# APPLICATION OF EXTREMUM SAMPLING IN SPEECH CODING

*Mohammad R. Nakhai and Farokh A. Marvasti*

Department of Electronic Engineering
King's College London, Strand, London WC2R 2LS, U.K.
E-mail:mohammad.nakhai@kcl.ac.uk

## ABSTRACT

The magnitude spectrum of speech is sampled to extract the extremum (maximum) points representing the underlying sine-wave amplitudes. The resulting nonuniform samples are, first, interpolated using a cubic spline function and, then, modelled by an all-pole magnitude spectrum. The gain factor and the line spectral frequency (LSF) domain representation of the coefficients of the all-pole model are quantized at the encoder. A faithful reconstruction of the spectral extremum envelope is obtained at the decoder using the dequantized all-pole model parameters.

## 1. INTRODUCTION

In sinusoidal speech coding, the extremum samples of the magnitude spectrum of speech and a smooth envelope passing through these samples are used for pitch estimation at the encoder and speech reconstruction at the decoder [1]-[4]. Since voiced speech waveform is quasi-periodic, its spectral domain extremum samples are nonuniformly spaced. Besides, the speech spectrum usually contains small noisy peaks which are irrelevant to its underlying sinusoidal structure. These noisy extrema must be detected and removed in order to avoid ambiguous pitch estimation at the encoder and to achieve a high signal-to-noise ratio of the reconstructed speech at the decoder. In harmonic sinusoidal speech coding, uniform samples of extremum envelope at harmonics of a fundamental frequency are uesd at the decoder to synthesize speech [2]-[4].

Efficient equantization of the extremum envelope of the magnitude spectrum of speech is of great importance in low rate sinusoidal speech coding. An all-pole transfer function is an efficient approach to model the extremum envelope since the resulting all-pole coefficients can be efficiently quantized in the LSF domain. On the other hand, the periodic energy of the voiced speech is most concentrated at low frequency range of spectrum. This demands a good accuracy and resolution in estimating the lower line spectral frequencies (LSFs) which could start from as low as 50 Hz in a male voiced speech. Unfortunately, the present efficient LSF estimation algorithms such as [5] does not have the required accuracy at low frequencies and lead to the coarse quantization of the low LSFs [5].

In this paper, the spline envelope fitted to the extremum points of the magnitude spectrum of speech is encoded by a gain factor and a hig-order set of LSFs. In Section 2, an extremum sampling and refining procedure is presented and an all-pole model of the cubic spline function interpolating the resulting samples is discussed. In Section 3, the quantization of the all-pole gain factor is explained. In Section 4, quantization procedure of the all-pole coefficients in the LSF domain is discussed. Finally, this paper is concluded in Section 5.

## 2. EXTREMUM SAMPLING OF MAGNITUDE SPECTRUM

The extremum points of the magnitude spectrum are determined by determining the location of points at which the slope changes from positive to negative. Let the magnitude of the spectrum be represented by $\mid S(f) \mid$. The values of $\mid S(f) \mid$ are evaluated at any three consecutive frequencies denoted by $f_{j-1}$, $f_j$ and $f_{j+1}$, where $j = 1, 2, ..., \frac{M}{2}$ and $M$ is the size of FFT. Let's define $\Delta_j^-$ and $\Delta_j^+$ as

$$\Delta_j^- = \mid S(f_j) \mid - \mid S(f_{j-1}) \mid, \quad (1)$$

$$\Delta_j^+ = \mid S(f_{j+1}) \mid - \mid S(f_j) \mid . \quad (2)$$

The frequency $f_j$ corresponds to an extremum point if the following relations hold

$$\Delta_j^- > 0, \quad (3)$$

$$\Delta_j^+ < 0. \quad (4)$$

In fact, this method finds all of the extrema including noisy ones by computing the slope of the magnitude spectrum at discrete points. A refining algorithm is then applied to remove the small noisy extrema. According to this algorithm, all small peaks of spectrum falling down to 25 dB of adjacent peaks are removed. Since several small peaks may appear consecutively, we cannot suppress all of them by simply comparing them with the adjacent peaks in order to decide if they are eligible to be removed. The refining algorithm removes the small peaks iteratively and one by one by repeating the following two steps:

1. If $\mid S_p(f_{j+1}) \mid > \mid S_p(f_j) \mid$ and $\frac{|S_p(f_{j+1})|-|S_p(f_j)|}{|S_p(f_{j+1})|} > \sigma$ then: $\mid S_p(f_j) \mid = 0$.
2. If $\mid S_p(f_j) \mid > \mid S_p(f_{j+1}) \mid$ and $\frac{|S_p(f_j)|-|S_p(f_{j+1})|}{|S_p(f_j)|} > \sigma$ then: $\mid S_p(f_{j+1}) \mid = 0$.

