All-Pole Model Estimation of Vocal Tract on the Frequency Domain

Luis Weruaga [♭] *and Amar Al-Khayat* [♯]

[♭] VISKOM, Austrian Academy of Sciences, Vienna, Austria,
[♯] Dictation Systems, PHILIPS Austria GmbH, Vienna, Austria
weruaga@ieee.org, amar.al-khayat@philips.com

Abstract

Probably the most powerful method for speech analysis is the linear prediction analysis, or LPC analysis, one of its main characteristics being the estimation of time-domain related parameters from time-domain samples. This paper proposes a novel speech analysis framework for estimating the spectral poles directly from spectral samples in voiced speech utterances. The method can be described in plain words as the task of fitting the spectral envelope of an all-pole model directly on the log energy of the harmonics. This problem is addressed with an analysis-by-synthesis mechanism supported on a Newton-Raphson algorithm of fast convergence. The proposed method differs clearly from previous approaches commonly used in Harmonic or Sinusoidal Coding. Comparative results on synthetic signals show the excellent performance of the novel analysis technique.

Index Terms: formant estimation, spectral analysis, all-pole model, logarithmic spectral distance.

1. Introduction

Autoregressive (AR) modelling is a popular parametric method for spectral analysis and adaptive filtering in speech processing [1, 2]. A short segment of discrete-time speech $s[n]$ is commonly described by the following difference equation

$$s[n] = \sum_{i=1}^{P} a_i \, s[n-i] + e[n] \,, \tag{1}$$

where $e[n]$ is considered a periodic impulse sequence for voiced utterances and a random stationary process for unvoiced ones, a_i are the AR coefficients and P is the model order. The autocorrelation method, also known as the Yule–Walker method, remains the most accepted method to obtain the AR or linear prediction coefficients (LPC) a_i. Its popularity is based largely on the fact that the estimated LPCs yield a stable all-pole model, which is accurate enough for most practical applications.

In spite of the time-domain character of this popular approach, the frequency domain is very often adopted as framework for assessing the estimation accuracy. Due to their very weak quantization properties, the AR coefficients are usually translated into a spectral-based parametrization, such as the line spectral frequencies (LSF) [4]. Furthermore, in spite of the stability of the Yule–Walker solution, the criterion for stability does not regard the AR coefficients, but frequency-domain parameters, such as the mentioned LSF or the poles of the model. These last ones are the roots of the P-order polynomial equation

$$1 - \sum_{i=1}^{P} a_i \, p^{-i} = 0 \,. \tag{2}$$

The poles have localized spectral sensitivity and inform explicitly on the system stability.[1] Apart from this well-known advantage, the pole-based characterization has many other appealing properties, such as robustness to quantization, interpolation-friendliness, simple signal synthesis mechanism, and meaningful descriptive features in many popular fields [5].

AR modelling exclusively in the spectral domain has not aroused much debate or interest so far in signal [6] or speech processing [7], with most of the speech coding techniques [8] relying on the algorithmic-simple time domain-based Yule–Walker mechanism. In harmonic or sinusoidal coding [9] the focus is on the contrary the compression of the amplitude and phase of the harmonics. The last works on harmonic coding wisely point out to the estimation of the all-pole model from the energy of the harmonics [7, 10] as an alternative to achieve better compression ratios. However, the technique proposed in those works cannot be considered a frequency-domain estimation, since the spectral envelope, obtained by spline-based interpolation of the spectral energy of the harmonics, is translated to the time domain so that Yule–Walker can be applied. Moreover, in case of non-stationary excitation the subjective spectral envelope delineated by the harmonics does not correspond strictly to that of an all-pole model [11], therein the mentioned method [7] becoming inaccurate.

This paper presents a novel all-pole estimation framework that deals with the direct estimation of poles from frequency samples in voiced speech utterances. The proposed method is based on an analysis-by-synthesis (ABS) mechanism in which the parametric poles and the energy of the harmonics are the main information pillars. This ABS framework allows theoretically to face more challenging scenarios, such as in-frame non-stationary excitation [11] or non-stationary vocal tract. To the best of our knowledge, the present paper is the first one addressing the full frequency-domain vocal tract estimation.

