# FREQUENCY-DOMAIN SPECTRAL ENVELOPE ESTIMATION FOR LOW RATE CODING OF SPEECH

*M. Jelinek and J.-P. Adoul*
University of Sherbrooke, Quebec, Canada, J1K 2R1
milan.jelinek@gel.usherb.ca

## ABSTRACT

Estimation of spectral envelope in frequency domain allows to avoid some problems of the Linear Prediction (LP) algorithms for voiced speech. We present a low complexity method of spectral envelope estimation from harmonics for low rate coding. The method consists in computing harmonic amplitude spectrum using pitch-synchronous DFT with length depending on voicing, modifying this spectrum outside the telephone bandwidth to simplify modeling of the useful bandwidth and interpolating it by a frequency-domain low-pass filter. An all-pole model is then fitted to this modified smoothed version of the harmonic spectrum. The method was implemented on the Harmonic-Stochastic Excitation (HSX) vocoder and the performance was compared with the LP algorithm similar to that used in the G.729 speech coding standard. A-B comparative tests show an important increase in perceptual quality.

## 1. INTRODUCTION

Most of modern speech coders use a parametric model of speech production in the form of an excitation signal filtered through a system. The excitation signal models the air pressure emanating from lungs through vocal cords. For voiced sounds, the vocal cords oscillate and the excitation signal is quasi-periodic. The harmonic structure of the excitation spectrum is responsible for the fine structure in voiced speech spectrum. If the vocal cords do not oscillate, the excitation looks like a white noise and unvoiced speech sounds are generated [1].

The spectral envelope is mainly determined by the shape of vocal tract and it is generally represented by a linear filter. Usually, it is an autoregressive (AR) filter that can be expressed in z-transform as

$$H(z) = \frac{1}{A(z)} = \frac{1}{1+\sum_{k=1}^{P} a_k \cdot z^{-k}} . \qquad (1)$$

This filter is often called synthesis filter. Its coefficients are generally estimated by means of Linear Prediction. Estimation of spectral envelope by LP uses the hypothesis that speech signal can be modeled as the output of the filter $H(z)$ when the input is a single impulse or a white noise [2]. This hypothesis is not exactly valid for voiced speech when the excitation signal has a quasi-periodic nature. Consequently, the LP representation suffer from some drawbacks that appear especially for high-pitch speakers.

In spite of some imprecision, the LP is still widely used for estimation of spectral envelope in all types of low rate speech coders. Its popularity is essentially due to its low complexity and to the fact that it gives an all-pole filter. The all-pole model represents a good compromise between accuracy of the model and number of bits necessary for its quantization.

The LP is predominant in CELP-type coders where the imprecision of spectral envelope estimation can be compensated to some extend by quantization of the excitation signal waveform. These coders dominate low rate speech coding for rates above approximately 5 kb/s. Below this rate, their performance decreases rapidly.

At rates below about 4 kb/s, speech coders called parametric become more efficient. Here, instead of quantizing excitation waveform, the excitation signal is modeled based on some parameters. These parameters are typically pitch and some information on the contribution of periodic and noisy components. As the information about excitation signal is rather simple, a good envelope representation is very important. To overcome shortcomings of the LP, several approaches have been described in the literature. One possibility consists in computing speech amplitude spectrum, specifying in some way its envelope and then fitting an all-pole filter to this envelope. This is the approach used in the method presented in this contribution.

The paper is organized as follows. Section 2 reviews the LP algorithm, discusses its shortcomings and summarizes other methods of spectral envelope modeling for low rate coding of speech. In section 3, the new method is presented. Results of modeling and subjective test results are given in section 4. Finally, section 5 contains concluding remarks.

## 2. LP AND OTHER SPECTRAL ENVELOPE REPRESENTATIONS

In the LP algorithm, every speech sample $s[n]$ is approximated by a linear combination of preceding samples. Prediction error $e[n]$ - the error between the original speech sample and its approximation - can be written as

$$e[n] = s[n] + \sum_{k=1}^{P} a_k \cdot s[n-k] , \qquad (2)$$

where $P$ is the prediction order and $a_k$ are the coefficients of the inverse filter $A(z)$ in Equation (1). The coefficients $a_k$ are obtained by minimization of mean square error on a time interval where speech signal can be considered stationary:

$$\min_{a_k} \left\{ \sum_n e[n]^2 \right\}. \quad (3)$$

Using Parceval's theorem, the mean squared error can be written in frequency domain

$$\sum_{n=-\infty}^{\infty} e[n]^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\left| S\left(e^{j\theta}\right)\right|^2}{\left| H\left(e^{j\theta}\right)\right|^2} d\theta, \quad (4)$$

