Optimization of the CELP Model in the LSP Domain

Khosrow Lashkari and Toshio Miki

DoCoMo USA Labs
lashkari@docomolabs-usa.com

Abstract

This paper presents a new Analysis-by-Synthesis (AbS) technique for joint optimization of the excitation and model parameters based on minimizing the closed loop synthesis error instead of the linear prediction error. By minimizing the synthesis error, the analysis and synthesis stages become more compatible. Using a gradient descent algorithm, LSPs for a given excitation are optimized to minimize the error between the original and the synthesized speech. Since the optimization starts from the LPC solution, the synthesis error is guaranteed to be lower than that obtained using the LPC coefficients. For the ITU G.729 codec, there is about 1dB of improvement in the segmental SNR for male and female speakers over 4 to 6 second long sentences. By adding an extra optimization step, the technique can be incorporated into the LPC, multi-pulse LPC and CELP-type speech coders.

1. Introduction

Low bit rate speech coders use the source filter model based on the speech production mechanism [1]. This model consists of a vocal tract (synthesis) filter driven by an excitation signal $u(n)$ as shown in figure 1. The filter takes into account the effects of the vocal tract consisting of the glottis, the mouth, the lips and the nasal cavity on speech production. The excitation signal tries to mimic the air pressure variations generated by the vocal folds. The vocal tract and the excitation are assumed to have fixed characteristics over an interval of 10ms to 30ms but can change between successive intervals. This model has been the basis for successful speech coding systems such as LPC [2], multi-pulse LPC [3], CELP [4] and MELP [5]. The vocal tract filter is approximated by an all-pole linear filter of the form $H(z) = G / A(z)$ where $A(z)$ is a polynomial of degree M and G is a gain term representing the speech energy over the frame.

$$A(z) = 1 + \sum_{k=1}^{M} a_k z^{-k} \qquad (1)$$

$$\hat{s}(n) = -\sum_{k=1}^{M} a_k s(n-k) + Gu(n) \qquad (2)$$

The excitation signal depends on the coding scheme. In the LPC vocoder, the excitation signal $u(n)$ is assumed to consist of a periodic pulse train at the pitch period (P) for voiced sounds and random noise for unvoiced sounds. For the multi-pulse LPC, the excitation signal consists of a set of pulses distributed within the analysis frame. For the CELP encoder, the excitation function is selected from a codebook of possible excitations. In the LPC-based techniques, filter parameters $\{a_k\}$, $k = 1,2,...M$ are computed by minimizing the total linear prediction (LP) error E_p [2]. The sample prediction error $e_p(n)$ and total prediction error E_p are defined as:

$$e_p(n) = s(n) + \sum_{k=1}^{M} a_k s(n-k) \qquad (3a)$$

$$E_p = \sum_{n=0}^{N-1} e_p^{\,2}(n) \qquad (3b)$$

Where, N is the length of the analysis window. Similarly, the sample synthesis error $e_s(n)$ and total synthesis error E_s shown in figure 1 are defined as:

$$e_s(n) = s(n) + \sum_{k=1}^{M} a_k \hat{s}(n-k) - Gu(n) \qquad (4a)$$

$$E_s = \sum_{n=0}^{N-1} e_s^{\,2}(n) \qquad (4b)$$

As seen from (3a), LP minimization uses only the speech signal $s(n)$. The synthesis error in (4a) on the other hand depends on the excitation $u(n)$, the speech signal $s(n)$ and the reproduced speech $\hat{s}(n)$. The prediction error at time n ($e_p(n)$) is a linear function of the filter parameters $\{a_k\}$, however, the synthesis error ($e_s(n)$) is *not* a linear function of these parameters; simply because the synthesized speech $\hat{s}(n)$ itself depends on these parameters.

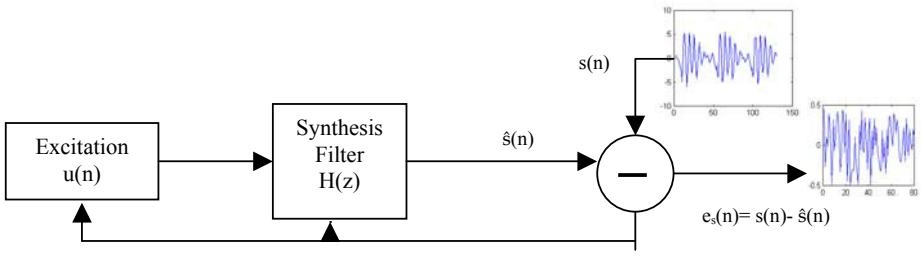

Figure 1: Source-filter model and the synthesis error

More specifically, $\hat{s}(n)$ is an n-th degree polynomial in parameters a_1 through a_M, making the synthesis error in (4b) a highly nonlinear function of these parameters. Hybrid speech coders such as CELP compute the parameters $\{a_k\}$ by minimizing the linear prediction error in (3b). Using these parameters, the excitation signal $u(n)$ is then determined by minimizing the synthesis error in (4b). The proposed technique computes both the filter parameters and the excitation signal by minimizing the synthesis error in (4b).

