

A Variational EM Method for Pole-Zero Modeling of Speech with Mixed Block Sparse and Gaussian Excitation

Liming Shi, Jesper Kjær Nielsen, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University

Emails: {ls, jkn, jrj, mgc}@create.aau.dk

Abstract—The modeling of speech can be used for speech synthesis and speech recognition. We present a speech analysis method based on pole-zero modeling of speech with mixed block sparse and Gaussian excitation. By using a pole-zero model, instead of the all-pole model, a better spectral fitting can be expected. Moreover, motivated by the block sparse glottal flow excitation during voiced speech and the white noise excitation for unvoiced speech, we model the excitation sequence as a combination of block sparse signals and white noise. A variational EM (VEM) method is proposed for estimating the posterior PDFs of the block sparse residuals and point estimates of modelling parameters within a sparse Bayesian learning framework. Compared to conventional pole-zero and all-pole based methods, experimental results show that the proposed method has lower spectral distortion and good performance in reconstructing of the block sparse excitation.

I. INTRODUCTION

The modeling of speech has important applications in speech analysis [1], speaker verification [2], speech synthesis [3], etc. Based on the source-filter model, speech is modelled as being produced by a pulse train or white noise for voiced or unvoiced speech, which is further filtered by the speech production filter (SPF) that consists of the vocal tract and lip radiation.

All-pole modeling with a least squares cost function performs well for white noise and low pitch excitation. However, for high pitch excitation, it leads to an all-pole filter with poles close to the unit circle, and the estimated spectrum has a sharper contour than desired [4], [5]. To obtain a robust linear prediction (LP), the Itakura-Saito error criterion [6], the all-pole modeling with a distortionless response at frequencies of harmonics [4], the regularized LP [7] and the short-time energy weighted LP [8] were proposed. Motivated by the compressive sensing research, a least 1-norm criterion is proposed for voiced speech analysis [9], where sparse priors on both the excitation signals and prediction coefficients are utilized. Fast methods and the stability of the 1-norm cost function for spectral envelope estimation are further investigated in [10], [11]. More recently, in [12], the excitation signal of speech is formulated as a combination of block sparse and white noise components to capture the block sparse or white

noise excitation separately or simultaneously. An expectation-maximization (EM) algorithm is used to reconstruct the block sparse excitation within a sparse Bayesian learning (SBL) framework [13].

A problem with the all-pole model is that some sounds containing spectral zeros with voiced excitation, such as nasals, or laterals, are poorly estimated by an all-pole model but trivial with a pole-zero (PZ) model [14], [15]. The estimation of the coefficients of the pole-zero model can be obtained separately [16], jointly [17] or iteratively [18]. A 2-norm minimization criterion with Gaussian residuals assumption is commonly used. Frequency domain fitting methods based on a similarity measure is also proposed. Motivated by the logarithmic scale perception of the human auditory system, the logarithmic magnitude function minimization criterion has been proposed [15], [19]. In [19], the nonlinear logarithm cost function is solved by transforming it into a weighted least squares problem. The Gauss-Newton and Quasi-Newton methods for solving it are further investigated in [15]. To consider both the voiced excitation and the PZ model, a speech analysis method based on the PZ model with sparse excitation in noisy conditions is presented [20]. A least 1-norm criterion is used for the coefficient estimation, and sparse deconvolution is applied for deriving sparse residuals.

In this paper, we propose a speech analysis method based on the PZ model with mixed excitation. Using the mixed excitation and PZ modeling together, we combine the advantages of non-sparse and sparse algorithms, and obtain a better fitting for both the excitation and SPF spectrum. Using the PZ model, instead of the all-pole model, a better spectral fitting can be expected. Moreover, we model both the voiced, the unvoiced excitation or a mixture of them by the mixed excitation. Additionally, block sparsity is imposed on the voiced excitation component, motivated by the quasi-periodic and temporal-correlated nature of the glottal excitation [12], [21]. The posterior probability density functions (PDFs) for the sparse excitation and hyper-parameters, as well as point estimates of the PZ model parameters are obtained using the VEM method.

