, we deal with (5) by using the following isotropic Gaussian prior

$$\mathbf{d}_{(L)} \sim \prod_{i=1}^{L} \mathcal{N}(0, \gamma^{-1}) = \mathcal{N}(0, \gamma^{-1}\mathbf{I}). \qquad (10)$$

The prior for $\mathbf{e} + \mathbf{w} + \mathbf{n}$ is the same prior as for $\mathbf{e} + \mathbf{w}$ in Section 3.1, as shown in (7). Using similar arguments as in section 3.1, the MAP estimate of $\mathbf{d}_{(L)}$ is

$$\begin{aligned} \hat{\mathbf{d}}_{(L)} &= \boldsymbol{\Sigma}_{(L)}(\bar{\mathbf{X}}\mathbf{T}_{(L)})^\top\mathbf{B}\mathbf{x}, \\ \boldsymbol{\Sigma}_{(L)} &= (\gamma\mathbf{I} + (\bar{\mathbf{X}}\mathbf{T}_{(L)})^\top\mathbf{B}\bar{\mathbf{X}}\mathbf{T}_{(L)})^{-1}. \end{aligned} \qquad (11)$$

and the update equations for the precisions are

$$\begin{aligned} \gamma^{new} &= \frac{L+2a}{\text{trace}(\boldsymbol{\Sigma}'_{(L)}+\hat{\mathbf{d}}_{(L)}\hat{\mathbf{d}}_{(L)}^\top)+2b}, \\ \beta_i^{new} &= \frac{1+2c}{(\mathbf{x}-\bar{\mathbf{X}}\mathbf{T}_{(\mathbf{L})}\hat{\mathbf{d}}_{(L)})_i^2+(\bar{\mathbf{X}}\mathbf{T}_{(L)}\boldsymbol{\Sigma}'_{(L)}(\bar{\mathbf{X}}\mathbf{T}_{(L)})^\top)_{ii}+2d}. \end{aligned} \qquad (12)$$

The derivation of the update equations is similar as before and hence not shown. In this case, we use a non-informative prior for $\mathbf{d}_{(L)}$ as we do not have a-priori knowledge about its properties.. We achieve this behavior by fixing the hyper-priors for $\boldsymbol{\gamma}$ as: $a \sim 10^{-3}$, $b \sim 10^4$. For the noise term $\mathbf{e} + \mathbf{w} + \mathbf{n}$ we still use a sparsity promoting solution, and hence we set their hyper-priors to $c = d \sim 10^{-3}$.

4. EXPERIMENTS

We evaluate the methods using synthetic signals as well as real speech where $P = 10$ is used and the sampling rate is 8 kHz. We considered window lengths of 20 ms, 40 ms and larger window length of 250 ms. The estimation methods we used are of two types: least-squares and Bayesian learning. In least-squares estimation, we compared following methods: [a] LP (stationary) - the so-called autocorrelation method using statistical stationarity by solving Yule-Walker equations, [b] LP (least-squares) - the so-called covariance method solving (5) for $L = P$ by using a standard least squares $\hat{\mathbf{d}}_{(L)} = [\bar{\mathbf{X}}\mathbf{T}_{(L)}]^\dagger\mathbf{x}$ (pseudo-inverse), [c] TVLP (least-squares) - solving (5) by pseudo-inverse, for $L = 30$ when window length is 20 ms, $L = 40$ when window length is 40 ms, $L = 60$ when window length is 250 ms. For Bayesian learning, we compared: [a] LP (Bayesian) - solving (5) for $L = P$ by using the relations in section 3.2, [b] TVLP (Bayesian) - solving (5) by using the relations in section 3.2, for $L = 30$ when window length is 20 ms, $L = 40$ when window length is 40 ms, $L = 60$ when window length is 250 ms, and [c] TVLP (U, Bayesian) - solving underdetermined setup (4) by using the relations in section 3.1 where the dimension of each block is $K = P = 10$. We also considered TVLP (U, Bayesian) for $K = 1$, that means using unstructured sparsity on $\mathbf{d}$, which provided highly degraded performance and we do not report the degraded results.

