Proc. Biennial Symp. Commun. (Kingston, ON), pp. 308-311, June 1990

LOW–RATE ANALYSIS–BY–SYNTHESIS WIDEBAND SPEECH CODING †

Guylain Roy [1] and Peter Kabal [1,2]

[1] Electrical Engineering
McGill University
Montreal, Quebec, H3A 2A7

[2] INRS-Télécommunications
Université du Québec
Verdun, Quebec, H3E 1H6

Abstract

This paper presents possible implementations for a low rate wideband *analysis–by–synthesis* speech coder. The wideband speech signals have a bandwidth of 8 kHz, and the target operating bit rate is 16 kbits/sec. The basic Residual Excited Linear Predictive coder (RELP) is used as a starting point to develop and test flexible pitch parameter optimization procedures, which can operate in either full-band or split-band mode. These procedures are then applied to an analysis–by–synthesis CELP (Code Excited Linear Prediction) model. The performance of full-band and split-band CELP structures are compared.

1. Introduction

In recent years, analysis–by–synthesis coders have been developed for narrowband (0–4 kHz) systems, and have achieved high quality speech reproduction at rates from 4.8 kbits/sec to 9.6 kbits/sec [1]. So far, little work has been done in wideband analysis–by–synthesis systems. However, since roughly 80% of the perceptually important speech spectral information is contained within the baseband (0.2 – 3.2 kHz) [2], it is safe to assume that the incremental cost incurred coding the extra bandwidth of a wideband signal should be relatively small. The added bandwidth increases the perceived speech quality, and helps discriminate between fricatives (e.g "f" vs "s"). Potential applications for this type of coder include mobile telephone, videophone and voice-mail services.

In this paper, the coding of the extra bandwidth is first analyzing using the RELP (Residual Excited Linear Predictive) speech coder model [3]. This is useful in two ways. First, it demonstrates that the upper 4 kHz of a wideband signal can be reproduced at low cost and at little perceived degradation. Second, it helps derive and test a set of flexible optimization procedures to be later used in either a full-band or split-band CELP (Code Excited Linear Prediction) context.

2. Enhanced RELP coding

In the original RELP configuration, the near sample redundancies of the discrete input speech $s(n)$ are removed by a short-term linear prediction filter $1 - F(z)$ (i.e. formant prediction). The filter coefficients a_k are obtained using a standard LPC analysis, done over frames of 10 to 20 ms duration [4]. The short-term prediction error $d(n)$ (formant residual) is a noise-like, spectrally flat signal, with an embedded train structure corresponding to the pitch period.

The formant residual signal is low-pass filtered. The upper band is simply discarded. The baseband residual signal is then decimated by R, quantized and sent to the receiver, along with the LPC coefficients. At the receiver, a High Frequency Regeneration (HFR) scheme is applied to the baseband residual to artificially recreate the discarded upper band information. The regenerated residual $\hat{d}(n)$ then excites the all-pole formant synthesis filter $H(z) = 1/(1 - F(z))$.

The quality of RELP coded speech strongly depends on the HFR scheme used. Various HFR methods have been proposed: spectral folding, spectral translation, non-linear functions [5]. No one method is clearly better than the others and all have been used with some degree of success for medium quality narrowband coders, operating in the 4.8 to 9.6 kbits/sec range.

2.1 Addition of pitch prediction

When pitch prediction is used, the far sample redundancies of the discrete input speech $s(n)$ are removed by a long-term linear prediction filter $1 - P(z)$. The predictor

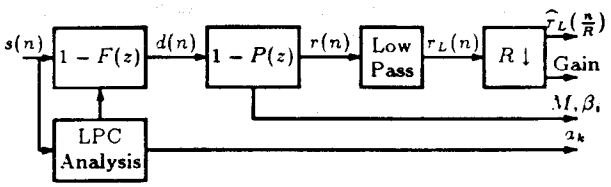
$$P(z) = \sum_{i=1}^{N_p} \beta_i z^{-(M+i-1)},$$

where N_p is the number of pitch coefficients (between 1 and 5) and M the pitch lag, is updated every pitch sub-frames of 2.5 to 5 ms duration. The pitch lag takes on values within a predetermined range. Its smallest value is usually limited by the pitch sub-frame size. Generally, the pitch lag ranges from 2.5 ms to 20 ms. When cascaded with the formant prediction filter, the pitch prediction error $r(n)$ (pitch residual) has little or no pitch structure left. The pitch structure is re-inserted by the pitch synthesis filter $G(z) = 1/(1 - P(z))$.