where $S\left(e^{j\theta}\right)$ is the Fourier Transform of the speech signal and $H\left(e^{j\theta}\right)$ is the frequency response of the synthesis filter [2]. Equation (4) shows that the minimization of the mean squared error tries to fit the synthesis filter frequency response to speech power spectrum and not to its envelope. This is not relevant for unvoiced sounds or for voiced sounds with large number of harmonics because the criterion in Equation (4) is more severe for regions where the signal spectrum amplitude is superior to the filter frequency response than for the regions where it is the inverse. Consequently, if the order of filter is maintained reasonably low, the estimation of filter coefficients by means of LP results actually in speech spectrum envelope. On the other hand, for high-pitch speakers and higher LP analysis order, the frequency response of the synthesis filter shows tendency to follow the fine structure of speech power spectrum. Formant frequencies are then often biased toward pitch harmonics and formant bandwidth is underestimated.

Several methods have been proposed to overcome this problem, both in time and frequency domain. In the approach of selective LP, the speech signal samples corresponding to regions around glottal pulses are excluded from filter coefficient estimation in order to reduce periodicity of the excitation signal [3]. Another approach consists of considering periodicity of the excitation in the criterion minimization (3) [4], [5]. A simple method widely used to reduce formant bandwidth underestimation is the artificial expansion of their bandwidth [6].

Spectral envelope estimation in frequency domain consists of specifying in some way the spectral envelope from pitch harmonics in the log domain. As the only information available about spectral envelope are its samples given at pitch harmonic frequencies, it is necessary to do some assumption about envelope behavior between these harmonics. It is generally supposed that sampling theorem is verified and that speech spectral envelope between pitch harmonics is smooth. A continuous description is usually derived by means of different types of interpolation ranging from simple linear interpolation to complex cubic spline interpolation [7]. An all-pole filter model can be then fitted to this smoothed envelope according to (4). Minimization of this criterion now tries to fit correctly the synthesis filter frequency response to speech power spectrum envelope.

Another interesting method called Discrete All-Pole Modeling (DAP) has been proposed by El-Jaroudi [8]. In this method it is accepted that the sampling of spectral envelope by pitch harmonics does not necessarily satisfy the sampling theorem and so that the autocorrelation function is aliased. The DAP thus requires matching the aliased autocorrelation to the autocorrelation of the all-pole filter model aliased in the same manner. This method needs however to resolve a set of non-linear equations. Yet another approach to fitting an all-pole filter to harmonic spectrum is based on Minimum Variance Distortionless Response [9].

Besides the problem that the input signal to the synthesis filter does not have always a spectrum that can be supposed white, the LP methods suffer also from a large speech signal spectrum dynamics. These dynamics are in addition increased by the low-pass filter used in the analog-to-digital conversion with sharp cut-off edges [10]. To reduce the dynamics, a small value is usually added to the main diagonal of covariance matrix [11].

Even if the all-pole filter model of spectral envelope is greatly prevailing, other spectral envelope representations have also been used. These are for example cepstral representation [12] or direct spectrum envelope quantization [13].

## 3. VOICING DEPENDENT FREQUENCY-DOMAIN MODEL

The presented method uses the approach where spectral envelope is derived from pitch harmonics with some special features. These are namely voicing dependent spectrum analysis, modification of harmonics outside the telephone bandwidth in order to reduce spectrum dynamics and low-pass interpolation between harmonics with an adjustment of the frequency scale. The system overview is given in Figure 1.

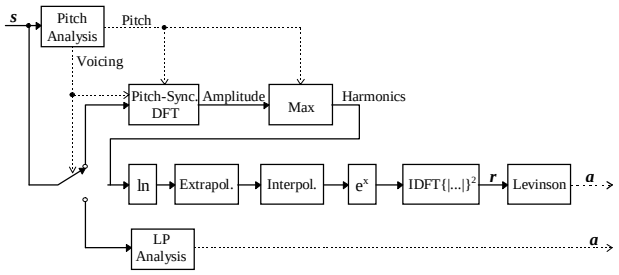


**Figure 1**. Block diagram of the voicing dependent frequency-domain spectral envelope estimation.

At first, pitch frequency is determined. A method estimating spectral envelope from pitch harmonics needs a robust pitch estimator. A small error in pitch is not very important (it can be corrected by a peak-picking algorithm) but a pitch sub-multiple would produce fake harmonics and consequently parasite oscillations in the envelope. In the presented algorithm, the pitch period is estimated in two stages by searching maximum of normalized correlation. In the first stage, pitch is roughly estimated to ensure a smooth pitch evolution. Particular attention is paid to avoid multiples and sub-multiples of pitch. The algorithm used is derived from the composite correlation method [14]. In the second stage, the pitch period is adjusted in the interval of <-5, 5> samples around the first estimation with 1/3 sample resolution.