4.1. Synthetic signals

4.1.1. Signal generation

Synthetic signals are generated by the model (1) where we used different types of driving noise q_n:

i. Sparse noise: a pulse train with frequency 200 Hz to model a normal pitch frequency. The signal-to-driving-noise ratio (SDNR) is $\frac{\sigma_x^2}{\sigma_q^2} = 14.5$ dB.
ii. Dense noise: iid Gaussian noise signal such that SDNR = 13.1 dB.
iii. Joint noise: additive sparse noise and dense noise with equal variance such that SDNR = 12.5 dB.

To generate a stable signal x_n, we used minimum phase TVLP coefficients, this means the Z-domain analysis filter $1 - \sum_{k=1}^{P} a_n(k)Z^{-k}$ has roots inside the unit circle for all n. Further, to have a speech like signal, the TVLP coefficients are drawn from real speech signal analysis. Using a window length of 160 samples and window shift of one sample for a real 8 kHz speech signal, we performed autocorrelation based $P = 10$ order LP analysis that is guaranteed to provide stable filter coefficients. As the coefficients are drawn for every sample of the real speech signal, they can be used as known TVLP coefficients in the generative model (1).

4.1.2. Performance

For every n'th sample we compute the spectral distortion (SD) and report the average SD as the performance measure. The SD for the n'th sample is defined as

$$SD = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{S_n(\omega)}{\hat{S}_n(\omega)}\right]^2 d\omega\right]^{\frac{1}{2}}, \qquad (13)$$

Table 1: Comparison of methods for synthetic signals using average spectral distortion (average SD)

Method	Noise types		
	Sparse	Dense	Joint
Window length = 20 ms			
Ground truth (subspace)	0.85	0.85	0.85
LP (stationary)	2.26	2.88	2.78
LP (least-squares)	2.12	2.59	2.49
TVLP (least-squares)	1.53	2.86	2.64
LP (Bayesian)	1.93	2.59	2.33
TVLP (Bayesian)	1.07	2.95	2.38
TVLP (U, Bayesian)	2.65	3.93	4.57
Window length = 40 ms			
Ground truth (subspace)	1.21	1.21	1.21
LP (stationary)	2.81	3.06	2.96
LP (least-squares)	2.80	2.87	2.82
TVLP (least-squares)	1.68	2.55	2.39
LP (Bayesian)	2.56	2.87	2.73
TVLP (Bayesian)	1.22	2.62	2.16
TVLP (U, Bayesian)	1.97	3.43	2.94

where $S_n(\omega) = 1/|1 - \sum_{p=1}^{P} a_n(p)e^{(-j\omega p)}|^2$ is the power spectrum and $\hat{S}_n(\omega) = 1/|1 - \sum_{p=1}^{P} \hat{a}_n(p)e^{(-j\omega p)}|^2$ is the reconstructed power spectrum. The average SD is computed for signal samples across 100 windowed signal frames.

Table 1 shows the performance of all competing methods. The 'ground truth' refers to the determined setup (5) where a baseline average SD is the minimum to arise due to the truncation of $\mathbf{d}$ to $\mathbf{d}_{(L)}$. For sparse driving noise, TVLP (Bayesian) provides the best performance. Further, for dense and joint driving noise types, the TVLP (Bayesian) provides good performance. The TVLP (U, Bayesian) turns out to be degraded. A possible reason for degraded performance in the underdetermined setup (4) can be the poor condition of system matrix $\bar{\mathbf{X}}$, as reflected in (3). The system matrix $\bar{\mathbf{X}}$ is nonideal as it is far away from the usual dense wide matrices used in standard sparse representations and compressed sensing. We have independently verified that the TVLP (U,Bayesian) learning algorithm provides good performance in experiments with ideal dense system matrices (with iid Gaussian entries) for the considered system sizes. For brevity, we do not report these experiments in this paper.

4.2. Real speech signals

We used clean speech signals from the Noizeus database. Considering the task of formant tracking, Fig. 1 shows spectrograms of different methods for a 250 ms speech signal instance. Fig. 1 (a) comprises of a series of periodograms where each periodogram is computed for a 10 ms window length to have a better time resolution and a 5 ms shift to have a better tracking of formants. It is clear that the standard LP method comes with a high level of granularity, as we observe in Fig. 1 (b). Finally, by visually comparing Fig. 1 (c) and (d), TVLP using Bayesian learning can be found to be the best method in the sense of formant tracking. The formant tracking is more prominent than the competing method of TVLP using least squares. We rely on visual inspection as there is no available performance measure to quantify the improvement of Bayesian learning over least-squares for real speech formant tracking. At this point, we mention that the standard LP used 250 coefficients, whereas the TVLP methods used 60 coefficients (i.e. $L = 60$).