2. Problem Statement

According to the speech production model [2], a short segment of a voiced utterance can be expressed in simplified form as

$$s(t) = \kappa \sum_{m} h(t - mT_o) \,, \tag{3}$$

where T_o is the period of the vocal cord excitation, $h(t)$ is the vocal tract impulse response and κ its amplitude. Let us consider N consecutive samples of the discrete-time speech signal $s[n]$

$$x[n] = w_N[n] \, s[n] \,, \tag{4}$$

[1] The reflection coefficients (RC) [3], which are computed directly from the autocorrelation, are appropriate also to detect system stability. However an intuitive interpretation of the spectral response from the RCs is difficult.

where $w_N[n]$ is the N-point analysis window. The Fourier transform of $\hat{x}[n]$ can be expanded as

$$X(\omega) = \kappa\,\omega_o\,H(\omega)\sum_k W_N(\omega - k\omega_o)\,, \tag{5}$$

where ω_o is the fundamental frequency or pitch, and $H(\omega)$ and $W_N(\omega)$ are the Fourier transforms of the discrete-time equivalent vocal tract impulse response and the window respectively. The frequency response of the vocal tract $H(\omega)$ is usually described by an all-pole system as

$$H(\omega) = \frac{1}{\prod_{i=1}^{P}\left(1 - p_i\,e^{-j\omega}\right)}\,. \tag{6}$$

where p_i are the poles of the system (and roots of (2)).

The following approximation

$$X(k\omega_o) = \kappa\,\omega_o H(k\omega_o)\,W_N(0)\,, \tag{7}$$

is feasible if $w_N[n]$ is a well-defined analysis window, such as Hamming or Gaussian. Formally speaking, (7) holds if $W_N(\omega)$'s main lobe width is smaller than the fundamental frequency ω_o. By taking the absolute squared value and logarithm over the spectral equality (7), the following relation holds

$$\log\left|\hat{X}(k\omega_o)\right|^2 = \log\left|\kappa\,\omega_o\,W_N(0)\right|^2 + \log\left|H(k\omega_o)\right|^2. \tag{8}$$

Fig. 1 shows the graphical interpretation of (8) on the complex Z-plane: the logarithm operation transforms the product (6) into a sum of (the same) functions centered at each pole location; the value of this sum at the unit circle corresponds to the log spectral response of the all-pole system.

The goal of this paper is to estimate the all-pole model $H(\omega)$ from the spectral samples $X(k\omega_o)$. We state this problem as the minimization of the following risk

$$\mathcal{J}(\mathbf{p}) \triangleq \sum_k {\chi_k}^2\,, \tag{9}$$

where $\mathbf{p} = (A, p_1, \ldots, p_P)$ is the vector with the log energy and the poles, and χ_k is the residual difference at the k-th harmonic

$$\chi_k \triangleq \log\left|\hat{X}(k\omega_o)\right|^2 - A + \sum_{i=1}^{P}\log\left|p_i - e^{jk\omega_o}\right|^2. \tag{10}$$

The desired solution results from the minimization of the risk (9) with respect to the poles p_i and log-power level A

$$\hat{\mathbf{p}} = \arg\min_{\mathbf{p}}\,\mathcal{J}(\mathbf{p})\,. \tag{11}$$

The proposed risk (9) involves the difference between the log spectral samples and the log AR spectrum. This log-based measure is usually considered as the natural indicator of the subjective closeness between speech spectra. On the contrary the Yule-Walker algorithm, classically used to draw the vocal tract model, yields the minimization of the squared spectral difference in the linear domain, and not in the logarithmic one.

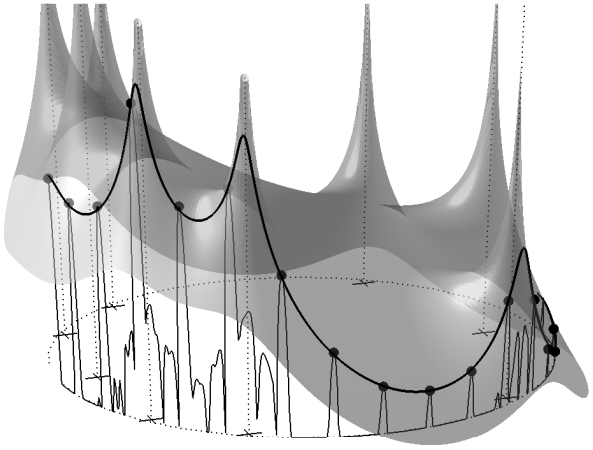

Figure 1: All-pole system in the complex plane: the complex function $\log|H(z)|^2$ is the sum of the function $-\log|z|^2$ centered on each pole location. The spectrum of the output of the all-pole system to a periodic pulse excitation meets that complex surface at the location of the harmonics along the unit circle.

3. Estimation Mechanism

The error surface in the least squares AR modelling is well-known to be quadratic, thus possessing a global minimum. However, the identification of the poles (11) is a nonlinear problem. Furthermore, the risk functional (9) cannot be assured to be convex in the entire domain. The question thus arises as which mechanism is suitable to reach the solution and to whether or not there are false local minima to which the mechanism can converge.