II. SIGNAL MODELS

Consider the following general speech observation model:

$$y(n) = s(n) + u(n), \quad (1)$$

This work was funded by the Danish Council for Independent Research, grant ID: DFF 4184-00056

where $y(n)$ is the observation signal and $u(n)$ denotes the noise. We assume that the clean speech signal $s(n)$ is produced by the PZ speech production model, i.e.,

$$s(n) = -\sum_{k=1}^K a_k s(n-k) + \sum_{l=0}^L b_l e(n-l) + m(n), \quad (2)$$

where a_k and b_l are the modeling coefficients of the PZ model with $b_0 = 1$, $e(n)$ is a sparse excitation corresponding to the voiced part and $m(n)$ is the non-sparse Gaussian excitation component corresponding to the unvoiced part. Assuming $s(n) = 0$ for $n \leq 0$ and considering one frame of speech signals of N samples, (1) and (2) can be written in matrix forms as

$$\mathbf{y} = \mathbf{s} + \mathbf{u}, \quad (3)$$

$$\mathbf{A}\mathbf{s} = \mathbf{B}\mathbf{e} + \mathbf{m}, \quad (4)$$

where \mathbf{A} and \mathbf{B} are the $N \times N$ lower triangular Toeplitz matrices with $[1, a_1, a_2, \dots, a_K, 0, \dots, 0]$ and $[1, b_1, b_2, \dots, b_L, 0, \dots, 0]$ as the first columns, respectively. The block sparse residuals are defined as $\mathbf{e} = [e(1) \dots e(N)]^T$, and \mathbf{m} , \mathbf{s} , \mathbf{y} and \mathbf{u} are defined similarly to \mathbf{e} . When $L = 0$, \mathbf{B} reduces to the identity matrix and (4) becomes the all-pole model. Combining (3) and (4), the noisy observation can be written as

$$\mathbf{A}\mathbf{y} = \mathbf{B}\mathbf{e} + \mathbf{m} + \mathbf{A}\mathbf{u}. \quad (5)$$

In [20], we assumed that only the sparse excitation was present ($\mathbf{m} = \mathbf{0}$, but $\mathbf{u} \neq \mathbf{0}$). The sparse residuals and model parameters were estimated iteratively. The sparse residuals were obtained by solving

$$\min_{\mathbf{e}} \|\mathbf{e}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}^{-1}\mathbf{B}\mathbf{e}\|_2^2 \leq C. \quad (6)$$

where C is a constant proportional to the variance of the noise. The model parameters was estimated using the l_1 norm of the residuals as the cost function (see [20] for details).

III. PROPOSED VARIATIONAL EM METHOD

We now proceed to consider the noise-free scenario but with mixed excitation ($\mathbf{u} = \mathbf{0}$, but $\mathbf{m} \neq \mathbf{0}$). We consider the pole-zero model parameters $\mathbf{a} = [a_1, a_2, \dots, a_K]^T$ and $\mathbf{b} = [b_1, b_2, \dots, b_L]^T$ to be deterministic but unknown. Utilizing the SBL [13] methodology, we first express the hierarchical form of the model as

$$\begin{aligned} \mathbf{A}\mathbf{y} &= \mathbf{B}\mathbf{e} + \mathbf{m}, & \mathbf{m} &\sim \mathcal{N}(\mathbf{0}, \gamma_m^{-1}\mathbf{I}_N), \\ \mathbf{e} &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}_e^{-1}), & \gamma_m &\sim \Gamma(c, d), & \boldsymbol{\alpha} &\sim \prod_{o=1}^O \Gamma(\alpha_o; e, f), \end{aligned} \quad (7)$$

where O is the number of blocks, $\mathbf{\Gamma}_e = \text{diag}(\boldsymbol{\alpha}) \otimes \mathbf{I}_D$, \otimes is the Kronecker product, D is the block size, $N = DO$, \mathcal{N} denotes the multivariate normal PDF and Γ is the Gamma PDF. The hyperparameter α_o is the precision of the o^{th} block, and when it is infinite, the o^{th} block will be zero. Note that it is trivial to extend the proposed method to any D . Moreover, when $D = 1$, each element in \mathbf{e} is inferred independently.

