BLOCK SPARSE EXCITATION BASED ALL-POLE MODELING OF SPEECH

Ritwik Giri and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
rgiri@ucsd.edu, brao@ucsd.edu

ABSTRACT

In this paper, it is shown that an appropriate model for voiced speech is an all-pole filter excited by a block sparse excitation sequence. The modeling approach is generalized in a novel manner to deal with a wide spectrum of speech signal; voiced speech, unvoiced speech and mixed excitation speech. In this context, the input sequence to the all-pole model is modeled as a suitable weighted linear combination of a block sparse signal and white noise. We develop the corresponding estimation procedure to reconstruct the generalized input sequence and model parameters via sparse Bayesian learning methods employing the Expectation-Maximization based procedure. Rigorous experiments have been performed to show the efficacy of our proposed model for the speech modeling task. By imposing a block sparse structure on the input sequence, the problems associated with the commonly used Linear Prediction approach is alleviated leading to a more robust modeling scheme.

Index Terms— Deconvolution, speech, Sparse bayesian Learning, Expectation-Maximization

1. INTRODUCTION

In speech modeling, an all pole model is most commonly used to model the vocal tract. Depending on the nature of the utterance, voiced, unvoiced or mixed, the input to the all-pole filter is either a glottal pulse train, white noise, or a combination of glottal pulses and white noise respectively. Estimation of the model parameters has a long history and a popular approach is the linear prediction (LP) based all pole model parameters estimation which involves minimizing the 2-norm of the residual, the difference between the observed signal and the predicted signal. The residual signal in all pole modeling is the input excitation sequence. Because of the 2-norm minimization approach, such estimation methods work well for unvoiced speech where the input to the filter is white noise. The 2-norm minimization based linear prediction approach suffers from some well known problems [1] in the case of voiced speech. The spectrum of the resulting model tends to overestimate the spectral powers at the formant frequencies, providing a sharper contour than the original vocal tract response. Several different methods have been proposed to alleviate these effects. Some of the proposed techniques involve a general rethinking of the spectral modeling problem [2, 3] while some others are based on changing the statistical assumptions made on the prediction error in the minimization process [4, 5]. Recently, instead of minimizing the 2-norm of the residual, methods based on minimizing the one norm of the residual, to accommodate the spike train nature of the input sequence, have been suggested with some success for voiced speech [6]. Interesting algorithms [6, 7] based on reweighted l_1 approaches have been employed to exploit the sparsity assumption on the input process.

In case of voiced speech, the excitation can be considered to be a sparse excitation of a quasi-periodic nature [8]. The excitation component of the voiced speech production model is known as the glottal excitation. The structure of this glottal excitation has been an interesting topic of research for several years. From Figure 1 the temporal extent of the glottal pulses show that a block sparse structure is more appropriate. Thus to make the voiced speech modeling task more robust and efficient we propose a framework where the excitation has a prior block sparse quasi-periodic structure. It is useful to note that block sparsity has been studied before in the context of sparse signal recovery, but they are usually for underdetermined problems and the block sparsity is imposed on the solution vector [9], not on the residual as discussed here. The model is then generalized to deal with the broad spectrum of speech signals. In our proposed model the residual is modeled as being a linear combination of two components: a block sparse component and a Gaussian i.i.d white noise component. By appropriately weighting the components, this model for the input can deal with all speech utterances; voiced, unvoiced speech and mixed excitation speech.

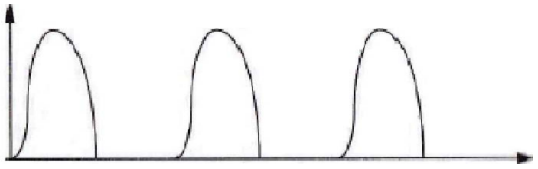

Fig. 1. Shape of Glottal Excitation

The rest of the paper is organized in the following way. Section 2 presents the model and discusses its advantages and disadvantages and Section 3 provides a detailed description of the estimation procedure of the parameters. Section 4 summarizes the performance of the proposed model over synthetic data, and Section 5 presents the results of the speech modeling problem over the Vowel dataset and finally Section 6 concludes the paper.

