SPEECH ANALYSIS BY ESTIMATING PERCEPTUALLY RELEVANT POLE LOCATIONS

Venkatraman Atti and Andreas Spanias

Department of Electrical Engineering, Ira A. Fulton School of Engineering,
Arizona State University, Tempe, AZ 85287-5706, USA
[atti, spanias]@asu.edu

ABSTRACT

An approach for estimating the perceptually-relevant pole locations is described. This "perceptual poles" are determined by using an auditory excitation pattern-matching method. The estimated perceptual poles are then used to construct a perceptually-motivated all-pole (PMAP) filter for use in speech analysis/synthesis. The proposed PMAP approach is compared against some of the existing perceptually-based linear prediction (LP) methods, i.e., the perceptual LP and the Warped LP. The PMAP approach compares well against the perceptual LP and the warped LP in terms of speech reconstruction quality and estimation of the formant frequencies.

1. INTRODUCTION

Limitations associated with the conventional linear prediction (LP) have been studied extensively and several extensions to LP-based speech analysis-synthesis have been proposed [1]-[3]. Because of the underlying source-system model in the conventional LP, it is not straightforward to fully integrate psychoacoustic principles in linear predictive coding (LPC). A simple, yet popular, approach employed in speech coding standards is to use a perceptual weighting filter (PWF) to shape the quantization noise according to the masking properties of the human ear [4] [5]. However, the PWF is not suitable when there is a pronounced spectral tilt, especially, in case of wideband speech [6]. Other popular LP methods that make use of certain perceptual constrains and the auditory psychophysics include, the perceptual LP (PLP) [7] and the warped LP (WLP) [8] [9]. We note that both the PLP and the WLP impose indirectly perceptual constrains either by manipulating the input speech spectrum or by scaling appropriately the frequency axis. Consequently, both these methods tend to sacrifice some of the properties of conventional LP and end up incorporating several inexplicit psychoacoustic models into the LP [1]-[3], [9]. In particular, in the PLP, a perceptually-based auditory spectrum is obtained by filtering the input speech spectrum using a filterbank that mimics the critical band structure of the auditory filterbank. An all-pole filter that approximates the auditory spectrum is then computed using the autocorrelation method [4]. On the other hand, in the WLP [8] [9], the main idea is to warp the frequency axis (usually, according to a Bark scale) prior to performing the LP analysis to effectively provide a better resolution at some frequencies than at others. This is typically done by replacing the unit-delay elements of the LP analysis filter with all-pass sections, i.e., $(z^{-1}-\lambda)/(1-\lambda z^{-1})$, where, λ is the warping coefficient.

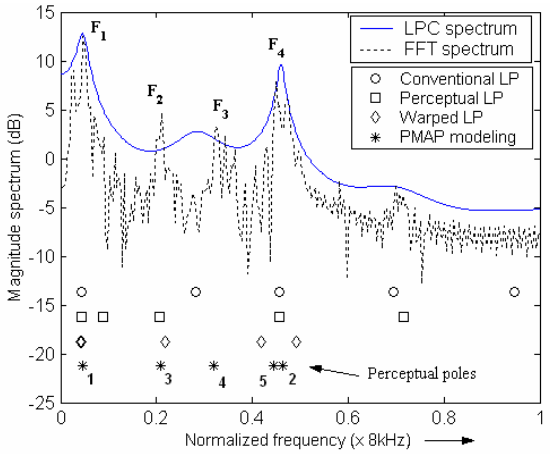

Figure 1. An example depicting the pole locations estimated by the LP ('o'), the PLP ('□'), the WLP (' ◊ ') and the PMAP modeling ('*') corresponding to a voiced speech segment