3.1. Numerical Algorithm

The minimization of the risk (9) can be accomplished with the following gradient algorithm

$$\mathbf{p}_{k+1} = \mathbf{p}_k - \frac{1}{2}\,\boldsymbol{\mu}\,\frac{\partial\mathcal{J}(\mathbf{p})}{\partial\mathbf{p}^H}\bigg|_k\,, \tag{12}$$

where $\boldsymbol{\mu}$ is the step-size diagonal matrix

$$\boldsymbol{\mu} \triangleq \frac{\mathbf{I}}{\operatorname{diag}\left(\dfrac{\partial^2\mathcal{J}(\mathbf{p})}{\partial {p_0}^2}, \dfrac{\partial^2\mathcal{J}(\mathbf{p})}{\partial p_1\partial p_1^H}, \ldots, \dfrac{\partial^2\mathcal{J}(\mathbf{p})}{\partial p_P\partial p_P^H}\right)}\,. \tag{13}$$

Here $\mathbf{I}$ is the identity matrix and $\operatorname{diag}(\mathbf{a})$ is the diagonal matrix built with vector $\mathbf{a}$. The step matrix (13) is built with the diagonal of the Hessian matrix, characteristic in the Newton–Raphson method. The gradient results in

$$\frac{\partial\mathcal{J}(\mathbf{p})}{\partial p_0} = -2\sum_k \alpha_k\,\chi_k\,, \tag{14a}$$

$$\frac{\partial\mathcal{J}(\mathbf{p})}{\partial p_i^H} = 4\sum_k \alpha_k\,\chi_k\,\frac{p_i - e^{jk\omega_o}}{\left|p_i - e^{jk\omega_o}\right|^2}\,, \tag{14b}$$

and the Laplacian as

$$\frac{\partial^2 \mathcal{J}(\mathbf{p})}{\partial {p_0}^2} = 2\sum_k \alpha_k \,, \tag{15a}$$

$$\frac{\partial^2 \mathcal{J}(\mathbf{p})}{\partial p_i \partial p_i^H} = 4\sum_k \alpha_k \frac{2-\chi_k}{\left|p_i - e^{jk\omega_o}\right|^2} \,. \tag{15b}$$

The update described by equations (14) can be intuitively interpreted as the pole being "pulled" by the points of the unit circle with an intensity depending on the error and the distance at each circle location.

3.2. Convergence Analysis

It will be shown that there are no false minima in the proposed risk functional except for trivial labelling ambiguities. An important part of the analysis here is supported by the analysis in [5]. The minima of the risk functional (9) are the solution of the following equation

$$\frac{\partial \mathcal{J}(\mathbf{p})}{\partial \mathbf{p}^H} = \frac{\partial \mathcal{J}(\mathbf{p})}{\partial \mathbf{a}^T}\frac{\partial \mathbf{a}}{\partial \mathbf{p}^H} = 0 \,. \tag{16}$$

where $\mathbf{a} = (a_1, \ldots, a_P)$ are the AR coefficients (1). The first term in (16) results in

$$\frac{\partial \mathcal{J}(\mathbf{p})}{\partial \mathbf{a}^T} = \sum_k \hat{\alpha}_k \chi_k \,, \tag{17}$$

where

$$\hat{\alpha}_k = \alpha_k \,\Re\left\{ H(\omega)\, e^{j\omega[1\ldots P]^T} \right\} . \tag{18}$$

The null in (17) is clearly met by the least square (LS) solution over the cepstral domain. However, note that if $\hat{p}_i$ is a solution, $1/\hat{p}_i^*$ can be also valid. On the other hand, the second source of minima in (16) is

$$\frac{\partial \mathbf{a}}{\partial \mathbf{p}^H} = 0 \,, \tag{19}$$

which has been object of careful analysis in [5]. It is easy to deduce that the optimal log-level parameter fulfils

$$\mathrm{E}\big[\hat{p}_0\big] = \log \omega_o^2 + \log\big|W_N(0)\big|^2 . \tag{20}$$

Thus, we can conclude that four classes of regions exist where the gradient (16) can be zero:

1. a point, such that $|p_i| < 1$, where the term (17) is zero: the least squares (LS) solution on the cepstral domain,
2. another regions where not all p_i are inside the unit circle and the term (17) is zero, which corresponds to a unstable all-pole filter,
3. regions (19) with two equal poles, $p_j = p_k$ for $j \neq k$, and
4. regions (19) where $|p_k| = 0$ for one or more k.