where $\{S_p(f_j), j = 1, 2, ...\}$ are the peaks of spectrum at the corresponding frequencies and $\sigma$ is the suppression factor (i.e., $\sigma = 0.95$). The refining algorithm finds the normalized

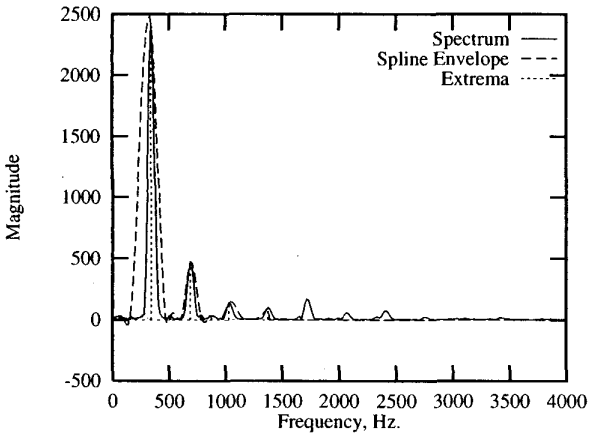


Figure 1: Extrema and the spline envelope in the baseband of 1500 Hz of magnitude spectrum of speech before refining.

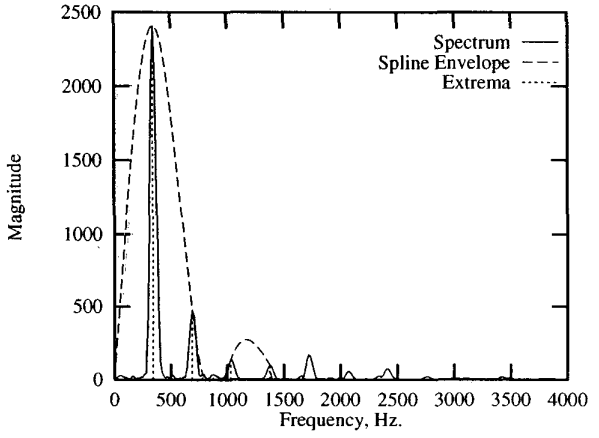


Figure 2: Extrema and spline envelope in the baseband of 1500 Hz after refining.

distance between two adjacent extrema by dividing their difference to the magnitude of the larger one. If the normalized distance is bigger than the suppression factor, then the smaller extremum is removed. To remove all of small extrema, the refining steps are repeated until no change in the number of extrema is observed during the last two iterations. The measured extrema are interpolated using the cubic spline function [6] and an all-pole magnitude spectrum given by

$$| H(e^{j\omega}) | = | \frac{G}{1 + \sum_{k=1}^{p} a_k e^{-j\omega k}} |, \qquad (5)$$

where $G$ is the gain factor and $\{a_k, k = 1, 2, ..., p\}$ are the linear predictive (LP) coefficients, is fitted to the resulting envelope. Spline interpolation is a valuable tool for representing empirical curves and it yields smooth interpolating curves without large oscillations. Figures 1 and 2 represent the speech magnitude spectrum and its spline envelope before and after applying the extremum refining recursion,

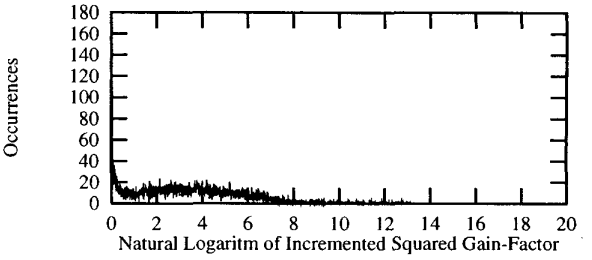


Figure 3: Distribution of $\ln(1 + G^2)$.

respectively. In the sequel, we describe quantization procedures of the all-pole model parameters.