In this paper, we introduce a new speech analysis/synthesis approach that employs a perceptually-motivated all-pole (PMAP) filter. The proposed PMAP modeling yields: *i*) improved speech reconstruction quality; *ii*) accurate estimation of the formant frequencies; and *iii*) improved spectral modeling. The main idea is to directly estimate the perceptually relevant poles based on an auditory excitation pattern (AEP)-matching method called excitation similarity weighting (ESW) [10]-[12]. The ESW methodology was first proposed in the context of sinusoidal modeling of audio in order to rank and select the perceptually relevant sinusoids for scalable audio coding. More details on the ESW technique are given in Section 2. As a preamble, it is interesting to analyze the pole locations estimated by the various LP-based speech analysis-synthesis methods. Simultaneously, we will also highlight some of the merits of the proposed PMAP modeling as opposed to the existing PLP and WLP. In Figure 1, a 20ms voiced speech segment sampled at 16kHz is used to illustrate the pole locations obtained from the conventional LP (shown as 'o'), the PLP ('□'), the WLP (' ◊ '), and the PMAP modeling ('*'). The first four formant frequencies are indicated as F_1, F_2, F_3, and F_4. This figure also shows the FFT spectrum (dotted line) and a tenth-order LP spectral envelope (solid line) corresponding to the voiced-speech segment. From Figure 1, we note that, *i*) the conventional LP fails to model accurately the formants F_2 and F_3. This is

primarily due to the psychoacoustically-blind least squares error minimization criterion used in LP (e.g., see [16] and [3]); *ii*) the PLP manages to model the formants F_1, F_2, and F_4, however, it totally misses the formant F_3. This is because of several approximations employed by the PLP in computing the auditory speech spectrum [7]; *iii*) the WLP ($\lambda = 0.57553$) gives more emphasis to the formant F_1 (which is not correct, see observation (*iv*) below) by placing two poles at the F_1 formant frequency, and manages to model formants F_2 and F_4. Nevertheless, the use of all-pass sections in the WLP would result in delay-free loops that leads to dissimilarity in the structures of the analysis and synthesis filters. This dissimilarity implies that computationally expensive recursive filters will be required for the WLP synthesis. Also, it has been pointed out in [9] that a drop of 1-15dB in the spectral flatness measure (SFM) associated with the WLP would occur compared to that of the conventional LP. In particular, the WLP sacrifices the whitening property in order to provide better modeling at some frequencies and to accommodate for the perceptual shaping of quantization noise. Several other limitations (e.g., numerical inaccuracies) associated with the WLP analysis-synthesis were described in [2]. *iv*) the PMAP modeling not only estimates accurately all the formant frequencies, but also ranks the estimated poles according to their perceptual relevancies (see perceptual ranks shown in Figure 1). Section 2 describes the algorithm for the computation of the PMAP filter. Experimental results are given in section 3. Section 4 presents a comparative study of the computational complexity associated with the conventional LP, the PLP, the WLP, and the PMAP modeling. Concluding remarks are also included in section 4.

2. PERCEPTUALLY-MOTIVATED ALL-POLE FILTER

The idea of constructing a perceptually-motivated all-pole (PMAP) filter is a rather simple one. First, we estimate the perceptually-relevant pole frequencies using the ESW measure [10]. Second, we use an iterative procedure to compute the corresponding pole amplitudes [15] [16]. Finally, we construct a PMAP filter in cascade-form, i.e.,

$$H(z) = \frac{1}{\prod_{i=1}^{p} (1 - 2r_i \cos\theta_i z^{-1} + r_i^2 z^{-2})} \qquad (1)$$

where (r_i, θ_i) denote the *i*-th pole location in polar coordinates and *p* is the total number of poles. Note that in the above equation, *H(z)* is represented as second-order factors in order to make use of the conjugate-symmetry associated with the poles. One consequence of this is that the prediction order will always be even. Nevertheless, the PMAP analysis presented in this paper can easily be extended to the case of first-order factors.