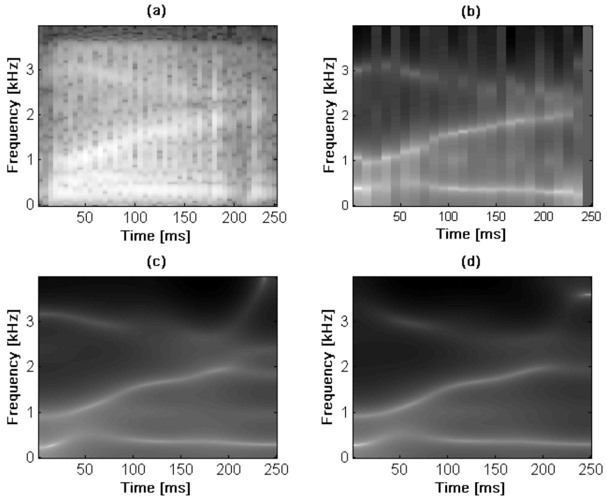

Fig. 1: Spectrograms as outputs of different methods for real speech signal. (a) DFT spectrum (periodogram) using window length of 10 ms and frame shift of 5 ms. (b) for LP using least-squares with window length of 20 ms and frame shift of 10 ms. (c) for TVLP using least-squares with window length of 250 ms. (d) TVLP using Bayesian learning with window length of 250 ms.

5. CONCLUSIONS

We deal with the problem of estimating TVLP coefficients from speech samples, corresponding to an underdetermined linear system. We have found that a direct assumption of unstructured sparsity does not lead to a good estimation performance. Instead a subspace approach provides reliable performance. Further, among several estimation methods, Bayesian learning is found to be the best, as it has higher generalization capability to model different statistics of model parameters and driving noise.

6. DERIVATION OF UPDATE EQUATIONS

To maximize $p(\mathbf{x}, \mathbf{d}|\boldsymbol{\theta})$ we use the EM algorithm with prior assumptions to its parameters $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$.

- 1. Choose an initial setting for parameters $\boldsymbol{\theta}'$

- 2. **E-step**. Evaluate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathcal{E}_{\mathbf{d}|\mathbf{x},\boldsymbol{\theta}'}[\ln p(\mathbf{x}, \mathbf{d}|\boldsymbol{\theta})]$

  From Bayes rule,

  $$\ln p(\mathbf{x}, \mathbf{d}|\boldsymbol{\theta}) = \ln p(\mathbf{x}|\mathbf{d}, \boldsymbol{\theta}) + \ln p(\mathbf{d}|\boldsymbol{\theta}),$$

  where we use the following distribution functions,

  $$\begin{aligned} p(\mathbf{x}|\mathbf{d}, \boldsymbol{\theta}) &\sim \mathcal{N}(\bar{\mathbf{X}}\mathbf{T}\mathbf{d}, \mathbf{B}), \\ p(\mathbf{d}|\boldsymbol{\theta}) &\sim \mathcal{N}(\mathbf{d}, \boldsymbol{\Gamma}). \end{aligned}$$

  Then

  $$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = &\ \text{constant} + \frac{1}{2}\ln\det(\mathbf{B}) \\ &- \frac{1}{2}(\mathbf{x} - \bar{\mathbf{X}}\mathbf{T}\hat{\mathbf{d}})^\top \mathbf{B}(\mathbf{x} - \bar{\mathbf{X}}\mathbf{T}\hat{\mathbf{d}}) \\ &+ \frac{1}{2}\text{tr}\left(\left(\bar{\mathbf{X}}\mathbf{T}\right)^\top \mathbf{B}\bar{\mathbf{X}}\mathbf{T}\boldsymbol{\Sigma}'\right) \\ &+ \frac{1}{2}\ln\det(\boldsymbol{\Gamma}) - \frac{1}{2}\text{tr}\left(\boldsymbol{\Gamma}\left(\boldsymbol{\Sigma}' + \hat{\mathbf{d}}\hat{\mathbf{d}}^\top\right)\right). \end{aligned}$$