## 3. QUANTIZATION OF GAIN FACTOR

The energy of underlying sinusoidal components of speech spectrum is directly proportional to $G^2$. Since the ear responds logarithmically to the sine-wave amplitudes, the logarithm of $G^2$ is quantized. Distribution of $\ln(1 + G^2)$ is plotted for the frames of 3 minutes of speech in Fig. 3. The high value of distribution at the origin corresponds to the silent frames. Ignoring this peak at the origin, we can assume that the resulting distribution is uniform. Hence, uniform quantization can be used to quantize the log-square of the gain factor (i.e., $\ln G^2$). A uniform quantization scheme using 6 bits with a step size of 0.1 (or 1 dB) according to

$$q_g = \begin{cases} 0, & \text{for: } \ln G^2 < 1.2 \\ int(\frac{\ln G^2 - 1.1}{0.1}), & \text{for: } 1.2 \leq \ln G^2 < 7.4 \\ 63, & \text{for: } \ln G^2 \geq 7.4, \end{cases} \qquad (6)$$

where $q_g$ is the quantization index, works well. Note that the function $int(x)$ in (6) returns the integer part of $x$.

## 4. QUANTIZATION OF LP COEFFICIENTS

The LSF representation of the LP parameters proposed by Itakura [7] contains some useful properties such as well-behaved dynamic range, localized spectral sensitivity and simple stability analysis of LP synthesis filter. These properties make the LSF domain suitable to be used for the quantization of the LP parameters. A computationally efficient method described in [5] is commonly used in speech coding to convert the low order (usually tenth order) LP coefficients to the corresponding LSFs. This method maps the upper semicircle in the z-plane to the real interval $x \in [-1, +1]$ using

$$x = \cos \omega, \qquad (7)$$

where $0 < \omega < \pi$ determines the LSFs. The efficiency of the method proposed in [5] is conditioned on avoiding non-linear mapping from $x$-domain into the $\omega$-domain. Therefore, the resulting LSFs are quantized in the $x$-domain. But, highly resolved low LSFs in the $\omega$-domain map on very closely spaced points in the $x$-domain due to the highly nonlinear behaviour of (7) at low frequencies. This requires high resolution $x$-domain quantization especially when the population of low LSFs in a high-order LSF set is considerable.

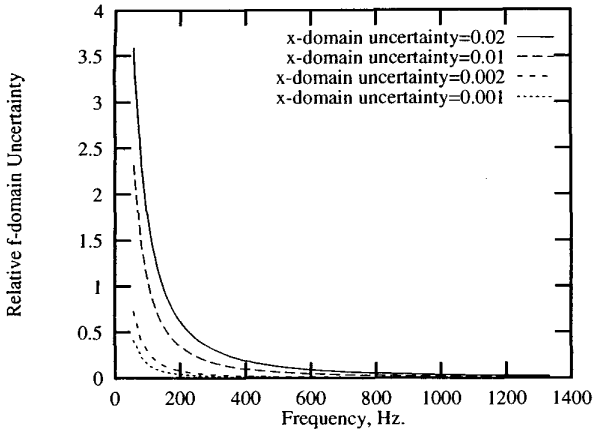


Figure 4: Relative frequency domain uncertainty variations over the low frequencies due to the cosine nonlinearaity.

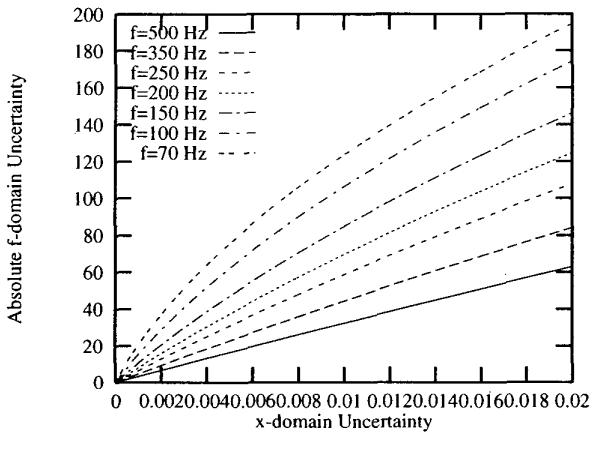


Figure 5: Absolute frequency domain uncertainty variations versus $x$-domain uncertainties due to the cosine nonlinearity.