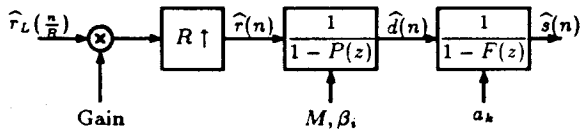
Figure 1 shows the structure of the pitch enhanced RELP model. Since the transmitted baseband residual has no harmonic structure, the HFR scheme is reduced to simple spectral folding (i.e. upsampling by R).

When compared, in a wideband context, to the basic RELP model, this approach strongly reduces the harmonic discontinuities at the spectral folding points. The resulting speech $\hat{s}(n)$ does not sound as metallic, and the clicks and pops have been replaced by a more uniform degradation. This degradation is most noticeable when a small residual baseband is preserved (e.g. 1000 Hz), but becomes less important when 4 kHz of baseband is transmitted. This confirms the original assumption that upper band portion of the excitation signal $\hat{r}(n)$ has a lower perceptual impact than the baseband.

† This work was supported by FCAR.



(a) Analysis phase



(b) Synthesis phase

Fig. 1 RELP Coder with Pitch Prediction

3. Split-band optimization

The quality of the reproduced speech is affected by the modification of the relation between the prediction coefficients and the residual. The excitation signal $\hat{r}(n)$ appearing before the pitch and formant synthesis filters is no longer optimal with respect to its original analysis coefficients. Given this known sub-optimal excitation signal, the problem is then to find a new set of synthesis parameters that produce the best possible reconstructed speech. The formant parameters, because of the low–delay feedback induced at the synthesis filter, yield a highly non-linear set of equations and are not re-optimized. However, the gain and pitch parameters can easily be re-adjusted under certain conditions explained below.

Consider the model shown in Figure 2. The excitation source now consists of two separate signals $\hat{r}_L(n)$ and $\hat{r}_H(n)$. These are obtained by splitting the interpolated pitch residual $\hat{r}(n)$ into its original baseband and its interpolated high band. This dual excitation scheme offers separate optimization control over each band. In the case of RELP, this is especially desirable since, when compared to the original full band residual, the transmitted baseband is not as distorted as the regenerated high band. The low band parameters are thus somewhat isolated from the potential adverse effects of the less optimal high band. The bandwidth expansion factor $\gamma = 1/0.75$ emphasizes the formant regions thus rendering the coding noise less perceivable.

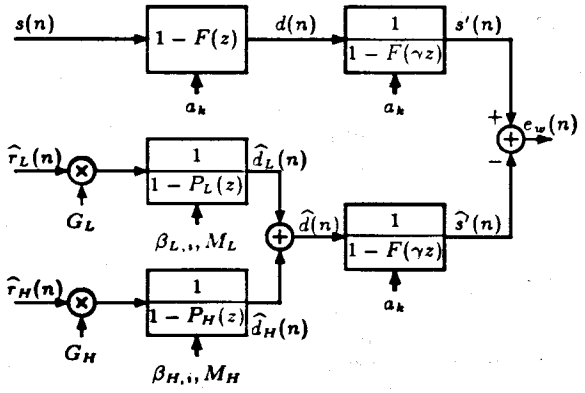


Fig. 2 Split-band optimization

The parameters subjected to optimization are the low and high band pitch coefficients $\beta_{L,i}$ and $\beta_{H,i}$, pitch lags M_L and M_H and gain factors G_L and G_H. This structure provides flexible control over N_{pL} and N_{pH}, the number of pitch coefficients in each band.

3.1 Mathematical description

Given the excitation signals, the goal is to minimize the energy of the error $e_w(n)$ for every pitch sub-frame of N samples. This error is the difference between the bandwidth expanded original and reconstructed speech signals $s'(n)$ and $\hat{s}'(n)$:

$$\begin{aligned} s'(n) &= \sum_{k=-\infty}^{\infty} d(k)h'(n-k), \\ \hat{s}'(n) &= \sum_{k=-\infty}^{\infty} \hat{d}(k)h'(n-k). \end{aligned} \tag{1}$$

where $h'(n)$ is the impulse response of the bandwidth expanded formant synthesis filter. This impulse response is time-varying. However, since the minimization is done at the pitch sub-frame level, $h'(n)$ is fully known and held constant for the duration of the sub-frame. Also, $s'(n)$ is also known for the duration of the sub-frame. Therefore, the summation limits can be changed to 0 and $N-1$, provided the contribution of past sub-frame excitation samples (i.e. $k < 0$) are preserved as initial conditions for the current sub-frame. This is achieved by saving the formant synthesis filter internal memory from one sub-frame to the next.