Here, block sparsity model is used to take the quasi-periodic and temporal-correlated nature of the voiced excitation into account. The \mathbf{m} is used for capturing the white noise excitation from unvoiced speech frame or a mixture of phonations.

Our objective is to obtain the posterior densities of \mathbf{e} , γ_m and $\boldsymbol{\alpha}$, and point estimates of the model parameters in \mathbf{a} and \mathbf{b} . First, we write the complete likelihood, i.e.,

$$\begin{aligned} p(\mathbf{y}, \mathbf{e}, \boldsymbol{\alpha}, \gamma_m) &= p(\mathbf{y}|\mathbf{e}, \gamma_m)p(\mathbf{e}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\gamma_m) \\ &= \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{e}|\mathbf{0}, \mathbf{\Gamma}_e^{-1}) \\ &\quad \times \Gamma(\gamma_m; c, d) \prod_{o=1}^O \Gamma(\alpha_o; e, f), \end{aligned} \quad (8)$$

where we used $\mathcal{N}(\mathbf{y}|\mathbf{A}^{-1}\mathbf{B}\mathbf{e}, \gamma_m^{-1}(\mathbf{A}^T\mathbf{A})^{-1}) = \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)$ when $\det(\mathbf{A}) = 1$. Instead of finding the joint posterior density $p(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m|\mathbf{y})$, which is intractable, we adopt the variational approximation [22]. Assume that $p(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m|\mathbf{y})$ is approximated by the density $q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m)$, which may be fully factorized as

$$q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m) = q(\mathbf{e})q(\gamma_m) \prod_{o=1}^O q(\alpha_o), \quad (9)$$

where the factors are found using an EM-like algorithm [22].

In the E-step of the VEM method, we fix the model parameters \mathbf{a} and \mathbf{b} , and re-formulate the posterior PDFs estimation problem as maximizing the variational lower bound

$$\max_q \mathbb{E}_q[\log p(\mathbf{y}, \mathbf{e}, \boldsymbol{\alpha}, \gamma_m)] + H[q], \quad (10)$$

where q is the shorthand for $q(\mathbf{e}, \boldsymbol{\alpha}, \gamma_m)$, $H[q]$ is defined as $H[q] = -\mathbb{E}_q[\log(q)]$, and $\mathbb{E}_{q(x)}[f(x)]$ denotes the expectation of $f(x)$ w.r.t. the random variable x (i.e., $\mathbb{E}_{q(x)}[f(x)] = \int f(x)q(x)dx$). Substituting (8) and (9) into (10), and following the derivation from [22], we obtain

$$\begin{aligned} q(\mathbf{e}) &\propto e^{\mathbb{E}_{q(\boldsymbol{\alpha}, \gamma_m)}[\log \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)\mathcal{N}(\mathbf{e}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha})^{-1} \otimes \mathbf{I}_D)]}, \\ q(\alpha_o) &\propto \Gamma(\alpha_o; e, f) e^{\mathbb{E}_{q(\mathbf{e})}[\log \mathcal{N}(\mathbf{e}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha})^{-1} \otimes \mathbf{I}_D)]}, \quad 1 \leq o \leq O, \\ q(\gamma_m) &\propto \Gamma(\gamma_m; c, d) e^{\mathbb{E}_{q(\mathbf{e})}[\log \mathcal{N}(\mathbf{A}\mathbf{y}|\mathbf{B}\mathbf{e}, \gamma_m^{-1}\mathbf{I}_N)]}. \end{aligned} \quad (11)$$

It is clearly seen that $q(\mathbf{e})$ in (11) is a Gaussian PDF, i.e.,

$$q(\mathbf{e}) = \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \quad (12)$$

where $\tilde{\boldsymbol{\Sigma}} = (\mathbb{E}[\gamma_m]\mathbf{B}^T\mathbf{B} + \mathbb{E}[\mathbf{\Gamma}_e])^{-1}$ and $\tilde{\boldsymbol{\mu}} = \mathbb{E}[\gamma_m]\tilde{\boldsymbol{\Sigma}}\mathbf{B}^T\mathbf{A}\mathbf{y}$. We also define the auto-correlation matrix $\tilde{\mathbf{R}} = \tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^T$. The posterior PDF of α_o in (11) is a Gamma probability density, i.e.,