2. PROPOSED MODEL

Since we are modeling the vocal tract using all-pole models, we will consider the signal to have been generated by an all-pole filter excited by an appropriate input, either block sparse, white noise or a combination. The all-pole model parameters and the nature of excitation input sequence are not known before hand. For instance, in speech this depends on the utterance. This production model can be described by the following difference equation,

$$x(n) = \sum_{k=1}^{M} a_k x(n-k) + w(n) + e(n) \tag{1}$$

Thus $x(n)$ is written as a linear combination of past M samples. Here a_k are the model parameters and $w(n)$ is the block sparse excitation sequence, whereas $e(n)$ is the non sparse white noise component. Now considering this production model for a segment of sample length N, for n=1 to N, we can represent this model in matrix form as,

$$\mathbf{Y} = X\mathbf{a} + \mathbf{w} + \mathbf{e} \tag{2}$$

Where, $\mathbf{Y} = [x(M+1), x(M+2) \cdots x(N)]^T$, X is the known data matrix which is constructed from the known time series data. A pictorial representation of this model is shown in Figure 2. The main idea behind this model is that $\mathbf{w}$ will capture the (block) sparse excitation and $\mathbf{e}$ will capture the standard non-sparse Gaussian excitation and provide a richer class of excitation sequences and richer class of models. In the context of speech, by appropriate weighting of these components we have the ingredients to deal with all types of speech signals. For voiced speech, $\mathbf{w}$ will dominate the residual. For unvoiced speech, $\mathbf{e}$ will dominate the residual. For mixed speech both components would be present at appropriate levels. For the block sparse structure of $\mathbf{w}$, we assume that the all the block sizes are equal and equal to d, and that the blocks are non-overlapping and contiguous, i.e. block boundaries known. Though a more general block structure can be imposed, our experiments indicate that the methods developed work reasonably with a properly chosen block size d.

3. PARAMETER ESTIMATION

To estimate the parameters of our model, we can proceed in two ways. First is a deterministic setting where an extension

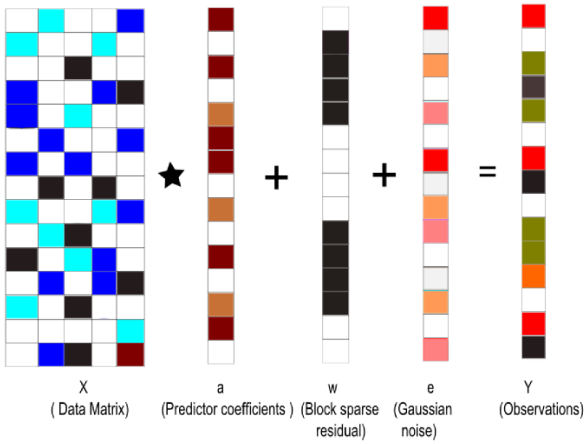

Fig. 2. Pictorial Representation of the proposed model

of the l_1 norm is considered such as a mixed norm l_1/l_2 norm, i.e. minimizing the l_1 norm of the l_2 norm of the blocks. In our work, we have chosen a probabilistic setting by adopting the empirical Bayes approach because of its flexibility and it also readily allows this type of two component noise modeling technique [10]. In particular, we utilize the Sparse Bayesian learning (SBL) [11] methodology. Detailed analysis of the original SBL for sparse signal recovery have been extensively discussed in several literatures[12] [13]. Interested readers are referred to these references for more details. We will use a standard EM algorithm to estimate the parameters of our model. It is assumes that

$$p(\mathbf{e}) = N(0, \sigma^2 I) \tag{3}$$

Thus,

$$p(\mathbf{Y} - X\mathbf{a}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp - \frac{|\mathbf{Y} - X\mathbf{a} - \mathbf{w}|^2}{2\sigma^2} \tag{4}$$

For this model framework we will assume that the error $\mathbf{w}$ has a normal distribution with mean zero and a block structure of block size d. Under the SBL formulation, the covariance matrix of these error blocks is modeled as $\gamma_i I, i = 1, .., L$. Hence the covariance matrix of the complete error sequence is

$$\Gamma = \text{diag}(\gamma_1 I, ..., \gamma_L I) \tag{5}$$

Here γ_i is the hyperparameter which controls the variance of the i^{th} block and have to be learnt. If $\gamma_i = 0$, it means that the corresponding block will also be zero.

To estimate the values of the parameters $\mathbf{a}, \sigma^2$ and γ_is we will use the EM algorithm and will consider $\mathbf{w}$ as the latent variable. The complete loglikelihood can be written as,

$$L = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}|\mathbf{Y} - X\mathbf{a} - \mathbf{w}|^2 - \frac{N}{2}\log 2\pi - \frac{1}{2}\log(\det(\Gamma)) - \frac{1}{2}\mathbf{w}^\intercal\Gamma^{-1}\mathbf{w}$$

The Q function is defined as,

$$Q = E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{\mathbf{t-1}}}[L] \tag{6}$$

Thus we need to know, $E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{\mathbf{t-1}}}[\mathbf{w}]$ and $E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{\mathbf{t-1}}}[\mathbf{w}^\intercal\mathbf{w}]$

