ANALYSIS BY SYNTHESIS SPEECH CODING WITH GENERALIZED PITCH PREDICTION[1]

Paul Mermelstein and Yasheng Qian

INRS-Télécommunications
Université du Québec, Verdun, Québec H3E 1H6
Canada

ABSTRACT

A new analysis-by-synthesis speech coding structure is presented for high-quality speech coding in the 4 to 8 kb/s range. CELP with generalized pitch prediction (GPP-CELP) differs from classical code-excited linear prediction (CELP) in that for voiced segments it is the speech signal that is decomposed into a component predictable with the aid of the adaptive codebook (ACB) and a nonpredictable aperiodic component, not the LPC residual. The spectrum of the aperiodic component is estimated by linear-prediction analysis. An approximation to the aperiodic component is synthesized from a stochastic codebook of sparse pulse sequences and its spectrum is shaped by the LPC synthesis filter. The ACB contains samples of the past reconstructed signal, low-passed to increase the pitch prediction gain. For voiced segments the new structure yields higher pitch prediction gain and lower linear-prediction gain than classical CELP. Subjective and objective comparisons reveal significant advantages for GPP-CELP over classical CELP.

1. INTRODUCTION

Most modern speech coding techniques are based on the CELP paradigm. It permits synthesis of a replica of the signal to be coded by linear filtering an excitation signal composed of predictable and nonpredictable (stochastic) components. The predictable component is selected from the ACB which contains samples of the past excitation signal. Accordingly, the sequence of analysis operations is linear-prediction analysis followed by pitch prediction analysis. We investigate an alternative analysis-by-synthesis structure, one in which pitch (long-term) prediction analysis precedes linear-prediction (short-term) analysis. In both structures closed-loop pitch analysis is used, i.e., the segment to be coded is correlated with information derived from the past reconstruction. To minimize the energy of the stochastic component, we introduce a linear filtering of the past reconstructed signal, a low-pass filter that attenuates the higher frequencies where the pitch harmonics are generally weaker. Pitch prediction with the aid of such a linear filter in the feedback loop is termed generalized pitch prediction (GPP) since it enables the exploitation of the frequency-dependence of the long-term correlation to enhance the predictable component.

The inversion of the order of short-term and long-term prediction was previously explored by [1] in an open-loop analysis environment and no gain in total prediction gain was noted. However, executing the long-term filtering first significantly increases the pitch prediction gain relative to the case where the short-term predictor is applied first. The prediction gain of the subsequent short-term predictor is simultaneously reduced so that the long and short-term prediction gains take on values that are much more comparable than in the classical structure [2]. The periodic component is generally more energetic than the aperiodic component for voiced speech. Since only the weaker aperiodic component is shaped by the LPC filter, the sensitivity of the reconstructed signal to quantization errors in the LPC filter is reduced, leading to higher quality or reduced quantization requirements.

We noted previously [3] that the pitch gain in voiced speech manifested significant frequency dependence. The pitch prediction error could be reduced by modeling the pitch gains in the frequency bands above 1 kHz as varying linearly in time with the pitch gain determined for the band up to 1 kHz. The pitch prediction filter which includes the low-pass filter in the feedback loop implements such a frequency-dependent pitch prediction gain.

This paper reports on the results of objective and subjective evaluations of the quality of speech achieved with the new structure when none of the coding pa-

[1] This work was supported by the Bell-Quebec/Nortel/NSERC Industrial Research Chair in Personal Communications and a grant from Nortel.

rameters is quantized. Only the size of the stochastic codebook is controlled in terms of the number of independently adjustable pulses permitted per time-frame. Theoretical considerations are reviewed in the next section. The representation of the stochastic codebooks is considered in section 3. Section 4 discusses details of the GPP and classical coding structures. Objective comparisons of extracted parameters and subjective comparisons of the resynthesized speech are given in Section 5. Section 6 provides discussion of the results and conclusions.