The maximum of the normalized correlation, computed on the low-pass filtered speech signal with a cut-off frequency of 1200 Hz, is also used as voicing parameter. If this maximum is low, it was experimentally found that the envelope estimation in frequency domain does not change synthesized speech quality. Classical LP is then used because of its low complexity. The threshold was fixed to 0.65.

If the maximum of normalized correlation is greater than 0.65, spectral envelope is estimated from harmonics. First, speech spectrum is computed using pitch-synchronous DFT on Hamming windowed speech signal. The length of the DFT is dependent not only on pitch period but also on voicing information. It was observed that for maximum of normalized correlation above about 0.9, speech signal is strongly periodic and the length of the DFT is chosen longer to ensure better spectral resolution. If the maximum is lower, this is not always true and it often corresponds to transitions in speech signal. The length of the DFT is then smaller.

Harmonic frequencies are next determined from amplitude spectrum by a peak picking algorithm. First, the theoretic position of harmonics is estimated using fractional pitch information. The position is then adjusted by looking for local maxima in a way that neighbor harmonics are not in the search interval. The maximum interval range is <-2, 2> with respect to the supposed harmonic position. This algorithm has better performance in combination with the following low-pass interpolation between harmonics than the algorithm where each new harmonic is searched in an interval defined with respect to the last found harmonic [7].

Given the frequencies and the amplitudes of harmonics, spectral envelope is determined in the log domain. Before interpolation between harmonics is applied, the harmonic spectrum is modified outside the telephone bandwidth in order to reduce the dynamics of speech spectrum. Harmonics outside the interval <100, 3700> Hz are extrapolated from harmonics in useful bandwidth using a half-period of cosine function. The cosine function amplitude was fixed in a way that the spectrum amplitude at 0 Hz is 5 dB below the first amplitude in the interval <100, 3700> and the spectrum amplitude at 4000 Hz is 2 dB below the last amplitude in this interval. These values were found experimentally. The presented modification allows better all-pole filter fit in the useful bandwidth because the filter frequency response does not have to reproduce sharp transitions often present on extremities of speech spectrum (especially in low frequencies). As only the telephone bandwidth is kept at the decoder output and the rest is filtered away, the modification does not affect synthesized speech quality. This method is well suited for parametric coders where it is necessary to actually filter the excitation signal. For coders where the all-pole filter serves only to describe spectral envelope in frequency domain, the all-pole filter can be obtained using some frequency axis transformation [2].

Harmonic spectrum is interpolated by a low-pass Hamming weighted sinc filter. The use of low-pass filter is complicated by the fact that harmonics are not exactly regularly spaced. The interpolation is hence done at first as if these harmonics were equidistant and the position of interpolated envelope points is then given by simple linear interpolation between frequencies of adjacent harmonics:

$$f[i] = f[k] + \frac{f[k+1] - f[k]}{u} \cdot i \,. \qquad (5)$$

In Equation (5), $f[i]$ is a frequency of interpolated envelope point, $f[k]$ is the $k^{th}$ harmonic frequency and $i$ varies between 1 and the upsampling factor. The performance of this relatively low complexity interpolator was comparable to the envelope representation by high-order cepstrum.

The interpolated envelope representation is then brought to the linear domain and the autocorrelation coefficients are found through the inverse DFT of envelope power spectrum. Filter coefficients are finally computed using the Levinson recursion as in the LP methods.

## 4. PERFORMANCE EVALUATION

The presented method was tested for several analysis orders (10, 12, 16 and 20) and modeling results were compared graphically to the performance of the LP algorithm refined by 60 Hz formant bandwidth expansion and by addition of a white noise correction factor. The results were generally better for any order, especially for some female speakers, and the precision was increasing with higher orders. An example is shown in the Figure 2. for the vowel “i” and analysis order of 10.

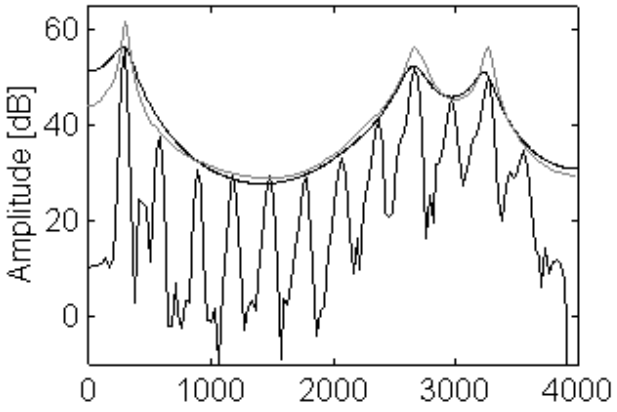


**Figure 2**. Spectral envelope modeling results. The estimation from harmonics is drawn in black line and the LP representation is drawn in gray line.

The frequency-domain spectral envelope estimation method has been implemented on a version of the HSX speech coder [15]. The subjective performance was evaluated by informal A-B comparative test with performance of the coder using LP. Apart from all-pole filter coefficients estimation, both coders were exactly the same.