2.1. Perceptual pole frequencies, θ_i

The frequency associated with a sinusoidal stimulus that provides the maximum AEP-matching between the original and reconstructed speech is selected as a perceptual pole-frequency. The motivation for this approach came from the ESW sinusoidal component selection strategy proposed by Painter and Spanias [10] [11]. The perceptual pole frequencies are computed as follows. First, the input speech, *s(n)*, is segmented into 20ms frames (320 samples at 16kHz). Next, a candidate set that consists of 30 sinusoids is estimated on each frame using the short-time Fourier transform (STFT) analysis. An iterative ranking procedure is performed next [10]. The objective on the *k*-th iteration is to extract from the candidate set the most perceptually salient sinusoid, given the previous (*k*-1) selections. The maximum perceptual salience is associated with the sinusoidal stimulus that is able to affect the greatest improvement in matching between the AEP[1] associated with the original signal (i.e., called the reference AEP) and the AEP that is associated with the reconstructed signal. In particular, in the first iteration, the AEPs associated with each of the estimated (*K*=30) sinusoids are calculated and compared against the reference AEP. At the end of the first iteration, the sinusoid that generates an AEP most closely resembling the reference AEP is chosen as the first component. During the second iteration, each of the remaining components is individually combined with the first component to determine which of the remaining components provides the greatest increase in AEP-matching. The process repeats until all the *p* perceptually-relevant pole frequencies have been obtained.

2.2. Perceptual pole amplitudes, r_i

Both the frequency domain and time domain approaches have been investigated to estimate the perceptual pole amplitudes. In the frequency domain approach, we minimize the prediction error, *E*, that is given by,

$$E = \frac{1}{N} \sum_{m=1}^{N} \frac{|P(\omega_m)|^2}{|H(\omega_m)|^2} \qquad (2)$$

with respect to r_i. In the above equation, $|P(\omega_m)|^2$ is the power spectral density (PSD) of the input speech segment, *N* is the number of discrete frequencies ω_m, and $|H(\omega_m)|^2$ is the PSD associated with the PMAP filter. The minimization of *E* with respect to r_i, $1 \le i \le p$, i.e., $\partial E / \partial r_i = 0$, yields,

$$A r_i^3 + B r_i^2 + C r_i + D = 0; \ 1 \le i \le p \qquad (3)$$

where

$$A = \frac{1}{N} \sum_{m=1}^{N} \xi_m; \ B = -3\cos(\theta_i) \left[\frac{1}{N} \sum_{m=1}^{N} \xi_m \cos(\omega_m) \right];$$
$$C = 2\cos^2(\theta_i) + \frac{1}{N} \sum_{m=1}^{N} \xi_m \cos(2\omega_m); \text{ and, } D = B/3 \qquad (4)$$

$$\text{and, } \xi_m = P(\omega_m) \left| \prod_{k=1, k \ne i}^{p} (1 - 2r_k \cos\theta_k e^{-j\omega_m} + r_k^2 e^{-2j\omega_m}) \right|^2 \qquad (5)$$

Since A, B, C, and D are all real, the roots of the equation (3) would contain at least one real value. Also, note that A, B, C, and D are all independent of r_i. This leads to a convenient iterative procedure and only requires that the pole amplitudes be initialized first. One possible approach is to initialize the pole amplitudes according to the amplitudes associated with the sinusoidal stimuli that are estimated in the ESW procedure. This frequency-domain approach, however, does not guarantee convergence to "true" values primarily because of the LP error criterion employed [16]. El-Jaroudi and Makhoul discussed the drawbacks associated with the LP error criterion and presented

[1] The auditory excitation patterns (AEPs) are generated using steps similar to [14]. For more information on the AEPs refer to [12] [13].

an all-pole modeling method based on a discrete-form of the Itakura-Saito (IS) distance measure [16]. Minimizing the discrete-form of the IS error measure with respect to r_i is not trivial.