2. THEORY OF GENERALIZED LINEAR PITCH PREDICTION

Let the speech signal segment i(n), n = 1,...,N be approximated as a sum of an ACB component a(n-L) processed by an FIR filter with impulse response $h_f(m)$ and an SCB component b(n) processed by an all-pole filter with impulse response $h_r(k)$,

$$i(n) = g_p \cdot a(n-L) * h_f(m) + r(n), \quad (1)$$
$$r(n) = g_{pl} b(n) * h_r(k) + e_s(n). \quad (2)$$

where $*$ denotes convolution, g_p and g_{pl} are appropriate pitch and pulse gains, $e_s(n)$ is the stochastic synthesis error and L is the pitch lag in samples. To determine the impulse response $h_f(m)$ of the fixed low-pass filter, we ignore the SCB component and minimize the energy of the unpredicted component,

$$\sum_n r^2(n) = \sum_\omega |I(\omega) - g_p \cdot A(\omega) \cdot H_f(\omega)|^2 \quad (3)$$

where $I(\omega)$, $A(\omega)$, and $H_f(\omega)$ are the Fourier transforms of i(n), a(n-L) and $h_f(m)$, respectively. For a linear-phase FIR filter, $H_f(\omega)$ is real and may be estimated from

$$H_f(\omega) = Re\{A(\omega) \cdot I^*(\omega)\}/g_p \cdot |A(\omega)|^2 \quad (4)$$

Suitable averaging of the desired transfer function over a large number of voiced frames yields a target transfer function with attenuation at 2 and 4 kHz of 3 and 12 dB, respectively.

3. REPRESENTATION OF THE STOCHASTIC COMPONENT

3.1 Voiced Segments

We now attempt to approximate the pitch residual in terms of a sparse sequence of pulses exciting an all-pole filter with impulse response $h_r(k)$. The pitch residual r(n) found above may be approximated as in Eq. 3 by minimizing the energy of the stochastic synthesis error $e_s(n)$.

$$\sum_n e_s^2(n) = \sum_n [r(n) - g_{pl} b(n) * h_r(k)]^2 \quad (5)$$

The impulse response $h_r(k)$ is obtained by the customary linear-prediction analysis of r(n) and the residual b(n) so derived can be approximated by a sparse sequence of pulses. such that

$$g_b b(n) = \sum_j g_{bj} \delta(n - N_j) + e_{sp}(n) \quad (6)$$

as in multipulse excitation [4]. The sequence $e_{sp}(n)$ represents the stochastic codebook approximation error. The best pulse sequence may be selected by minimizing the energy of a perceptually weighted error,

$$e_{pw}(n) = [i(n) - g_p a(n-L) * h_f(m) - \sum_j g_{bj} \delta(n - N_j) * h_r(k)] * h_{pw}(l), \quad (7)$$

where $h_{pw}(l)$ is the impulse response of the perceptual weighting filter [5]. Available techniques for this purpose are discussed by Salami [6]. A suboptimal but practically acceptable sequential pulse placement technique is one that first minimizes the perceptually weighted error using one pulse alone, removes the contribution of this pulse from the input signal and then places a second pulse so as to minimize the remaining error.

The LPC filtered stochastic codebook contribution is generally not orthogonal to the ACB contribution. A small improvement in the quality of voiced sounds may be obtained by readjusting the pitch gain g_p after the stochastic codebook contribution has been determined. Comparative evaluations for sequential pulse placement and unquantized parameters are provided in section 5.

4.2 Unvoiced Segments

For unvoiced segments the sparse pulse representation may lead to a significant underestimate of the required excitation energy. Also, the perceptually weighted error drops very slowly as additional pulses are introduced to the SCB representation. In general, it is more efficient to select one of a small number of white-noise sequences spanning the frame duration to excite the LPC synthesis filter. Thus the error function to be minimized becomes

$$e_{sn}(n) = [i(n) - g_n c_{xn}(n) * h_i(k)] * h_{pw}(l) \quad (8)$$

where g_n is an appropriate noise gain, $c_{xn}(n)$ is a white noise sequence selected from a Gaussian codebook and