In order to avoid the areas 2, 3 and 4, the evolution of the pole should be monitored in the algorithm (12) according to:

— *The poles have to be inside the unit circle*. Enforcing $|p_i| < 1$ prevents case 2.

— *The initial value of any two poles must be dissimilar*. If $p_i \simeq p_j$ the convergence tends to slow down.

— *The initial value of the poles must be* $p_i \neq 0$. Large initial values as $|p_i| \approx 1$ favor a faster convergence.

Although the risk has no false local minima it does exhibit multiple minima. In fact it contains $P!$ minima, all valid solutions that correspond to the possible permutations of the P poles. In spite of this multiple choice, the algorithm presents a rapid convergence to one of the $P!$ solutions, which is usually the closest one to the pole initialization. Finally, it is simple to prove that the risk (9) is a convex function around the minimum.

4. Results

The experiments address the all-pole filter estimation in the presence of a deterministic pulsed excitation. The synthetic signal was generated accordingly, of which $N = 384$ samples were considered (which is equivalent to 48 ms at 8 kHz sampling rate). The signal segment was Hamming-windowed, zero-padded to 1024 samples and DFTed. The fundamental frequency was estimated from the log-spectrum according to the method described in [11], whereby the log-energy of the harmonics was obtained and used as input to the all-pole estimation. Fig. 2 contains the results for two different cases of fundamental frequency: a) T_o larger than the time response of the all-pole filter, and b) a period much lower than the filter time response, which yields severe time overlapping among the response of adjacent pulses. The central pictures of Fig. 2 show the all-pole spectral envelope estimated with the method proposed in this paper (in solid line) and with the method for harmonic coding proposed in [7] (in dotted line). This last method is based on interpolating the harmonic energy with a high-resolution spline over a log-warped frequency domain, Fourier inverting the result and using finally Yule-Walker in order to obtain the time-domain LPC coefficients a_i.

It is pleasant to see that the estimation delivered by the method proposed in this paper does not get affected by the fundamental frequency ω_o. Especially in case of the wide-spaced harmonics undersampling the spectral formants, the estimated all-pole envelope is very accurate. On the contrary, the spline-based technique used in harmonic coding does get affected severely by the fundamental frequency: while in case of a low-pitched segment the estimation is accurate, in case of high pitch the envelope delivered results quite naïve. On the other hand, the evolution of the poles during the analysis-by-synthesis mechanism shows two main phases: the initial one, in which the risk functional is not necessarily convex and thus the pole evolution suffers sudden jumps, and the final one, characteristic for a direct evolution to convergence. Two poles from the 10-order model tend to zero in both examples, indicating that the actual all-pole model order is smaller than initially thought.

We also carried out experiments with real speech segments and the previous discussion can be extended here. However, in several cases the all-pole fitting did not result satisfactory. Following are our interpretations on the problems encountered and some hints on how to address them with the proposed technique:

— Since the length of the speech segment ranged within 30-40 ms, the pitch may not be stationary within that interval. In order to obtain a highly detailed harmonic spectrum, we used the Fan-Chirp transform proposed in [11]. As proven in that work, the resulting spectral envelope is actually a smoother version of the actual one, the more the higher the pitch rate and frequency. This means that the energy of the harmonics do not delineate strictly an all-pole system.

— Because of that relatively long segment, the vocal tract in a real speech segment may not remain stationary within that