Let the relative frequency domain (i.e., $f$-domain) uncertainty be defined by

$$\frac{\delta f}{f} = \frac{|\arccos(x) - \arccos(x - \delta x)|}{\arccos(x)}, \tag{8}$$

where $\delta x$ and $\delta f$ are the $x$-domain and the absolute $f$-domain uncertainties. Fig. 4 represents the relative $f$-domain uncertainty plotted versus frequency $f$ given by

$$f = \frac{f_s}{2\pi}\arccos(x), \tag{9}$$

where $f_s$ is the sampling frequency, for a range of the $x$-domain uncertainties. As shown in this figure, the relative uncertainty increases steeply over the frequencies below 400 Hz. The absolute $f$-domain uncertainty given by

$$\delta f = \frac{f_s}{2\pi}[|\arccos(x) - \arccos(x - \delta x)|], \tag{10}$$

is also plotted in Fig. 5 versus $x$-domain uncertainties. As a result of these demonstrations, it can be concluded that a uniform uncertainty in the $x$-domain leads to a highly non-uniform and frequency-dependent uncertainty in the $f$-domain particularly at low frequencies. Hence, the method described in [5] leads to the coarse quantization of the low LSFs especially when a high order LSF set is to be quantized. Vocal tract envelope of speech spectrum is usually modelled by the transfer function of a tenth-order all-pole filter to reflect up to first 5 formants in the bandwidth of 4 KHz of speech spectrum. The lowest two LSFs fall in the vicinity of the first formant which varies in the range between 270 Hz and 730 Hz for vowels [8]. Since this range is broad and, and furthermore, most of the first formants are greater than 400 Hz [8], the coarse quantization of the low LSFs, due to the cosine nonlinearity, has been tolerable in the LPC based speech codecs.

On the other hand, since the human pitch period varies between 2.5 msec (or pitch frequency of 400 Hz) and 20 msec (or pitch frequency of 50 Hz), the first peak location of a voiced spectrum can start from a frequency as low as 50 Hz. Therefore, the peak envelope could contain peaks at frequencies much lower than the average location of the first formant. This, in turn, moves the LSFs towards lower frequencies. Furthermore, encoding the low frequency spectral peaks in the sinusoidal analysis demands a good accuracy and resolution, as the periodic energy of the voiced speech is most concentrated at low frequency range of spectrum.

In an algorithm proposed in [9], the LSFs are estimated in the $f$-domain with any uniform accuracy over the whole range of speech bandwidth. In this algorithm, we exploit the following concept from complex analysis [10].

> *The number of zeros of a function $F(z) = 0$ enclosed by a closed curve $\Delta$ in the z-plane is equal to the number of times (n) that $F(\Delta)$, i.e., the curve obtained by evaluating the polynomial $F(z)$ along the closed curve $\Delta$, rotates about the origin.*

Since we are looking for the roots on the unit circle, the closed curve of interest is defined as a radial sector, consisting of two radial boundaries and a closing arc, such that it encloses the roots. The algorithm starts with an initial radial sector covering the whole interval of $\theta = (0, \pi)$. The initial sector is then bisected into two sectors and the number of roots of function $F(z)$ fallen inside each sector is computed. The bisection procedure followed by computation of the number of enclosed roots is repeated until a single root is isolated within a sector with an angle not larger than the required accuracy [9]. The number of complex polynomial evaluations to estimate a single LSF using this algorithm is approximated by [11]

$$n_b = 12\log_2(\frac{f_s}{p\delta_f}), \tag{11}$$

where $\delta_f$ is the frequency domain resolution. Fig. 6 (top) shows the magnitude spectrum of a voiced speech frame with a fundamental frequency of 136.7 Hz and its spline envelope. The spline envelope is parameterized using 42-nd order linear predictive analysis. The resulting LP parameters are transferred to the LSF domain, using the algorithm proposed in [9], with a uniform accuracy of $\delta_f = 0.5$ Hz.