2. Minimization in the LSP domain

In a previous paper [6], an iterative minimization technique based on the gradient descent algorithm in the root domain was presented. The algorithm in [6] computes the filter coefficients $\{a_k\}$ using the autocorrelation method of linear prediction. Next, roots of the LPC filter (A(z)) are computed. The synthesis error in (4b) is then minimized starting from these initial LPC roots. An efficient version of this algorithm with applications to the ITU G.729 codec is presented in [7]. The main disadvantages of the algorithms in [6] and [7] are that complex roots have to be found and complex arithmetic must be used in the optimization process. Furthermore, it is not clear how to incorporate the filter memory and the LSP interpolation (used in the CELP-type speech codecs) into the optimization. The algorithm presented in this paper performs the minimization in the LSP domain and alleviates the above-mentioned problems.

Let $\Gamma^{(i)}$ denote the LSP vector at the i-th iteration of the gradient descent algorithm:

$$\Gamma^{(i)} = \left[\gamma_1^{(i)} ..... \gamma_k^{(i)} ..... \gamma_M^{(i)}\right]^T \qquad (5a)$$

Here, $\gamma_k^{(i)}$ is the value of the k-th LSP at the i-th iteration of the gradient descent algorithm and T stands for transpose. The algorithm starts from:

$$\Gamma^{(0)} = \left[\gamma_1^{(0)} ..... \gamma_k^{(0)} ..... \gamma_M^{(0)}\right]^T \qquad (5b)$$

where, $\Gamma^{(0)}$ is the LSP vector corresponding to the LPC solution. To compute $\Gamma^{(0)}$ we convert the LPC coefficients to LSPs using the transformations given in equations (12) and (13). The LSPs at iteration (i+1) are given as:

$$\Gamma^{(i+1)} = \Gamma^{(i)} + \mu_i \frac{\nabla_i E_s}{\left|\nabla_i E_s\right|} \qquad (6a)$$

At the i-th iteration, μ_i is the step-size and $\nabla_i E_s$ is the gradient vector of the synthesis error relative to the LSP vector $\Gamma^{(i)}$. As seen from (6a), we normalize the gradient vector by its norm. This normalization ensures that the difference between the LSP vectors at successive iterations is bounded by μ_i, that is:

$$\left|\Gamma^{(i+1)} - \Gamma^{(i)}\right| = \mu_i \qquad (6b)$$

From (4b), the gradient vector can be computed as:

$$\nabla_i E_s = \sum_{k=1}^{N-1} (s(k) - \hat{s}(k)) \nabla_i \hat{s}(k) \qquad (7a)$$

Thus, the gradient of the synthesis error can be computed in terms of the gradients of the synthesized speech samples $\hat{s}(k)$.

$$\nabla_i \hat{s}(k) = [\partial \hat{s}(k) / \partial \gamma_1^{(i)} ..... \partial \hat{s}(k) / \partial \gamma_M^{(i)}] \qquad (7b)$$

The terms $\partial \hat{s}(k) / \partial \gamma_r^{(i)}$ are the partial derivatives of $\hat{s}(k)$ with respect to the r-th LSP at the i-th iteration. From (2), these terms can be computed recursively as follows.

$$\partial \hat{s}(k) / \partial \gamma_r^{(i)} = -\partial \left[\sum_{j=1}^{M} a_j \hat{s}(k-j)\right] / \partial \gamma_r^{(i)} \qquad (8a)$$

Carrying the partial derivative inside the sum, we get:

$$\partial \hat{s}(k) / \partial \gamma_r^{(i)} = -\sum_{j=1}^{M} \left[\hat{s}(k-j) \partial a_j / \partial \gamma_r^{(i)} + a_j \partial \hat{s}(k-j) / \partial \gamma_r^{(i)}\right] \qquad (8b)$$

Define:

$$p(k,r) = \partial \hat{s}(k) / \partial \gamma_r \qquad (9a)$$

Similarly, define the partial derivatives of coefficients a_j relative to the r-th LSP γ_r as:

$$d(j,r) = \partial a_j / \partial \gamma_r \qquad (9b)$$

Now (8b) can be written as (we have dropped the superscript i for notational simplicity):

$$p(k,r) = -\sum_{j=1}^{M} \left[\hat{s}(k-j) d(j,r) + a_j p(k-j,r)\right] \qquad (10a)$$

With the initial conditions:

$$p(k,r) = 0 \qquad \text{for} \qquad k < 0 \qquad (10b)$$

Hence, the partial derivatives $p(k,r)$ can be recursively calculated once the quantities $d(j,r)$ are known. The values of $p(k,r)$ for $k = 0,1,...M-1$ depend on the initial conditions $\hat{s}(-M),..... \hat{s}(-1)$. Thus, the initial conditions are taken into account.