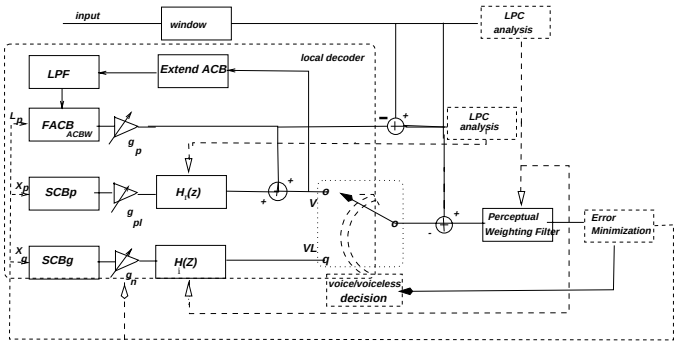

Figure 1: GPP-CELP encoder structure

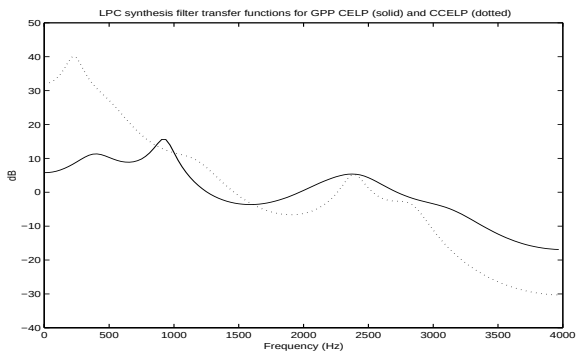

Figure 2: Comparison of LPC filter transfer functions determined for a speech segment and its pitch-prediction residual

$h_i(k)$ is the impulse response of the all-pole filter derived from the input signal.

The closed-loop pitch search will find some small correlation even for noiselike fricative segments. Once frication is no longer present in the input signal, an SCB component is sought that cancels this noisy ACB contribution. Sparse pulse sequences cannot do so effectively. Also, when the input signal contains only background noise it is preferable to eliminate the ACB contribution to the output altogether. Therefore, we cut the ACB feedback loop for frames in which the excitation is generated from noise segments. Keeping the ACB free of excitation noise has the added benefit of reducing fluctuations in the pitch lag estimates near the voicing onsets.

4. THE GPP-CELP CODING STRUCTURE

The new GPP-CELP coding structure is presented in Fig 1. Note that the decoder block is contained within the encoder structure. The ACB is prefiltered by $h_f(m)$ to allow the filtering to be ignored when carrying out the closed-loop pitch search. When the pitch lag is shorter than the windowed segment selected for analysis, it is desirable to extend the ACB into the future assuming a periodic continuation. Such extension may introduce small discontinuities in the selected ACB component when the segment that best matches the input straddles the time-reference point. Best results are obtained when the past reconstructed samples are placed into an unfiltered buffer, the contents of the buffer are extended into the future, and the extended ACB sequence is low-pass filtered and realigned in time. The voiced/unvoiced decision can be based on the relative energy of the input signal within the frame and the extent to which the allotted pulses suffice to reduce the stochastic synthesis error. For the moment, however, our subjective evaluations were carried out with sentences coded with hand-marked voiced/voiceless information. Synthesis of the unvoiced segments follows the classical CELP structure, the synthesis filter is a quantized form of the analysis filter derived from the input segment. The perceptual weighting filter is always derived from the LPC analysis of the input. By extracting the ACB feedback signal prior to the voiced or unvoiced component selection, feedback into the ACB of the gaussian SCB is avoided.

A comparison of the transfer functions of $H_i(z)$ and $H_r(z)$ for a typical voiced segment, as shown in Fig. 2, reveals that the pole bandwidths of $H_r(z)$ are broadened resulting in more rapidly decaying impulse responses $h_r(k)$. One effect of this difference is that LPC synthesis does not spread a pulse from the stochastic codebook as widely in time in the new structure, which in turn allows better representation of transient events in the signal. More rapidly decaying impulse responses also reduce the interaction between pulses in terms of their contributions to minimizing the perceptually weighted error and thereby allow the results of sequential pulse placement to better approximate the results of the optimal search over all possible pulse vectors.

5. COMPARATIVE EVALUATIONS

The performance of the GPP coder was compared to a classical CELP coder designed to resemble the GPP coder in all aspects not essential to the structural modifications. The same source data is used as input to both coders and all windowing operations are maintained identical. For best quality time-overlapped windowing is used and the synthesized signal is generated by overlapped addition of the segments derived for successive time frames. Input segments are windowed to 6.25 ms and spaced 5 ms apart.