After some simple manipulations we obtain,

$$\begin{aligned}\hat{W}_1 &= E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{\mathbf{t-1}}}[\mathbf{w}] \\ &= (I + \sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1}(\mathbf{Y} - X\mathbf{a_{t-1}})\end{aligned}$$

and,

$$\begin{aligned}\hat{W}_2 &= E_{\mathbf{w}|\mathbf{Y}-X\mathbf{a_{t-1}},\sigma^2_{t-1},\gamma_{\mathbf{t-1}}}[\mathbf{w}^\intercal\mathbf{w}] \\ &= (I + \sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1}(\mathbf{Y} - X\mathbf{a_{t-1}})(\mathbf{Y} - X\mathbf{a_{t-1}})^\intercal \\ &\quad (I + \sigma^2_{t-1}\Gamma^{-1}_{t-1})^{-1} + (\sigma^{-2}_{t-1}I + \Gamma^{-1}_{t-1})^{-1}\end{aligned}$$

In the M-step we will maximize the Q function with respect to our model parameters. So after taking derivative with respect to the parameters and setting them to zero we get,

$$\gamma_i = \frac{1}{d}\sum_{j=(i-1)d+1}^{id} \hat{w_j^2} \text{ where, } \hat{w_j^2} = [\hat{W}_2]_{j,j} \tag{7}$$

$$\sigma^2 = \frac{1}{N}[|\mathbf{Y} - X\mathbf{a}|^2 - 2(\mathbf{Y} - X\mathbf{a})^\intercal\hat{W}_1 + tr(\hat{W}_2)] \tag{8}$$

$$\mathbf{a} = (X^\intercal X)^{-1}X^\intercal(\mathbf{Y} - \hat{W}_1) \tag{9}$$

Hence by using these update rules the parameters of the model can be estimated in each iteration.

4. EXPERIMENTS ON SYNTHETIC DATA

In this section we will discuss the experiments over the synthetic data to validate our above mentioned models. Here, we will use an all pole model that has been obtained after modeling a speech segment using LPC technique, to produce the synthetic speech signal by passing three different types of excitations through it. As we are dealing with block sparse excitations, the period of these block excitation becomes an important factor and this can be viewed as the pitch period. Thus, in the language of speech domain all the experiments have been performed using two pitch frequencies, 100 Hz and 200 Hz. Now as this pitch frequency changes with time in case of speech signals, a little randomization has also been introduced when using this pitch frequency. We did the experiments for two cases where case 1 is $f_1 = 100 + N(0, 9)$ and case 2 is $f_2 = 200 + N(0, 9)$ where $N(0, 9)$ is normal random variable with mean 0 and variance 9. For all these experiments we have used block size= 6 (empirically chosen).

The performance of a spectral envelope estimation method can be measured in many ways. An often used criterion for measuring quality is the spectral distortion between estimated all pole model $S'(\omega, \mathbf{a})$ and the true all pole model $S(\omega)$ which is the ground truth where, $S(\omega) = \frac{1}{|A(e^{j\omega})|^2}$ and $A(e^{j\omega})$ is defined by the filter coefficient vector $(a_0.....a_M)$.

This Spectral Distortion measure is defined as,

$$SD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}[10\log_{10}S(\omega) - 10\log_{10}S'(\omega, \mathbf{a})]^2 d\omega} \tag{10}$$

For a pair of spectra $S(\omega)$ and $S'(\omega, \mathbf{a})$, by applying Parseval's Theorem we can relate the l_2 cepstral distance of the spectra to the previously defined log spectral distortion,

$$SD^2 = \sum_{n=-\infty}^{\infty}(c_n - c'_n)^2 \tag{11}$$

For these experiments over synthetic data, cepstral coefficients are determined from the all pole model coefficients using the recursive relation [14] and the spectral distortion is measured using the above mentioned cepstral distance of the spectra. For three different types of input signals these experiments are performed. (Input 1: Block sparse signal, Input 2: Block sparse signal plus additive white Gaussian noise, Input 3: white Gaussian noise)

In Table 1 the spectral distortion measures are tabulated, using the mean of 200 frames of these three input signals.

Table 1. Spectral Distortion Measure over synthetic data

Inputs	Frequency	Std of noise	Spectral Distortion	
			Proposed Model	LPC
Input1	100 Hz		**1.0484**	1.0651
	200 Hz		**1.0279**	1.0660
Input2	100 Hz	0.1	**0.6155**	0.7010
		0.4	**0.2814**	0.3541
		0.6	**0.2776**	0.2989
	200 Hz	0.1	**0.6562**	0.8363
		0.4	**0.3639**	0.4320
		0.6	**0.3019**	0.3069
Input3		0.2	0.2683	**0.2432**

From the results shown in Table 1 it is evident that our proposed modeling method is very effective for voiced and mixed excitation signals.