$$q(\alpha_o) = \Gamma(\tilde{e}_o, \tilde{f}_o), \quad (13)$$

where $\tilde{e}_o = e + D/2$, $\tilde{f}_o = f + \sum_{i=(o-1)D+1}^{oD} \tilde{\mathbf{R}}_{i,i}/2$ and $\tilde{\mathbf{R}}_{i,i}$ denotes the (i, i) element of $\tilde{\mathbf{R}}$. The expectation of the precision matrix is $\mathbb{E}[\mathbf{\Gamma}_e] = \text{diag}(\tilde{e}_1/\tilde{f}_1, \dots, \tilde{e}_O/\tilde{f}_O) \otimes \mathbf{I}_D$. Similar to α_o , the posterior PDF of γ_m is

$$q(\gamma_m) = \Gamma(\tilde{c}, \tilde{d}), \quad (14)$$

where $\tilde{c} = c + N/2$, $\tilde{d} = d + (\text{tr}(\tilde{\Sigma}\mathbf{B}^T\mathbf{B}) + \|\mathbf{A}\mathbf{y} - \mathbf{B}\tilde{\boldsymbol{\mu}}\|_2^2)/2$. The expectation of γ_m can be expressed as $\mathbb{E}[\gamma_m] = \tilde{c}/\tilde{d}$.

In the M-step, we maximize the lower bound (10) w.r.t. the modeling parameters \mathbf{a} and \mathbf{b} , respectively. The optimization problems can be shown to be equivalent to $\min_{\mathbf{a}} \mathbb{E}_{q(\mathbf{e})} \|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$ and $\min_{\mathbf{b}} \mathbb{E}_{q(\mathbf{e})} \|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$, respectively. To obtain the estimate for \mathbf{a} , we first note that $\mathbf{A}\mathbf{y}$ can be expressed as $\mathbf{A}\mathbf{y} = \mathbf{C}\mathbf{a} + \mathbf{y}$, where \mathbf{C} is a $N \times K$ Toeplitz matrix of the form

$$\mathbf{C} = \begin{bmatrix} 0 & \cdots & 0 \\ y(1) & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \vdots & & y(1) \\ \vdots & & \vdots \\ y(N-1) & \cdots & y(N-K) \end{bmatrix}_{N \times K}$$

Using this expression and $q(\mathbf{e})$ obtained in the E-step, the minimization problem w.r.t. \mathbf{a} can be re-formulated as

$$\begin{aligned} \min_{\mathbf{a}} \mathbb{E}_{q(\mathbf{e})} \|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2 &\iff \min_{\mathbf{a}} \|\mathbf{A}\mathbf{y} - \mathbf{B}\tilde{\boldsymbol{\mu}}\|_2^2 + \text{tr}(\tilde{\Sigma}\mathbf{B}^T\mathbf{B}) \\ &\iff \min_{\mathbf{a}} \|(\mathbf{B}\tilde{\boldsymbol{\mu}} - \mathbf{y}) - \mathbf{C}\mathbf{a}\|_2^2. \end{aligned} \quad (15)$$

As can be seen, (15) is the standard least squares problem and has the analytical solution as

$$\mathbf{a} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T(\mathbf{B}\tilde{\boldsymbol{\mu}} - \mathbf{y}). \quad (16)$$

We can obtain the solution of \mathbf{b} , like \mathbf{a} , by setting the derivative of $\mathbb{E}_{q(\mathbf{e})} \|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2$ w.r.t. \mathbf{b} to zero, i.e.,

$$\begin{aligned} \frac{\partial \mathbb{E}_{q(\mathbf{e})} \|\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e}\|_2^2}{\partial \mathbf{b}} &= \mathbb{E}_{q(\mathbf{e})} [2\mathbf{F}^T(\mathbf{A}\mathbf{y} - \mathbf{B}\mathbf{e})] \\ &= \mathbf{0}_{L \times 1}. \end{aligned} \quad (17)$$

where \mathbf{F} is an $N \times L$ lower triangular Toeplitz matrix of the form

$$\mathbf{F} = \begin{bmatrix} 0 & \cdots & 0 \\ e(1) & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \vdots & & e(1) \\ \vdots & & \vdots \\ e(N-1) & \cdots & e(N-L) \end{bmatrix}_{N \times L}$$