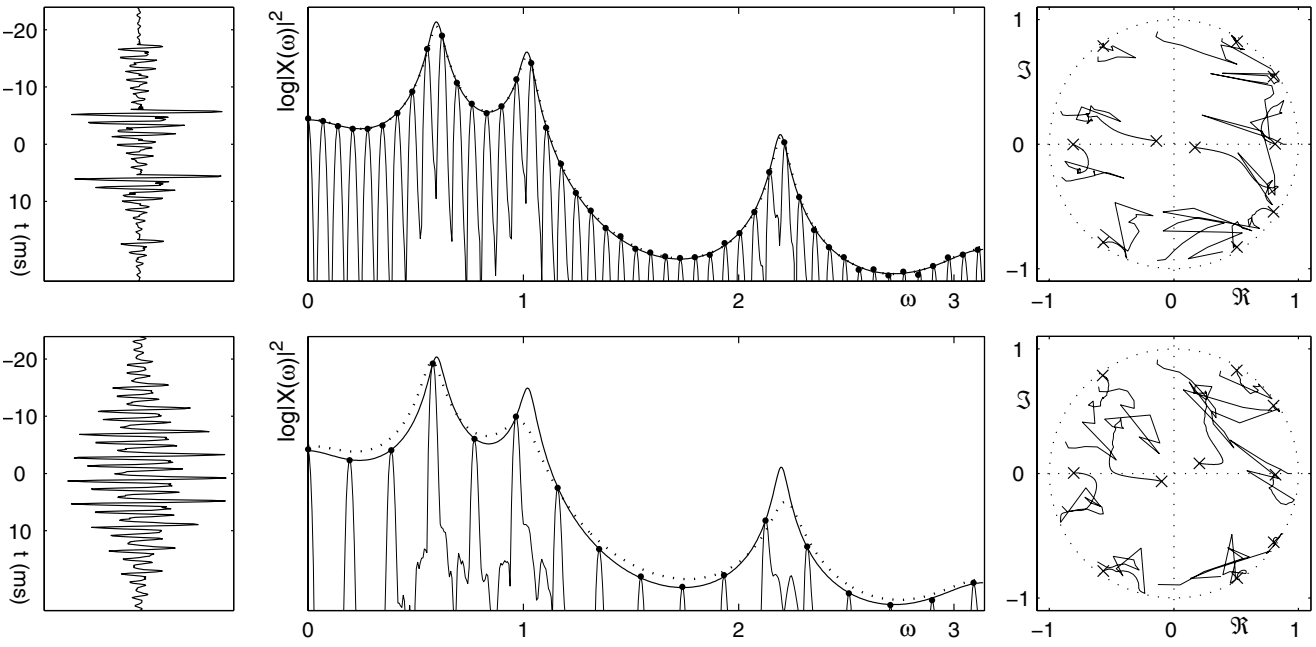

Figure 2: Estimation of an all-pole filter excited with periodic pulses: top – $F_0 = 88$ Hz, bottom – $F_0 = 246$ Hz (actual analog bandwidth 4kHz). For each case, from left to right: signal samples (rotated clockwise), spectrum in log scale with estimated spectral envelope by the proposed method (solid line) and that of [7] (in dotted line), and trajectory of the poles during 50 iterations (final location marked with "$\times$"). The synthetic signal sounds like the /a/ voiced utterance.

interval. As previously, that implies that the spectral envelope does not correspond strictly to an all-pole system.

The analysis-by-synthesis (ABS) technique proposed in the paper may be adequately modified in order to cope with the mentioned non-stationary scenarios (while the harmonic coding technique [7] on the contrary cannot cope with that challenge). These problems are covering currently our research activity.

5. Conclusions

This paper has proposed a novel method to estimate the all-pole model from the log energy of the harmonics in voiced speech utterances. The preliminary analysis-by-synthesis method is the first one addressing the estimation of frequency parameters from spectral samples in a direct fashion. Results over synthetic voiced utterances reveal the accuracy and robustness against different pitch values of the proposed technique, in contrast to the current techniques used in harmonic coding. This work represents a modest contribution to shift the attention to the frequency domain in linear prediction analysis. A new version of the method to cope with the non-stationary nature of real speech is currently under investigation. The integration of this technique in Harmonic Coding is also foreseen in future steps.

6. References

[1] B. Rust and D. Donelly, "The fast Fourier transform for experimentalists, part IV: Autoregressive spectral analysis," *Computing in Science & Engineering*, vol. 7, pp. 85–90, Nov.–Dec. 2005.

[2] T. F. Quatieri, *Discrete-time speech signal processing*. Prentice Hall, Upper Saddle River, NJ, 2002.

[3] J. Leroux and C. Gueguen, "A fixed point computation of partial correlation coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, pp. 257–259, Jun. 1977.

[4] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, no. 1, p. 535, 1975.

[5] A. Nehorai and D. Starer, "Adaptive pole estimation," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 38, pp. 825–838, May 1990.

[6] R. Pintelon and J. Schoukens, "Time series analysis in the frequency domain," *IEEE Trans. Signal Processing*, vol. 47, pp. 206–210, Jan. 1999.

[7] T. G. Champion, R. J. McAulay, R.J., and T. F. Quatieri, "High-order all-pole modelling of the spectral envelope," in *Proc. IEEE ICASSP 1994*, pp. 529–532.

[8] A. M. Kondoz, *Digital speech: Coding for low bit rate communication systems*. John Wiley & Sons, 2004.

[9] R. J. McAulay and T. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech Signal Processing*, ASSP-34, pp. 744–754, 1986.

[10] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech coding and synthesis*, by Kleijn and Paliwal (Eds.), pp. 121–174, Amsterdam: Elsevier Science, 1995.

[11] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech Commun.*, vol. 48, no. 5, pp. 474-492, May 2006.