5. EXPERIMENTS OVER VOWEL DATASET

As discussed before, our proposed model can deal with all the aspects of speech: voiced, unvoiced and also the mixed excitations. The experiments in the previous section using synthetic data also endorses our claim. So in this section we will continue our experiments over the vowel dataset using the proposed model and we will compare the performance of

our model with widely used LPC speech modeling technique. This dataset has audio recording of 12 Vowels i.e /i/, /I/, /ϵ/, /æ/, /ʌ/, /a/, /ɔ/, /U/, /u/, /ɜ/, /e/, /o/ spoken by a male speaker. The sampling frequency is 16 Khz. Speech signals are quasi-stationary, so they are divided into segments within which the signal can be regarded as stationary. We will use a 20 ms window as each segment, hence it will consist of 320 samples. All pole model of order(M)=20 has been used to model each of these segments. The spectral distortion measure for each vowel is computed as the mean over all the speech segments of that vowel. For both the models, the spectral distortion measure for each vowel is tabulated in Table 2. For 8 cases out of 12 vowels, our model performs better than well known LPC technique in terms of spectral distortion measure. Figure 3 shows the estimated envelopes using both the models along with the periodogram of a speech segment of vowel /a/. One can observe that the modeling technique results in formants that do not have the peaky behavior, LPC techniques are known to suffer from.

Table 2. Spectral Distortion Measure over Vowel data

Vowels	Models	
	Proposed Model	LPC
/i/	**4.1492**	4.6053
/I/	4.0753	**4.0511**
/ϵ/	4.0985	**3.8473**
/æ/	**3.7462**	3.8677
/ʌ/	**4.4092**	4.4179
/a/	**3.2895**	3.4036
/ɔ/	5.2601	**5.2598**
/U/	**4.6470**	4.8754
/u/	5.7985	**5.6795**
/ɜ/	**4.8576**	5.0481
/e/	**3.6325**	3.6431
/o/	**5.0795**	5.1003

6. CONCLUSION

In this paper, we have proposed a novel model to reconstruct block sparse excitation from the output of an all pole filter. We have used our model for the speech modeling task and the spectral distortion measure of the estimated envelope establishes our claim, that this is a more generalized and efficient modeling approach than linear prediction. As this problem is closely related to a more general deconvolution problem, applying these models in several other applications along with theoretically establishing the optimality of this model will be the direction of the future works.

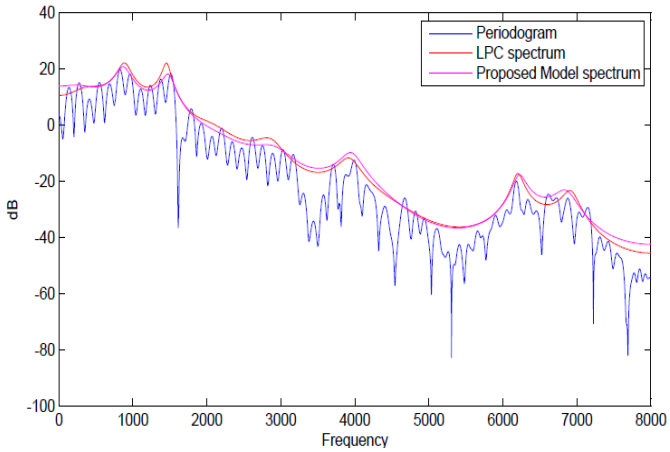

Fig. 3. Spectrum of a segment of vowel /a/

Acknowledgment

This research was supported by the National Science Foundation grant CCF-1144258. Also thanks to the reviewers for their useful comments.

7. REFERENCES

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

[2] Manohar N Murthi and Bhaskar D Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 221–239, 2000.

[3] L Anders Ekman, W Bastiaan Kleijn, and Manohar N Murthi, "Regularized linear prediction of speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 65–73, 2008.

[4] Etienne Denoël and J-P Solvay, "Linear prediction of speech with a least absolute error criterion," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 6, pp. 1397–1403, 1985.

[5] J Schroeder and R Yarlagadda, "Linear predictive spectral estimation via the l_1 norm," *Signal processing*, vol. 17, no. 1, pp. 19–29, 1989.

[6] Daniele Giacobello, Mads Græsbøll Christensen, Manohar N. Murthi, Søren Holdt Jensen, and Marc Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[7] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[8] John R Deller, John G Proakis, and John HL Hansen, *Discrete-time processing of speech signals*, Ieee New York, NY, USA:, 2000.

[9] Zhilin Zhang and Bhaskar D. Rao, "Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2009–2015, 2013.

[10] Yuzhe Jin and Bhaskar D Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 3830–3833.

[11] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[12] David Wipf, Jason Palmer, and Bhaskar Rao, "Perspectives on sparse bayesian learning," *Advances in neural information processing systems*, vol. 16, pp. 249–256, 2004.

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

[14] Richard J Mammone, Xiaoyu Zhang, and Ravi P Ramachandran, "Robust speaker recognition: A feature-based approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 5, pp. 58, 1996.