From (17), we obtain the estimate of \mathbf{b} , i.e.,

$$\mathbf{b} = (\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{F}])^{-1}(\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T]\mathbf{A}\mathbf{y} - \mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{e}]), \quad (18)$$

where $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{F}]$ is an $L \times L$ symmetric matrix with the $(i, j)^{\text{th}}$, $j \geq i$ element given by $\sum_{k=1}^{N-j} \tilde{\mathbf{R}}_{k, k+j-i}$. The $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T]$ can be obtained by simply replacing the stochastic variable $e(n)$, $1 \leq n \leq N-1$ in \mathbf{F}^T with the mean estimate $\tilde{\mu}(n)$ (the n^{th} element in $\tilde{\boldsymbol{\mu}}$). The $\mathbb{E}_{q(\mathbf{e})}[\mathbf{F}^T\mathbf{e}]$ is an $L \times 1$ vector with the l^{th} element given by $\sum_{k=1}^{N-l} \tilde{\mathbf{R}}_{k, k+l}$. Note that the estimation of \mathbf{b} in (18) requires the knowledge of \mathbf{a} and vice

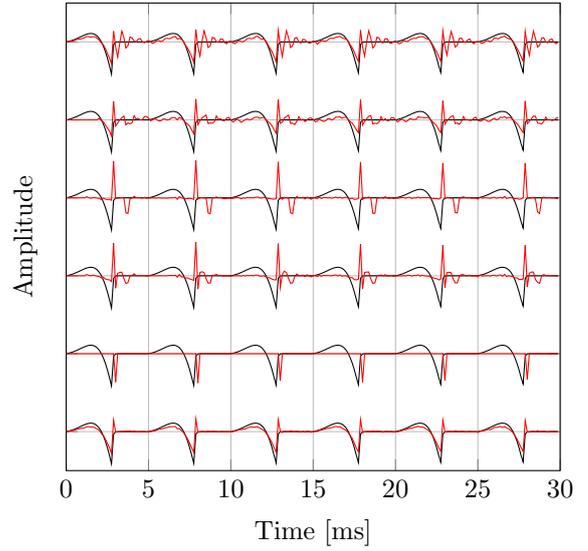


Fig. 1: residuals estimate for synthetic /n/. The black line is the LF excitation. The red lines shown correspond to, from top to bottom, residuals of (1) 2-norm LP, (2) TS-LS-PZ, (3) 1-norm LP, (4) EM-LP, D=8, (5) VEM-PZ, D=1, (6) VEM-PZ, D=8.

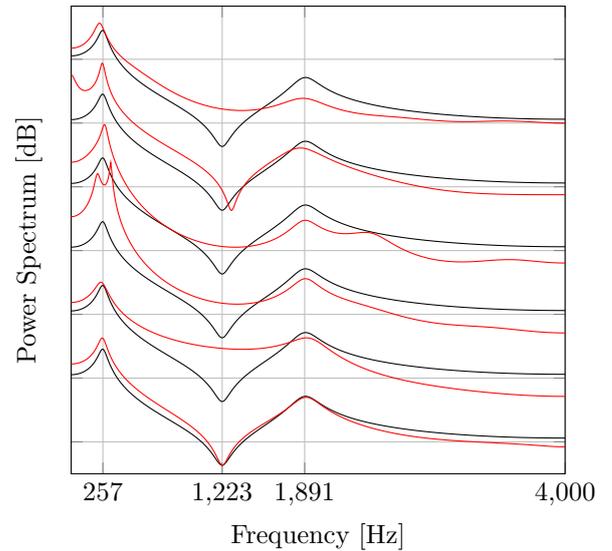


Fig. 2: corresponding spectral estimates for synthetic /n/. The red lines have the same setting as Fig. 1.

versa (see (16)). This coupling is solved by replacing them with their estimates from previous iteration. The algorithm is initialized with $\mathbf{a} = [1, 0, \dots, 0_K]^T$, $\mathbf{b} = [1, 0, \dots, 0_L]^T$, $\gamma_m = 10$ and $\alpha_o = 1, o = 1, \dots, O$, and starts with the update of \mathbf{e} . We refer to the proposed variational expectation maximization pole-zero estimation algorithm as the VEM-PZ.