The partial derivatives from (10a) are substituted into (7b) to compute the vector of partial derivatives $\nabla_i \hat{s}(k)$. Next, (7b) is substituted into (7a) to compute the gradient vector $\nabla_i E_s$. Finally, (7a) is substituted in (6a) to update the LSP vector $\Gamma^{(i)}$. The algorithm starts from the initial vector $\Gamma^{(0)}$ corresponding to the LPC solution and continues this process until some termination or optimality criterion has been satisfied such as when the step-size μ_i becomes

smaller than a predefined value ε or after a predetermined number of iterations is reached. In general, the step-size μ_i is not constant; it is adaptively adjusted at each iteration. Initially (at iteration zero) we start with a large step-size. If at any iteration, the synthesis filter becomes unstable or the synthesis error exceeds its value at the previous iteration, we go back to the parameter vector at the previous iteration and reduce the step-size by a known factor (of two for example). This process of adjusting the step-size continues until the filter becomes stable again or the error becomes smaller than its value at the previous iteration.

2.1. Partial Derivatives of $\{a_k\}$ w.r.t the LSPs

LSPs are the roots of the two polynomials $P(z)$ and $Q(z)$ defined below. For even model order M, polynomials $P(z)$ and $Q(z)$ can be written as:

$$P(z) = 1 + \sum_{k=1}^{M} p_k z^{-k} = \prod_{k=1}^{M/2} (1 - 2\cos\omega_{2k-1} z^{-1} + z^{-2}) \quad (11a)$$

$$Q(z) = 1 + \sum_{k=1}^{M} q_k z^{-k} = \prod_{k=1}^{M/2} (1 - 2\cos\omega_{2k} z^{-1} + z^{-2}) \quad (11b)$$

Where, ω_k, $k = 1,2,....M$ are the Line Spectrum Frequencies (LSFs) and $\gamma_k = \cos(\omega_k)$, $k = 1,2,....M$ are the Line Spectrum Pairs (LSPs). Roots of $P(z)$ give the odd LSPs and roots of $Q(z)$ the even LSPs, $\{p_k\}$ and $\{q_k\}$ are the coefficients of $P(z)$ and $Q(z)$ respectively.

Because of their desirable quantization properties, LSPs are used as the transmission parameters in most parametric speech codecs. For even M, we can express the coefficients $\{p_k\}$ and $\{q_k\}$ in terms of the filter coefficients $\{a_k\}$ as:

$$p_M = p_0 = 1 \quad (12a)$$

$$p_{M-k} = p_k = -p_{k-1} + (a_k + a_{M+1-k}) \quad k = 1,2..M/2 \quad (12b)$$

$$q_M = q_0 = 1 \quad (13a)$$

$$q_{M-k} = q_k = q_{k-1} + (a_k - a_{M+1-k}) \quad k = 1,2..M/2 \quad (13b)$$

The coefficients $\{a_k\}$ can be expressed in terms of the coefficients $\{p_k\}$ and $\{q_k\}$ as follows:

$$a_k = (p_k + p_{k-1} + q_k - q_{k-1})/2 \quad k = 1,2...M \quad (14)$$

Using (14), we can write:

$$d(j,r) = \partial a_j / \partial \gamma_r = (\partial p_j / \partial \gamma_r + \partial q_j / \partial \gamma_r + \partial p_{j-1} / \partial \gamma_r - \partial q_{j-1} / \partial \gamma_r)/2 \quad (15)$$