The test material consisted of 8 short sentences of total length of about 24.5 seconds where 4 sentences were pronounced by male speakers and 4 sentences by female speakers. Each sentence

was coded using the LP method and the frequency-domain method and these pairs were presented to listeners twice in reverse orders. 8 listeners were asked if they prefer the first sentence or the second sentence or if they judge them of the same quality. The results of the test in percentage of preferences are shown in Figure 3 for the analysis order of 10 and 16, separately for male and female speakers. The percentage of preferences for the LP method is on the left side of each chart, for the frequency-domain estimation on the right side, and the percentage of equal quality judgments is in the middle.

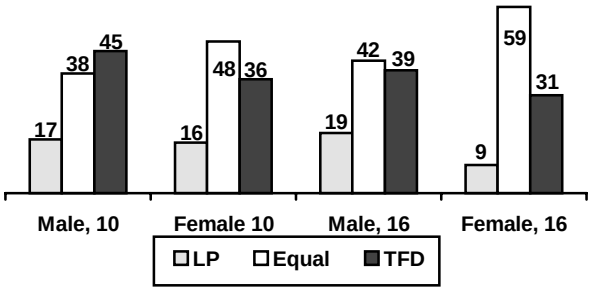


**Figure 3**. Subjective test results in percentage of number of preferences.

It can be seen from Figure 3, that the frequency-domain envelope estimation method is generally preferred to the LP method. As the LP performs well for some sounds, a detailed analysis was made to see the dependency of the test on speech signal. It has been observed that for some sentences, the performance was practically the same for both methods but for other sentences, the frequency-domain method was largely preferred. Interesting is that the percentage of preferences for the LP method was not superior to the presented method for any tested sentence.

## 5. CONCLUDING REMARKS

In this paper, we presented a new spectral envelope estimation method for low rate coding of speech using a frequency-domain approach. This technique is particularly suited for parametric coders where the excitation signal is filtered through an all-pole synthesis filter. In comparison to the LP method, better spectral envelope representation has been observed. This result was confirmed by subjective test using an HSX speech coder.

The parameters of this method has been developed in a sequential manner, mostly using an ACELP coder. Some additional quality gain is thus expected from parameter reoptimization.

## 6. REFERENCES


[1] W.B. Kleijn and K.K. Paliwal, "An Introduction to Speech Coding". *Speech Coding and Synthesis (Kleijn W.B. and Paliwal K.K., editors)*, pages 1-47, Elsevier Science B. V., Amsterdam, 1995.

[2] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, no. 4, pages 561-580, 1975.

[3] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida, and O. Kakusho, " Analysis of Speech Signals of Short Pitch Period by a Sample-Selective Linear Prediction," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-35, no. 9, pages 1233-1239, 1987.

[4] S. Singhal and B. S. Atal, "Optimizing LPC Filter Parameters for Multi-Pulse Excitation," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages. 781-784, Boston, USA, 1983.

[5] M. R. Zad-Issa and P. Kabal, "Smoothing the Evolution of the Spectral Parameters in Linear Prediction of Speech Using target Matching," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages. 1699-1702, Munich, Germany, 1997.

[6] R. Viswanathan and J. Makhoul, " Quantization Properties of Transmission Parameters in Linear Predictive Systems," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-23, no. 3, pages 309-321, 1975.

[7] D. B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-29, no. 4, pages 786-794, 1981.

[8] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. on Signal Process.*, vol 39, no. 2, pages 411-423, 1991.

[9] M. N. Murthi and B. D. Rao, "Minimum Variance Distortionless Response (MVDR) Modeling of Voiced Speech," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1687-1690, Munich, Germany, 1997.

[10] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-27, no. 3, pages 247-254, 1979.

[11] J.-H. Chen, "Low-Delay Coding of Speech," *Speech Coding and Synthesis (W. B. Kleijn and K. K. Paliwal, eds.)*, pages 209-256, Elsevier Sciences B. V., Amsterdam, 1995.

[12] J. H. Chung and R. W. Schafer, "A 4.8 Kbps Homomorphic Vocoder Using Analysis-by-Synthesis Excitation Analysis," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 144-147, Glasgow, Great Britain, 1989.

[13] A. Das and A. Gersho, "Variable Dimension Spectral Coding of Speech at 2400 Bps and Below with Phonetic Classification," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 492-495, Detroit, USA, 1995.

[14] ITU-T SG 15 Contribution, "Description of AT&T 4 Kbit/s Coder," *Source: AT&T*, contact R. Cox, june 1996.

[15] C. Laflamme, R. Salami, R. Matmti and J.-P. Adoul, "Harmonic-Stochastic Excitation (HSX) Speech Coding below 4 Kbit/s," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 204-207, Atlanta, USA, 1996.