IV. RESULTS

In this section, we compare the performance of the proposed VEM-PZ, the two-stage least squares pole-zero (TS-LS-PZ)

TABLE I: The spectral distortion

F0	200	250	300	350	400
2-norm LP	1.79	2.14	2.12	2.53	2.13
TS-LS-PZ	2.41	4.77	1.88	1.46	2.86
1-norm LP	2.43	3.15	3.60	3.29	4.29
EM-LP	5.62	6.68	4.68	3.96	4.83
VEM-PZ, D=1	4.50	7.14	2.29	1.54	2.31
VEM-PZ, D=5	1.55	4.47	0.69	2.01	4.50
VEM-PZ, D=7	2.08	4.07	2.18	1.41	1.29
VEM-PZ, D=8	0.77	5.56	2.52	4.86	0.53

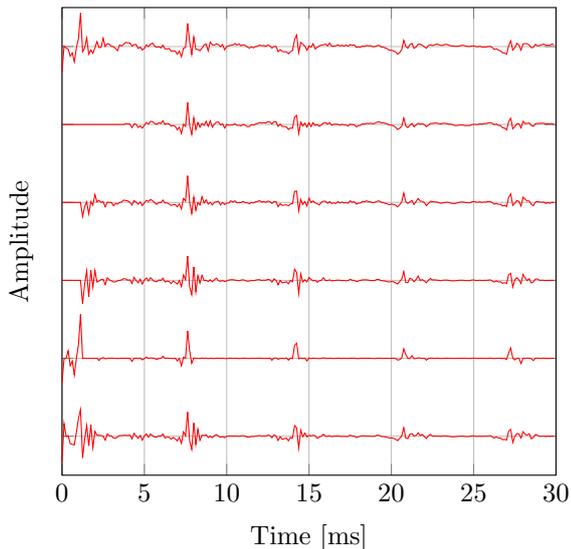


Fig. 3: residuals estimate for /n/ in the word "manage". The red lines have the same setting as Fig. 1.

method [14], 2-norm linear prediction (2-norm LP) [1], 1-norm linear prediction (1-norm LP) [9] and expectation maximization based linear prediction (EM-LP) for mixed excitation [12] in both synthetic and real speech signals analysis scenarios.

A. Synthetic signal analysis

We first examine the performance of the VEM-PZ with different block size D and compares it with traditional algorithms using synthetic voiced consonant /n/, as shown in Fig. 1 and Fig. 2. The synthetic signals are generated by convolving an artificial glottal source waveform with a constructed filter. The first derivative of the glottal flow pulse is simulated with the Liljencrants-Fant (LF) waveform [23] with the modal phonation mode, whose parameter is taken from Gobl [24]. The voiced alveolar nasal consonant /n/ is synthesized at 8 kHz sampling frequency with the constructed filter having two formant frequencies (bandwidths) of 257 Hz (32 Hz) and 1891 Hz (100 Hz) and one antiformant of 1223 Hz (52 Hz) [25]. N is set to 240 (30 ms), $e = 1$ and $c = d = f = 10^{-6}$ are used for all the experiments. The power ratio of the block sparse excitation e and Gaussian component m is set to 30 dB, and the pitch is set to 200 Hz. The K and L are set to 5 for the pole-zero modeling methods (i.e., VEM-PZ and