2.1.1. Derivatives of $\{p_k\}$ and $\{q_k\}$ w.r.t the LSPs

Let $\gamma_r = \cos(\omega_r)$ be the r-th LSP. Then the polynomials $P(z)$ and $Q(z)$ in (11a) and (11b) can be written as:

$$P(z) = \prod_{k=1}^{k=M/2} (1 - 2\gamma_{2k-1} z^{-1} + z^{-2}) \quad (16a)$$

$$Q(z) = \prod_{k=1}^{k=M/2} (1 - 2\gamma_{2k} z^{-1} + z^{-2}) \quad (16b)$$

Let $P_r(z)$ be the polynomial of degree $(M-2)$ defined by dropping the r-th factor $(1 - 2\gamma_{2r-1} z^{-1} + z^{-2})$, that is:

$$P_r(z) = 1 + \sum_{k=1}^{M-2} p'_{kr} z^{-k} = \prod_{k=1, k\neq r}^{M/2} (1 - 2\gamma_{2k-1} z^{-1} + z^{-2}) \quad (17a)$$

where, p'_{kr} are the coefficients of $P_r(z)$ and are calculated from the right hand side of (17a) once the LSPs are known. It is easily verified from (16a) that:

$$\partial P(z) / \partial \gamma_r = -2 z^{-1} P_r(z) \quad (17b)$$

From (17b), we see that the coefficients $\{p'_{kr}\}$ of the polynomial $P_r(z)$ are scaled derivatives of the coefficients $\{p_k\}$ with respect to the r-th LSP. More specifically:

$$\partial p_j / \partial \gamma_r = -2\, p'_{j-1,r} \quad , j = 1,3,...(M-1) \quad (17c)$$

For $j = 0$, we have:

$$\partial p_0 / \partial \gamma_r = 0 \quad (17d)$$

where $p'_{0r} = 1$ for all r.

Let $Q_r(z)$ represent polynomial of degree $(M-2)$ defined by dropping the r-th factor $(1 - 2\gamma_{2r} z^{-1} + z^{-2})$, that is:

$$Q_r(z) = 1 + \sum_{k=1}^{M-2} q_{kr} z^{-k} = \prod_{k=1, k\neq r}^{M/2} (1 - 2\gamma_{2k} z^{-1} + z^{-2}) \quad (18a)$$

Where, q'_{kr} are the coefficients of $Q_r(z)$ and are calculated from the right hand side of (18a) once the LSPs are known. It is easily verified from (16b) that:

$$\partial Q(z) / \partial \gamma_r = -2 z^{-1} Q_r(z) \quad (18b)$$

From (18b), we see that the coefficients $\{q'_{kr}\}$ of the polynomial $Q_r(z)$ are scaled derivatives of the coefficients $\{q_k\}$ with respect to the r-th LSP. More specifically:

$$\partial q_j / \partial \gamma_r = -2 q'_{j-1,r} \quad , \quad j = 2,4,...M \quad (18c)$$

and

$$\partial q_0 / \partial \gamma_r = 0 \quad (18d)$$

where, $q'_{0r} = 1$ for all r. The largest step-size $\mu_{i,\max}$ of the gradient descent algorithm can be determined as:

$$\mu_{i,\max} = \min \left| \gamma_{j+1} - \gamma_j \right| / 2 \quad , \; j = 1,2,.....M-1 \quad (19)$$

Derivatives of the coefficients $\{p_k\}$ and $\{q_k\}$ from (17c) and (18c) are substituted into (15) to compute the partial derivatives $d(j,r)$ of $\{a_j\}$ relative to the LSPs. For joint optimization, the excitation signal $u(n)$ can be recomputed after LSPs are updated. Alternatively, the excitation signal may be recomputed only after a certain number of iterations. For example, the excitation signal could be updated only after the optimization criteria for the LSPs are met.

3. Results

Figure 2 shows the optimization result for a 25ms segment of voiced speech (part of nasal "m"). Here, the original waveform is shown together with the 10th order multi-pulse LPC and SEM synthesized waveforms using one pulse per millisecond. As shown, the SEM synthesized speech is closer to the original speech waveform than the synthesis using multi-pulse LPC. Figure 3 shows that after only six iterations, the synthesis error is reduced by 37% corresponding to 2dB of improvement in the signal-to-noise ratio. The top (solid) graph in figure 3 shows the synthesis error at different iterations of the gradient descent algorithm. Since the search algorithm starts from the LPC solution, synthesis error at the first iteration is the same as the LPC synthesis error, which is assumed to be 100%. The error increases between the second and the third iteration indicating that the minimum has been overshot. In such a case, the algorithm uses the parameters at the previous step, re-adjusts the step-size μ_i and continues with the search. As figure 3 demonstrates, this leads to a reduction in the synthesis error after the third iteration. The bottom (dashed) graph in figure 3 shows the percent improvement in the synthesis error at different iterations. Figure 4 shows the spectrum of the original speech and the spectra derived from the LPC and the SEM-optimized filter coefficients. The SEM-based spectrum produces a peak at 280Hz that is closer to the spectral peak of the original speech. This is consistent with figure 2 that shows a closer match to the low frequency part of the speech waveform. Figure 5 shows the improvement in the frame SNR for the ITU G.729 codec for a 6-second long female sentence. Here, order is 10, frame length is 10ms and 8 pulses per frame are used. As seen here, for some frames, the SNR improvement is well above 1 dB (2-5dB)

4. Conclusion

A new technique for joint optimization of the LSPs and the excitation signal was presented. Minimization in the LSP domain offers several unique advantages.