The ouputs of the low and high band pitch synthesis filters are combined to form the regenerated formant residual signal $\hat{d}(n)$, expressed as:

$$\begin{aligned} \hat{d}(n) = G_L\hat{r}_L(n) &+ \sum_{i=1}^{N_{pL}} \beta_{L,i}\hat{d}_L(n - M_L - i + 1) \\ + G_H\hat{r}_H(n) &+ \sum_{i=1}^{N_{pH}} \beta_{H,i}\hat{d}_H(n - M_H - i + 1). \end{aligned} \tag{2}$$

The pitch lags M_L and M_H must both be larger than the pitch sub-frame size. This prevents any feedback that would cause the equations to become non-linear. Then, $\hat{d}(n)$ can be viewed as a linear combination of all the known waveforms $\hat{r}_L(n)$, $\hat{r}_H(n)$, $\hat{d}_L(n-M_L-i)$ and $\hat{d}_L(n-M_H-i)$. The bandwidth expanded regenerated speech $\hat{s}'(n)$ can then be expressed as:

$$\begin{aligned} \hat{s}'(n) = &\sum_{k=-\infty}^{-1} \hat{d}_L(k)h'(n-k) + \sum_{k=-\infty}^{-1} \hat{d}_H(k)h'(n-k) \\ &+ \sum_{k=0}^{\infty} \hat{d}_L(k)h'(n-k) + \sum_{k=0}^{\infty} \hat{d}_H(k)h'(n-k), \end{aligned} \tag{3}$$

The anti-causal terms in the above equation are the zero-input responses of the bandwidth expanded formant synthesis filter, and account for the initial conditions of each band at the pitch sub-frame boundaries. The impulse response $h'(n)$ is causal, and the upper limit in both causal terms summations can be set to $N-1$. Defining the following terms,

$$\begin{aligned} x_L(n) &= \sum_{k=0}^{N-1} \hat{r}_L(k)h'(n-k), \\ x_H(n) &= \sum_{k=0}^{N-1} \hat{r}_H(k)h'(n-k), \end{aligned} \tag{4}$$

and

$$\begin{aligned} y_{L,i}(n) &= \sum_{k=0}^{N-1} \hat{d}_L(k - M_L - i + 1)h'(n-k), \\ y_{H,i}(n) &= \sum_{k=0}^{N-1} \hat{d}_H(k - M_H - i + 1)h'(n-k), \end{aligned} \tag{5}$$

the weighted error $e_w(n)$ can be expressed as:

$$e_w(n) = s^*(n) - G_L x_L(n) - \sum_{i=1}^{N_{pL}} \beta_{L,i} y_{L,i}(n), \\ - G_H x_H(n) - \sum_{i=1}^{N_{pH}} \beta_{H,i} y_{H,i}(n), \quad (6)$$

where $s^*(n)$ contains all the terms not subjected to optimization:

$$s^*(n) = s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k) h'(n-k). \quad (7)$$

The optimization is done in the mean-square sense. Let the energy of the weighted error signal $e_w(n)$ in the pitch sub-frame be:

$$\xi = \sum_{n=0}^{N-1} e_w(n)^2. \quad (8)$$

Differentiating the above equation with respect to the gain and the pitch coefficients and setting it equal to 0 yields, for any given pitch lag values, a linear system of equations. This is best represented in matrix form, $\Phi \mathbf{v} = \mathbf{b}$. Let †:

$$\mathbf{v}^T = [G_L, G_H, \beta_{L,1}, \ldots, \beta_{L,N_{pL}}, \beta_{H,1}, \ldots, \beta_{H,N_{pH}}], \\ \mathbf{b}^T = [s^* x_L, s^* x_H, s^* y_{L,1}, \ldots, s^* y_{L,N_{pL}}, \\ s^* y_{H,1}, \ldots, s^* y_{H,N_{pH}}], \quad (9)$$

and

$$\Phi = \mathbf{q}\mathbf{q}^T, \quad (10)$$

where

$$\mathbf{q}^T = [x_L, x_H, y_{L,1}, \ldots, y_{L,N_{pL}}, y_{H,1}, \ldots, y_{H,N_{pH}}].$$

The matrix Φ is symmetric and can be solved using the Cholesky factorization technique. Since the solution vector $\mathbf{v}$ depends on the pitch lags (from Eq. 5), the overall optimal solution is obtained through an exhaustive search of all possible lag values.

3.2 Evaluation

With no quantization and as little as 2 kHz of transmitted baseband, the split-band optimization scheme yields higher quality speech than the model in Section 2.1. This demonstrates the usefulness of the re-optimizing the synthesis parameters.