To overcome this problem, we present a computationally-efficient time-domain approach to estimate the perceptual pole amplitudes. The method was inspired by the cascade-form LP proposed by Jackson and Wood [15]. In [15], the conventional direct-form LP was reformulated to compute the roots of the predictor polynomial in an iterative manner. In particular, the pole angles, θ_i, and the corresponding pole amplitudes, r_i, were estimated iteratively by solving p simultaneous non-linear equations using a modified steepest-descent algorithm. In our case, we have already estimated the perceptual pole frequencies, θ_i. Therefore, in order to compute r_i, we slightly modify the update equations presented in [15] to incorporate the already estimated pole frequencies, θ_i. The resulting updates for, r_i, are given by,

$$\Delta r_i = -\mu r_i \frac{\sum_{k=0}^{2p}\sum_{l=1}^{2p-1} h(k)\zeta_i(l)\phi(k,l)}{\sum_{k,l=1}^{2p-1}\zeta_i(k)\zeta_i(l)\phi(k,l)}; \quad 1 \le i \le p \qquad (6)$$

where μ is the convergence factor, $\phi(k, l)$ denotes the covariance matrix of order $[2p+1, 2p+1]$ associated with the input speech segment, $h(n)$ is the impulse response of the PMAP filter $H(z)$, and $\zeta_i(n)$ is given by,

$$\zeta_i(n) = -2r_i\cos(\theta_i)h_{1i}(n) + 2r_i^2 h_{2i}(n) \qquad (7)$$

and $h_{ti}(n); t = 1,2$ denote the impulse response of $H_{ti}(z)$,

$$H_{ti}(z) = \frac{z^{-t}H(z)}{(1-2r_i\cos\theta_i z^{-1} + r_i^2 z^{-2})} \qquad (8)$$

Although the update equations presented above look deceptively complex, note that the only computations required are to estimate the covariance matrix $\phi(k, l)$, the impulse response, $h(n)$, and the gradient-related parameter, $\zeta_i(n)$. Jackson and Wood have pointed out that an appropriate choice of μ would ultimately guarantee convergence to true values [15]. Typically, μ values in the range of 0.2-0.4 resulted in good convergence rates (see Figure 3). Section 4 further elaborates on the computational complexity associated with the PMAP modeling.

3. EXPERIMENTAL RESULTS

For all the experiments described in this section, the input speech was sampled at 16kHz and segmented as 20ms frames. A tenth-order predictor was employed, i.e., p = 5 poles. The perceptual pole amplitudes were estimated using the time-domain approach and the convergence factor μ was taken as 0.3.

i) *AEP-matching* – In Figure 2(a), the AEPs generated by the reconstructed speech signals from the LP, the PLP, the WLP, and the PMAP modeling were compared against the AEP associated with the input speech. From this figure, it can be noted that the AEP obtained from the PMAP matches closely the reference AEP. On the other hand, the AEPs generated by the LP, the PLP and the WLP differ from the reference AEP. In order to clearly differentiate the performance of the PMAP approach from the others, in Figure 2(b), we compare the AEPs generated by the prediction residuals of these methods against a

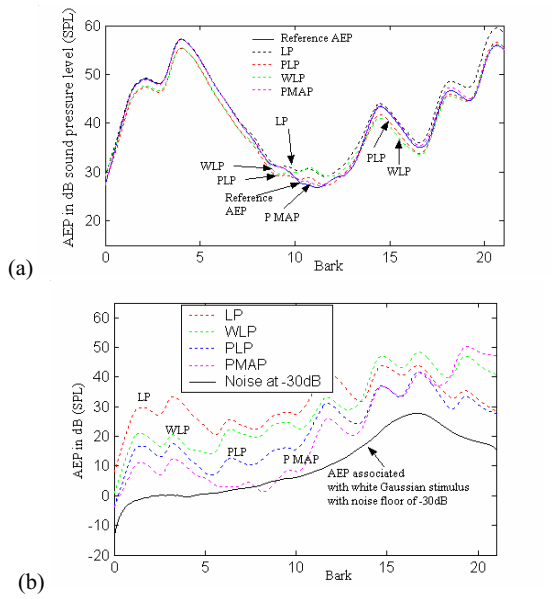

Figure 2. (a) AEPs generated by the input speech segment, the LP reconstructed speech (RS), the PLP RS, the WLP RS, and the PMAP RS; (b) AEPs associated with the prediction errors from the LP, the PLP the WLP, and the PMAP modeling.