First, the root domain optimization described in [7] assumes distinct roots, that is A(z) has no repeated roots. Second, LSP interpolation at the subframe boundaries can be easily incorporated into the optimization procedure. Third, LSP computation is simpler than root finding and uses real arithmetic. Fourth, LSPs are a more appropriate set of parameters for optimization. In the root domain optimization, if the initial LPC roots are distinct real roots, the optimization cannot make them complex. With LSPs this is possible. Fifth, LSPs can be optimized using real arithmetic. Finally, in most of the state-of-the-art speech codecs such as the ITU G.729 and ETSI AMR, LSPs are used as the final parameters for quantization and transmission. The resultant synthesis error is guaranteed to be lower than that obtained using the LPC filter for both clean and noisy speech and the optimized synthesis filter is guaranteed to be stable. The technique is applicable to any excitation signal and can be incorporated into LPC, multi-pulse LPC, CELP and MELP-type speech coders.

5. References

[1] Fant, G.C.M., "Acoustic Theory of Speech Production", Mouton and Co.,s-Gravenhage, The Netherlands, 1960.

[2] Atal, B.S. and Hanauer, S.L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, Vol. 50, pp. 637-655.

[3] Atal, B.S. and Remde, J.R. "A New Model of LPC Excitation for Producing Natural- Sounding Speech at Low Bit Rates," *Proc. ICASSP'82, pp*. 614-617.

[4] Schroeder, M.R. and Atal, B.S. "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates," *Proc. IEEE ICASSP'85, 25.1.1, pp*. 937-940, April 1985.

[5] McCree, A.V. and Barnwell III, T.P. "Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," *IEEE Trans. On Speech and Audio Processing, vol. 3,* pp. 242-250, July 1995.

[6] Lashkari, K and Miki, T." Joint Optimization of Model and excitation in Parametric Speech Coders," *Proc. ICASSP 2002*.

[7] Lashkari, K. and Miki, T. "Joint Optimization of Model and Excitation in CELP-type Speech Coders," *Record of the36th Asilomar Conference*, pp.191-195.

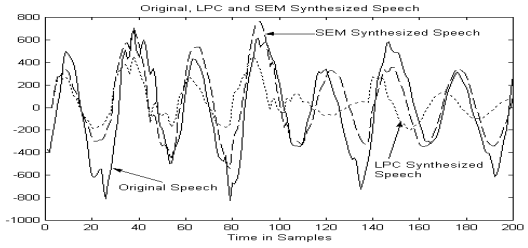

Figure 2: LPC and SEM waveforms

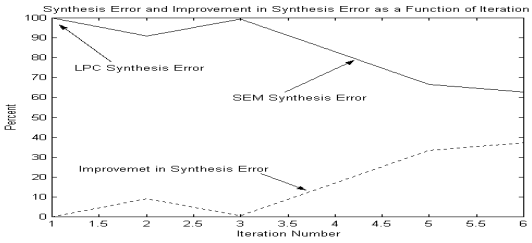

Figure 3: Synthesis Error at Different Iterations

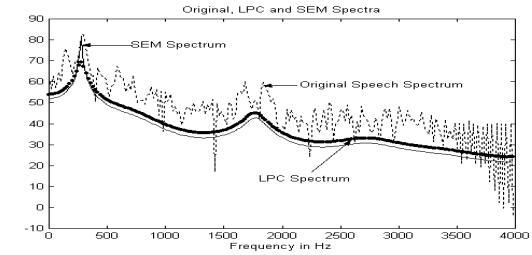

Figure 4: Original, LPC and SEM Spectra

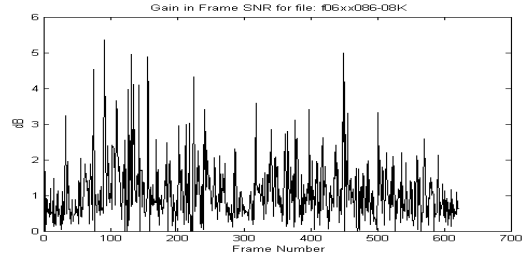

Figure 5: SNR Improvement for SEM-optimized G.729