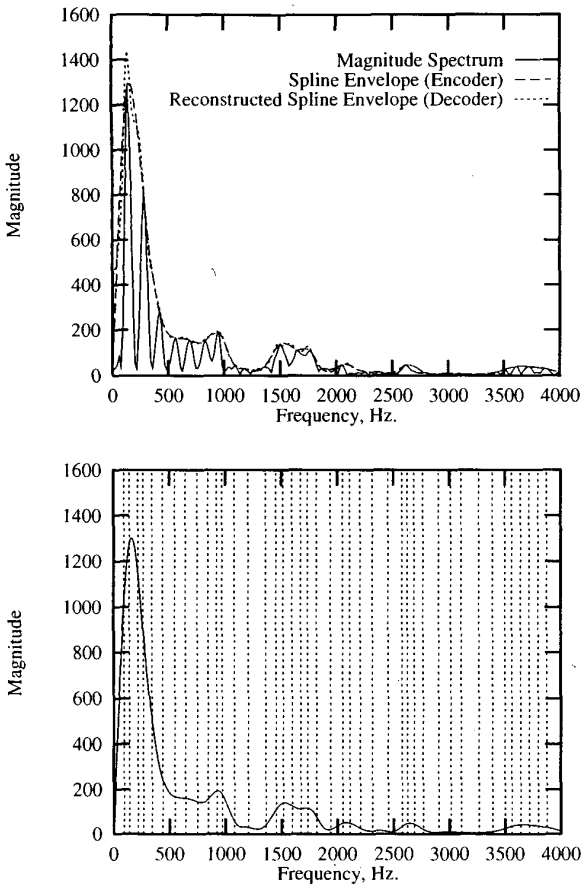


Figure 6: Reconstruction of the spline peak envelope; Top: speech magnitude spectrum with original spline peak envelope and its 42-nd order reconstruction, Bottom: original spline peak envelope and the corresponding 42 LSFs obtained by the proposed algorithm with accuracy 0.5 Hz.

The resulting 42-nd order LSFs are represented in Fig. 6 (bottom) by the vertical dashed lines. The spline envelope is, then, reconstructed from the resulting LSFs using the LSF to LP conversion algorithm [11] and shown in Fig. 6 (top). One point of interest, here, is the concentration of the LSFs at low frequency range.

Due to the polynomial evaluations at complex points, the computaional burden of the LSF estimation algorithm introduced in [9] is higher than that of the method described in [5]. However, these computations do not need high precision arithmetic, as a coarse evaluation which could locate the value of the polynomial anywhere within an area of 45° radial sector in the complex plane would be sufficient. The algorithm introduced in [9] is flexible to be used only over the low frequency range of spectrum where the method in [5] does not have a reliable performance. Since, the mid-range frequency LSFs can be efficiently estimated by the method given in [5], the joint application of these two algorithms leads to complexity and accuracy compromise, where necessary, especially in estimating high-order LSF sets.

The resulting LSFs can be quantized using split vector quantizatin [11].

## 5. CONCLUTIONS

Extremum samples of magnitude spectrum of speech are picked and then refined to obtain those ones which represent the amplitudes of the underlying sine waves. The spline envelope fitted to the resulting extrema is modelled in terms of a gain factor and a set of high-order LP coefficients. To account for the low frequency extrema representing the amplitudes of the underlying high energy sine wave components of speech spectrum, a new algorithm is used to convert the LP coefficients to the corresponding LSFs. In this way, the LSFs can be estimated with any desired uniform accuracy along the bandwidth of speech spectrum.

## 6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Sinusoidal Coding," *Speech Coding and Synthesis*, Chapter 4, W. B. Kleijn and K. K. Paliwal, Editors, Elsevier, The Netherlands, 1995.

[2] M. R. nakhai and F. A. Marvasti, A Hybrid Speech Coder Based on CELP and Sinusoidal Coding," *IEICE Trans. Information and Systems*, Vol. E82-D, No 8, pp. 1190-1199, August 1999.

[3] M. R. Nakhai and F. A. Marvasti, Split Band CELP (SB-CELP) Speech Coder," *IEEE Int. Conf. Acoustic, Speech and Signal Processing*, Vol. 1, pp. 461-464, March 1999.

[4] M. R. Nakhai and F. A. Marvasti, A 4.1 kb/s Hybrid Speech Coder," *IEEE Int. Symp. Circuits and Systems*, Vol. 3, pp. 110-113, May-June, 1999.

[5] P. Kabal and P. Ramachadran, The Computation of Line Spectral Frequencies using Chebyshev Polynomials," *IEEE Trans. Acoustic, Speech, and Signal Processing*, Vol. 34, No 6, pp. 1419-1426, Dec. 1986.

[6] J. Stoer and R. Bulirsch, "Introduction to Numerical Analysis," *Springer-Verlag*, 1980.

[7] F. Itakura, Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Amer.*, Vol. 57, p. S35, April. 1975.

[8] L. Rabiner and R. Schafer, "Digital Processing of Speech Signals," *Prentice-Hall, Inc., Englewood Cliffs, NJ*, 1978.

[9] M. R. Nakhai and F. A. Marvasti, "A Novel Algorithm to Estimate the Line Spectral Frequencies from LPC Coefficients," *Proc. IEEE Int. Symp. on Circuits and Systems* , Monterey,California, Vol. 4, pp. 198-201, June 1998.

[10] P. Henrici, "Applied and Computational Complex Analysis," *Wiley*, Volume 1, 1974.

[11] M. R. Nakhai, "A Low Bit Rate Speech Codec for Wireless Applications," *PhD Thesis, submitted to the University of London*, Chapter 6, December 1999.