random white Gaussian stimulus with noise floor -30dB. Results that further validate and justify the selection criteria used in the ESW methodology are given in [11]. Also note that the ESW methodology does not seek to satisfy a noise threshold criteria, but guarantees maximal matching between the modeled and the original excitation patterns.

ii) *Exact formant frequency estimation* – Given the success of the ESW methodology to rank and select the perceptually-relevant pole frequencies, the PMAP modeling yields accurate information regarding the formant frequencies and bandwidths. An example that illustrates this was presented earlier in section 1, Figure 1. The PMAP approach also yields an improved spectral fitting, especially, at the perceptually-relevant formant regions and compares well against the symmetric LP method proposed in [3].

iii) *Spectral envelope modeling and whitening* – The PSDs of the speech frame, the LP residual, the WLP prediction error, and the PMAP residual are shown in Figure 4(a) through (d), respectively. Note that the WLP residual was filtered using $D_0^{-1}(z) = (1-\lambda z^{-1})/(\sqrt{1-\lambda^2})$ [9] in order to allow for reasonable comparisons. From this figure, it is clear that both the WLP and the PMAP filter can pick out most of the peaks at low frequencies because of the perceptual constraints employed. On the other hand, the conventional LP provides equal emphasis to all frequencies in minimizing the error energy. Also, note that the update equations given in (6)-(8) were derived by reformulating the conventional LP for a cascade LP case. Therefore, the use of the ESW methodology for estimating the pole frequencies and the time-domain iterative algorithm for computing the pole amplitudes results in an efficient modeling of the spectral peaks and in maintaining the whitening property.

iv) *Speech analysis/synthesis and integration of the PMAP modeling in the ITU-T G.729 speech standard* – Figure 5 presents a comparison of the prediction residuals obtained from the PLP, the WLP, and the PMAP modeling. The results

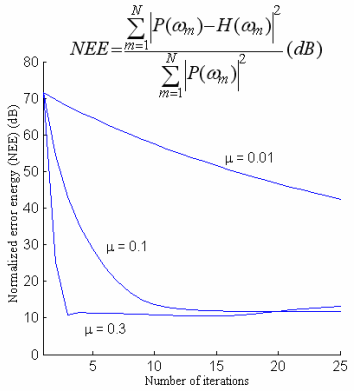

Figure 3. Normalized error energy (NEE) convergence for various values of μ

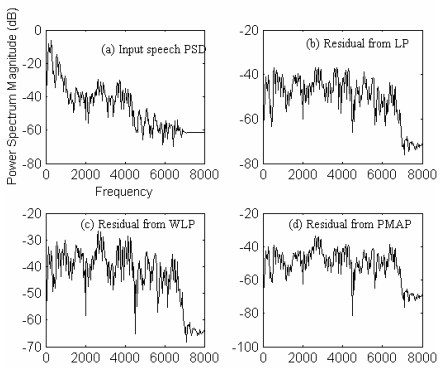

Figure 4. PSD of (a) the input speech, (b) the LP residual, (c) the WLP residual, and (d) the PMAP residual.

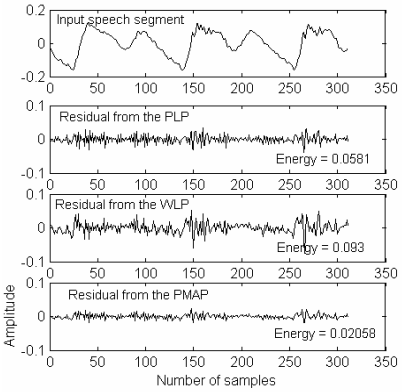

Figure 5. Input speech segment and the residuals obtained from the PLP, the WLP and the PMAP modeling

achieved with the voiced speech segments were consistent with the unvoiced case as well. As a preliminary testing, we replaced the conventional LP in the ITU-T G.729 standard [5] with the PMAP modeling. The experiments consistently revealed that, relative to the LP, the PMAP modeling reduces the prediction error energy significantly; therefore, requiring fewer bits to model the residual. In general, a gain of 5-6 bits per frame (80 samples in the ITU-T G.729) was noted.