Problems arise however, when the transmitted baseband is quantized. Although optimization reduces the overall error, this optimized RELP system still requires very good coding of the baseband. This is possible in narrowband systems where typically 1 kHz of baseband is transmitted. In wideband, at least 2, preferably 4 kHz of baseband should be well modeled. Unfortunately, no scalar quantization scheme can do this economically enough (i.e. < 2 bits/sample).

Consequently, a vector quantization (VQ) approach must be used. If the parameters are optimized with respect to a single selected codebook vector, the overall performance is highly limited by the size and quality of the VQ codebook. It is then just a natural step to search the whole codebook to find the codeword that generates the best speech rather than the codeword that best matches the original baseband residual. This is CELP.

† For clarity, all index references are left out. Thus, $s^* x_L$ refers to $\sum_{n=0}^{N-1} s^*(n) x_L(n)$

4. Wideband CELP

The above optimization procedure can be adapted to a CELP coder structure. In essence, the pitch residual signals $\hat{r}_L(n)$ and $\hat{r}_H(n)$ are modeled by Gaussian waveforms selected from pre-defined codebooks. Thus, for each codebook entry, the linear system in Eq. 10 is solved. The index of the codewords yielding the lowest error are then transmitted to the receiver.

4.1 Frame and sub-frame sizes

The input speech is sampled at 16 kHz. The frame and sub-frame sizes control the update rate of all the coding parameters. Two approaches are used in this research. In the first, the frame and sub-frame sizes are set to 320 and 40 samples (50 and 400 Hz) respectively . In the second case, the frame rate is increased to 64 Hz while the sub-frame rate is set at 320 Hz. These values are typical for narrowband coders.

4.2 LPC coefficients coding

The LPC coefficients a_k are coded using Line Spectral Frequencies (LSF's) [6]. These are a mathematical transformation of the direct form coefficients a_k. Moreover, LSF's are always ordered for stable synthesis filters and thus, stability can easily be ensured after quantization. For wideband speech, 16 coefficients are used to model the spectral envelope, and a non-uniform differential scalar quantization scheme is used [7]. Since the LSF's are related to the formants positions, allocating more bits for the lower LSF's emphasizes the perceptually important lower frequencies. During the simulations, 50 to 70 bits/frame are used. This figure could be reduced through inter and intra-frame interpolation [1].

4.3 Pitch coefficients coding

The allocation of pitch coefficients for the low and high bands varies between 1:1, 3:1 and 3:0. In the last case, no pitch coefficient is used for the high band (i.e. $N_{pH} = 0$). The coder is then operating in full-band mode, since the speech must be reconstructed strictly from the low band contributions.

The computed optimal pitch coefficients are coded with 3 to 5 bits non-uniform scalar quantizers. Quantization is done before the error energy ξ (Eq. 8) is calculated. The quantization error is thus accounted for within the optimization. Note that this is optimal if no more than one pitch coefficient is used in each band. Otherwise, a fully optimal solution would involve cascaded searches through all the possible quantized values for each pitch tap.

4.4 Lag estimate and coding

The optimal split-band solution is computationally heavy. Indeed, the system of equations must be solved for each band, each lag value and each codebook waveform. This is impractical. A slightly less optimal, yet more efficient approach, is to solve for the optimal lag values with both gains set to zero [1]. This eliminates the burden of nested lag and waveform index searches. The lag estimate is then performed by letting the pitch synthesis filters *free-wheel* with the past regenerated formant residual signals $\hat{d}_L(n)$ and $\hat{d}_H(n)$.

For practical purposes, a common lag value is optimized for both bands. Transmission of the lag usually requires a heavy 7 or 8 bits per pitch sub-frame. There is some indication however, that the lag could be updated every other sub-frame. Indeed, from sub-frame to sub-frame, the computed optimal lags either remain around a certain value, or jump erratically to near-multiples of this value (e.g. 48,100,153,...).

4.5 Codeword design and selection

For this study, the codebook size is varied from 32 to 1024 waveforms. When operating in full-band mode, the codewords are normalized Gaussian sequences, and the optimal entry is found through an exhaustive search.

In split-band mode, the lower and higher band codebooks are respectively low and high pass filtered Gaussian sequences. The cutoff frequency is set at 4 kHz. This restricts the contribution of each codeword to its respective band. The lower band codeword selection is exhaustive. The upper band selection can either be exhaustive, in which case an index must be transmitted, or arbitrary, in which case the optimal low band index is used to pick the high band codeword.

4.6 Gain estimate and coding

In split-band mode, the gains G_L and G_H can be separately coded, or forced to be the same. A 3 to 5 bits differential quantizer with a leaky predictor (1 tap $\alpha = 0.9$) is used to code the difference in sucessive sub-frame gain magnitudes. An extra bit codes the sign. The computed gains are quantized before calculating the error energy ξ (Eq. 8). This ensures the overall best solution under quantization constraints.