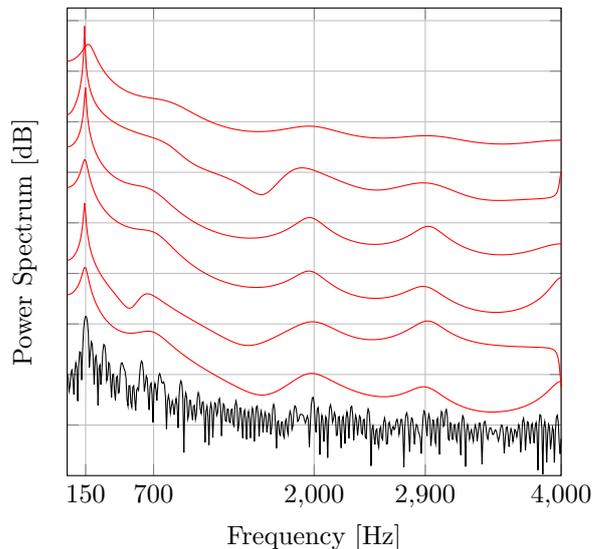


Fig. 4: corresponding spectral estimates for /n/ in the word "manage". The red lines have the same setting as Fig. 1. The black line is the periodogram.

TS-LS-PZ), but $K = 10$ is used for the all-pole modeling methods (i.e., 2-norm LP, 1-norm LP and EM-LP). In Fig. 1, the means of the residuals of EM-LP and VEM-PZ are plotted. Note that the residuals of the TS-LS-PZ, 1-norm LP and EM-LP are prepended with zeros due to the covariance-based estimation methods. As can be seen in Fig. 1, the residual estimate of the proposed VEM-PZ with $D = 8$ is closest to the true block sparse excitation. Moreover, when $D = 1$, the residuals of the VEM-PZ are the sparsest as expected. Although the EM-LP also produces block sparse residuals compared with the 1-norm LP, the proposed method achieves the best approximation to the true one due to the usage of the pole-zero modeling and the block-sparse motivated SBL method. The corresponding spectral estimates are shown in Fig. 2. First, as can be seen, the VEM-PZ with $D = 1$ has a smooth power spectrum due to the sparse residuals in Fig. 1. Second, the 1-norm LP tends to produce a better estimate of the formants than the 2-norm LP and TS-LS-PZ. Third, although the EM-LP has two peaks around the first formant, it performs well for second formant estimation. Finally, the proposed VEM-PZ with $D = 8$ has good performance for both formant, antiformant and bandwidth estimates because of the block sparse excitation and the pole-zero model.

Second, the spectral distortion is tested under different fundamental frequencies and block sizes. The measure is defined as the truncated power cepstral distance [26], i.e.,

$$d_{\text{ceps}} = \sum_{n=-S}^S (c_n - \hat{c}_n)^2, \quad (19)$$

where c_n and \hat{c}_n are the true and estimated power cepstral coefficients, respectively. Cepstral coefficients are computed by first reflecting all the poles and zeros to the inside of the unit circle and then using the recursive relation in [25]. In

our experiments, we set $S = 300$. The fundamental frequency rises from 200 to 400 Hz in steps of 50 Hz. The experimental results in TABLE I are obtained by the ensemble averages over 500 Monte Carlo experiments. $D = 6$ is used for the EM-LP [12]. As can be seen from TABLE I, the 2-norm LP has a lower spectral distortion than the 1-norm LP, EM-LP and TS-LS-PZ (except for 300 and 350 Hz). The proposed VEM-PZ achieves the lowest spectral distortion for 200, 300, 350 and 400 Hz. However, note that the good performance of the VEM-PZ depends on a good choice of the block sizes for different fundamental frequencies, and there is a fluctuation when the frequency changes. This is because the length of correlated samples changes with the fundamental frequency. But, as can be seen from Fig. 2 and from our experience, the VEM-PZ usually produces better formant, antiformant and bandwidth estimates than traditional ones.

B. Speech signal analysis

We examine the residuals and spectral estimate of the VEM-PZ for a nasal consonant /n/ in the word "manage" from the CMU Arctic database [27], [28], pronounced by an US female native speaker, sampled of 8000 Hz. The results are shown in Fig. 3 and Fig. 4. To improve the modeling flexibility, the K and L are set to 10 for the PZ methods (i.e., VEM-PZ and TS-LS-PZ), but $K = 10$ is still used for the all-pole methods (i.e., 2-norm LP, 1-norm LP and EM-LP). As can be seen from Fig. 3, the residuals of the EM-LP and 1-norm LP are sparser than the 2-norm LP. The residuals of the proposed VEM-PZ with $D = 1$ are the sparsest. The proposed VEM-PZ with $D = 8$ is block sparse and is sparser than the TS-LS-PZ. From Fig. 4, we can see that all the algorithms have formant estimates around 150 Hz. However, the TS-LS-PZ, 1-norm LP and VEM-PZ with $D = 1$ have very peaky behaviour. Also, the 2-norm LP produces bad bandwidth estimates around 2000 and 2900 Hz. Furthermore, compared to the EM-LP, the proposed VEM-PZ with $D=8$ has good antiformant estimates around 500 and 1500 Hz.