4. COMPUTATIONAL COMPLEXITY AND CONCLUDING REMARKS

It has been pointed out in [9] that relative to the conventional LP, the WLP requires approximately 2-3 times more computations for calculating the autocorrelations and 4-8 times for implementing the warped synthesis filter. The PMAP modeling involves the following computations: estimating a candidate set of K sinusoids, selecting p perceptual pole frequencies, and computing the corresponding pole amplitudes. In general, 5-10 iterations are sufficient to compute the pole amplitudes when μ is set around 0.3 (see Figure 3). Typically, the PMAP modeling requires 6-8 times more computations relative to the conventional LP. Note that the computations are mainly from the two iterative searches performed. A computationally fast algorithm that employs perceptual pruning techniques [11] can be employed to speed up the pole frequency search. Sinusoidal trajectory smoothing techniques and matching pursuit algorithms can also be employed to reduce the number of iterative searches.

This paper described a method to obtain the perceptually-relevant pole locations for use in speech analysis-synthesis. An AEP-matching method was employed to estimate the "perceptual poles." Results that justify the use of the AEP matching method to estimate the perceptual poles were given. Experiments that demonstrate the improved speech reconstruction quality and accurate estimation of the formant frequencies were presented.

5. REFERENCES

[1] A. Harma and U. K. Laine, "Linear predictive coding with modified filter structures," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 8, pp. 769-777, Nov. 2001.

[2] A. C. den Brinker, *et al.*, "IIR-based pure linear prediction," *IEEE Trans. Speech Audio Proc,* vol. 12, pp. 68-75, Jan. 2004.

[3] P. Alku and T. Backstrom, "Linear predictive method for improved spectral modeling of lower frequencies of speech with small prediction orders," *IEEE Trans. Speech Audio Proc.,* vol. 12. no. 2, pp. 93-99, Mar. 2004.

[4] P. Kroon and W. B. Kleijn, "Linear prediction-based analysis-synthesis coding" in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., Elsevier, 1995.

[5] R. Salami, *et al.*, "Design and description of CS-ACELP: A toll quality 8kb/s speech coder," *IEEE Trans. Speech Audio Proc.,* vol. 6, no. 2, pp. 116-130, Mar. 1998.

[6] B. Bessette, *et al.*, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech Audio Proc.,* vol. 10, no. 8, pp. 620-636, Nov. 2002.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.,* vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[8] H. W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Amer.,* vol. 68, pp. 1071-1076, 1980.

[9] A. Harma and U. K. Laine, "A comparison of warped and conventional linear prediction," *IEEE Trans. Speech Audio Proc.,* vol. 9, no. 5, pp. 579-588, July, 2001.

[10] T. Painter and A. Spanias, "Perceptual segmentation and component selection in compact sinusoidal representations of audio," *Proc. IEEE ICASSP*, vol. 5, pp. 3289-3292, May 2001.

[11] T. Painter and A. Spanias, "Perceptual segmentation and component selection for sinusoidal representations of audio," (To be published in *IEEE Trans. Acoust. Speech, Sig. Proc.*)

[12] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer-Verlag, 1990.

[13] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, Fifth Edition, Jan. 2003.

[14] B. Paillard, *et al.*, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Aud. Eng. Soc.*, v. 40, n. 1/2, pp. 21-31, Jan./Feb. 1992.

[15] L. B. Jackson and S. L. Wood, "Linear prediction in cascade form," *IEEE Trans. ASSP,* vol. ASSP-26, no. 6, pp. 518-528, Dec. 1978.

[16] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. on Sig. Proc.,* vol.39, no.2, pp.411-423, Feb 1991.