5. Experimental results

The wideband CELP coder is simulated in floating-point on a general purpose workstation. The wideband speech signals are bandlimited to 8 kHz, digitized and stored in files, and contain material from 2 male and 2 female speakers. Two utterances from each speaker were used to train the pitch, gain and LSF quantizers. Tests were first performed without quantizing the pitch, gain and LSF's. This helped determine the relative performance of split and full-band structures, as well as the effects of the frame update rate, number of pitch taps and seperate gain optimization. Then, the parameters were quantized to produce a split and a full-band coder, both operating at 16 kbits/sec.

5.1 No parameter quantization

Here, the first observation is that, either in split or full-band mode, the speech quality is slightly better with faster frame and sub-frame updates. Second, the split-band method yields a better reproduced speech than the full-band approach. The segmental Signal to Noise Ratio (segSNR) increases by a few dBs. Also, the use of 3 low band pitch parameters instead of 1 considerably increases the performance. Thus, it seems worth to reduce the update rate slightly in order to accomodate the extra parameters.

Finally, in split-band operation, performance is increased when the upper band gain is separately optimized, particularly when the upper band waveform is obtained arbitrarily.

Figure 3 shows the segSNR variations for 10 ms frames with no parameter quantization. The top trace is the energy of the original signal. The bottom dashed trace is the segSNR for a full band coder, operating with a codebook size equal to 32 and 1 pitch tap. The bottom solid trace if for a split-band, gain optimized coder with a pitch tap in each band, and a codebook size equal to 32. The split-band segSNR is on average 2 dB higher than the full-band segSNR.

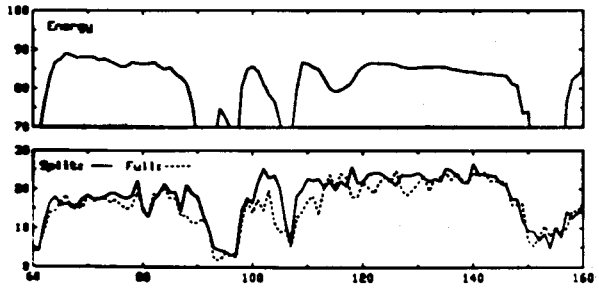


Fig. 3 SegSNR track for split and full band models

5.2 With parameter quantization

At high frame update rates, there is an obvious tradeoff between the number of bits/parameters and the output quality. With formant frames of 10 ms and sub-frames of 2.5 ms, the performance is best if a single pitch parameter is used (one in both bands for split-band operation), and if the gains are identical. The resulting speech is degraded and is comparable to 5-bit wideband log-PCM. However, the coder distorsion is of a *hollow* nature while the reference wideband log-PCM degradation is of a *hiss* nature.

At slower update rates (20 ms formant, 5 ms pitch), there is added flexiblity. Using 3 pitch taps and larger codebooks (256–512 levels), the resulting speech lies somewhere around 7 bit wideband log-PCM. There is still nonetheless, a noticeable difference between the original and coded speech. However, this could probably be cured by more efficient parameter quantization techniques, especially for LSFs.

6. Conclusion

The implementation of a wideband speech coder operating at 16 kbits/sec has been demonstrated. Pitch parameter optimization techniques have been developed for both split and full-band systems. These procedures are most useful for analysis-by-synthesis models, but could also improve the performance of simpler narrowband RELP systems. In wideband CELP, the use of split-band is generally performs better than a similar full-band system.

References

1. P. Kabal, J.L. Moncet and C.C. Chu, "Synthesis Filter Optimization and Coding: Applications to CELP," *ICASSP-88*, section S4.2, pp. 147–150
2. D. O'Shaughnessy, *Speech Communication. Human and Machine*, Addison-Wesley, 1987
3. C.K. Un and D.T. Magill, "The Residual Excited Linear Predictive Vocoder with Transmission Rate Below 9.6 kbits/s," *IEEE Trans. Communications*, vol. COM-23, pp. 1466–1474, December 1975.
4. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
5. C.K. Un and J.R. Lee, "On Spectral Flattening Techniques in Residual Excited Linear Prediction Vocoding," *ICASSP-82*, pp. 216–219, May 1982.
6. F.K. Soong and B.H. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression," *ICASSP-84*, pp. 1.10.1–1.10.4.
7. N. Sugamura and N. Favardin, "Quantizer Design in LSP Speech Analysis-Synthesis," *IEEE Journal on Selected Areas in Communication*, vol. 6, no. 2, pp.432–440, February 1988.