V. CONCLUSION

A variational expectation maximization pole-zero speech analysis method has been proposed. By using the pole-zero model, it can fit the spectral zeros of speech signals easily. Moreover, block sparse residuals are encouraged by applying the sparse Bayesian learning method. By iteratively updating parameters and statistics of residuals and hyperparameters, block sparse residuals can be obtained. The good performance has been verified by both synthetic and real speech experiments. The proposed method is promising for speech analysis applications. Next, further research into the formant, antiformant and bandwidth estimation accuracy, stability, and unknown sparse pattern should be conducted.

REFERENCES

- [1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.
- [2] J. Pohjalainen, C. Haniçci, T. Kinnunen, and P. Alku, "Mixture linear prediction in speaker verification under vocal effort mismatch," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1516–1520, dec 2014.
- [3] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.
- [4] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 221–239, 2000.
- [5] T. Drugman and Y. Stylianou, "Fast inter-harmonic reconstruction for spectral envelope estimation in high-pitched voices," *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1418–1422, 2014.
- [6] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.
- [7] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," vol. 16, no. 1, pp. 65–73, 2008.
- [8] P. Alku, J. Pohjalainen, M. Vainio, A. M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1295–1313, August 2013.
- [9] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," vol. 20, no. 5, pp. 1644–1657, jul 2012.
- [10] D. Giacobello, M. G. Christensen, T. L. Jensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Stable 1-norm error minimization based linear predictors for speech modeling," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 5, pp. 912–922, may 2014.
- [11] T. L. Jensen, D. Giacobello, T. V. Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Commun.*, vol. 76, pp. 143–156, 2016.
- [12] R. Giri and B. D. Rao, "Block sparse excitation based all-pole modeling of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014.
- [13] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 2001, pp. 211–244, 2001.
- [14] P. Stoica and R. Moses, *Spectral Analysis of Signals*. New Jersey: Prentice Hall, Upper Saddle River, 2004.
- [15] D. Marelli and P. Balazs, "On pole-zero model estimation methods minimizing a logarithmic criterion for speech analysis," vol. 18, no. 2, pp. 237–248, 2010.
- [16] J. Durbin, "The fitting of time-series models," *Rev. Int'l Statistical Inst.*, vol. 28, no. 3, pp. 233–244, 1960.
- [17] E. Levy, "Complex-curve fitting," *IRE Trans. Automat. Contr.*, vol. AC-4, no. 1, pp. 37–43, 1959.
- [18] K. SteTiange, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.
- [19] T. Kobayashi and S. Imai, "Design of IIR digital filters with arbitrary log magnitude function by WLS techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 247–252, 1990.
- [20] L. Shi, J. R. Jensen, and M. G. Christensen, "Least 1-norm pole-zero modeling with sparse deconvolution for speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, to be published.
- [21] P. Alku, "Glottal inverse filtering analysis of human voice production," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, 2011.
- [22] C. M. Bishop, *Pattern recognition and machine learning*, M. Jordan, J. Kleinberg, and B. Scho, Eds. Springer, 2006.
- [23] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR4 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden)*, pp. 1–13, 1985.
- [24] C. Gobl, "A preliminary study of acoustic voice quality correlates," *STL-QPSR4 (Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden)*, pp. 9–22, 1989.
- [25] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012.
- [26] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.
- [27] J. Kominek and A. Black, "The CMU Arctic speech databases," in *5th ISCA Speech Synthesis Workshop*. Pittsburgh, PA, 2004.
- [28] CMU-ARCTIC Speech Synthesis Databases. [Online]. Available: http://festvox.org/cmu_arctic/index.html