- 3. **M-step**. Evaluate $\boldsymbol{\theta}^*$ given by

  $$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ \ Q(\boldsymbol{\theta}, \boldsymbol{\theta}') + \ln p(\boldsymbol{\theta}),$$

  where

  $$\begin{aligned} \ln p(\boldsymbol{\theta}) &= \ln p(\boldsymbol{\gamma}) + \ln p(\boldsymbol{\beta}), \\ \ln p(\gamma_i) &= a\ln(\gamma_i) + b\gamma_i + \text{constant}, \\ \ln p(\beta_i) &= c\ln(\beta_i) + d\beta_i + \text{constant}. \end{aligned}$$

  By setting derivatives to zero, we find the update equations as follows

  $$\begin{aligned} \frac{\partial}{\partial\beta_i} =& \frac{1}{2\beta_i} - \frac{1}{2}(\mathbf{x} - \mathbf{X}\mathbf{T}\hat{\mathbf{d}})_i^2 \\ & - \frac{1}{2}(\mathbf{X}\mathbf{T}\boldsymbol{\Sigma}'(\mathbf{X}\mathbf{T})^\top)_{ii} + \frac{c}{\beta_i} - d = 0, \\ \beta_i =& \frac{1 + 2c}{(\mathbf{x} - \mathbf{X}\mathbf{T}\hat{\mathbf{d}})_i^2 + (\mathbf{X}\mathbf{T}\boldsymbol{\Sigma}'(\mathbf{X}\mathbf{T})^\top)_{ii} + 2d}. \\ \frac{\partial}{\partial\gamma_i} =& \frac{K}{2\beta_i} - \frac{1}{2}\sum_{j=(i-1)K+1}^{iK}(\boldsymbol{\Sigma}' + \hat{\mathbf{d}}\hat{\mathbf{d}}^\top)_{jj} \\ & + \frac{a}{\gamma_i} - b = 0, \\ \gamma_i =& \frac{K + 2a}{\sum_{j=(i-1)K+1}^{iK}(\boldsymbol{\Sigma}' + \hat{\mathbf{d}}\hat{\mathbf{d}}^\top)_{jj} + 2b}. \end{aligned}$$

REFERENCES

[1] Mark G. Hall, Alan V. Oppenheim, and Alan S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, no. 3, pp. 267 – 285, 1983.

[2] D. Rudoy, T.F. Quatieri, and P.J. Wolfe, "Time-varying autoregressions in speech: Detection theory and applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 977–989, May 2011.

[3] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[4] J. Markel and A. Gray, *Linear Prediction of Speech*, Springer-Verlag, 1976.

[5] Louis A Liporace, "Linear estimation of nonstationary signals," *The Journal of the Acoustical Society of America*, vol. 58, no. 6, pp. 1288–1295, 1975.

[6] Y. Grenier, "Time-dependent arma modeling of nonstationary signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 31, no. 4, pp. 899–911, Aug 1983.

[7] M.K. Tsatsanis and G.B. Giannakis, "Time-varying system identification and model validation using wavelets," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3512–3523, Dec 1993.

[8] S.R. Chetupalli and T.V. Sreenivas, "Time varying linear prediction using sparsity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6290–6293.

[9] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1644–1657, July 2012.

[10] R. Giri and B.D. Rao, "Block sparse excitation based all-pole modeling of speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3754–3758.

[11] M.E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, vol. 1, pp. 211–244, 2001.

[12] D.P. Wipf and B.D. Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153 – 2164, aug. 2004.

[13] M. Sundin, S. Chatterjee, M. Jansson, and C.R. Rojas, "Relevance singular vector machine for low-rank matrix sensing," in *Signal Processing and Communications (SPCOM), 2014 International Conference on*, July 2014, pp. 1–5.

[14] M. Sundin, S. Chatterjee, and M. Jansson, "Combined modeling of sparse and dense noise improves bayesian rvm," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, Sept 2014, pp. 1841–1845.