The variation of pitch prediction gain with time is

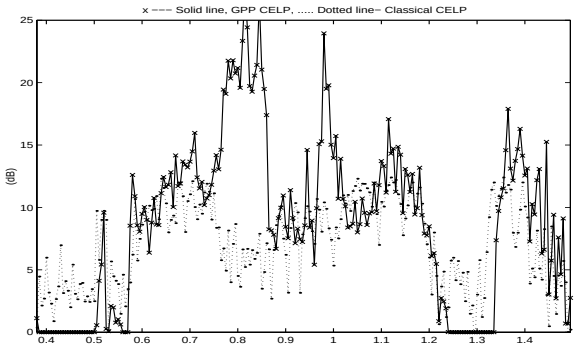

Figure 3: Pitch prediction gains for GPP and classical CELP encoders for sucessive 5 ms analysis frames

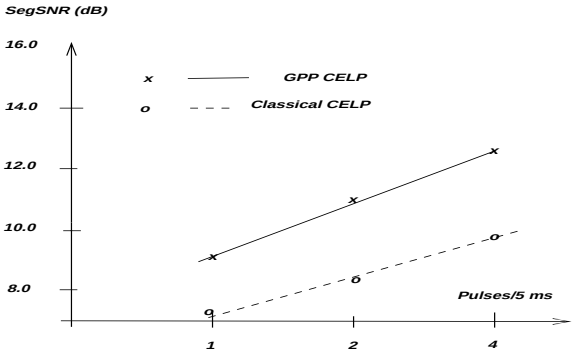

Figure 4: Segmental SNR values averaged over one sentence for GPP and classical CELP coders as functions of the SCB pulse densities

shown for the two coders in Fig.3. The increased gains for the GPP structure for voiced intervals are apparent. Decreased gains are noted for the unvoiced intervals since the pitch gain there is set to zero. The segmental SNR (average of the log SNR over individual frames) is calculated for the two coders and shown in Fig.4 for codebooks of one, two and four pulses per 40 samples, respectively. Similar segmental SNR values near 9 dB are attained by one pulse per frame with the GPP coder and by two pulses per frame with the classical CELP coder. The SNR advantage of the new structure increases with higher pulse densities.

A small subjective evaluation experiment was run with 6 listeners rating for preference 8 presentations of each of 4 sentences in ABAB or BABA formats. In each case A represented the product of the GPP-CELP synthesis and B classical CELP synthesis, both using 2 pulses per 5 ms and no other parameter quantization. In 63% of the presentations the GPP product was preferred. Experienced listeners showed an even stronger preference for GPP-CELP.

6. DISCUSSION AND CONCLUSIONS

The most important advantage of the GPP synthesis model is that little SCB energy is required for voiced sequences with nearly constant pitch and spectrum once the oscillations have built up. The aperiodic component only serves to start up the oscillations and to adjust the output as changes are required. The LPC information is used to shape the difference signal between the current input segment and that delayed by one pitch period. It may be argued that the spectrum of this difference is less likely to be all-pole than the spectrum of the actual speech. However, in both cases the all-pole spectral estimation serves only to approximate the desired spectral characteristics. The stochastic codebook sequences provide the phase information that allows the output signal to be generated with the correct spectral magnitude and phase, as ensured by the time-domain error minimization in each frame.

The GPP structure provides coded speech with good quality for pulse densities of 2 to 4 pulses per 5 ms. Vowel sounds appear slightly muffled due to the widened bandwidths of the higher frequency formants. The voicing onsets typically sound clearer than in classical CELP. We estimate that the quantization requirements can be reduced to 4 kb/s with 2 pulses and to 6.4 kb/s with 4 pulses. Our current goal is to determine the quality attainable with the GPP-CELP coder at those rates.

7. REFERENCES

[1] Ramachandran R.P. and Kabal P. "Pitch Prediction Filters in Speech Coding". *IEEE Trans. ASSP*, vol. ASSP-37, pp. 466-478. 1989.

[2] Nafei Y. "Analysis of Prediction Structures for Speech Coding," *M.Sc. thesis,* INRS-Telecommunications, Université du Quebec. 1998.

[3] Mermelstein P., Zheng P. and Saikaly M. "Multiband Residual Coding of Celp Codecs at 8 kb/s", *IEEE International Conference on Acoustics,Speech and Signal Processing.* Adelaide, Australia, pages Vol. II - 117-120. April, 1994.

[4] Kroon P. and Deprettere E.F. "Experimental Evaluation of Different Approaches to the Multipulse Coder". *IEEE International Conference on Acoustics,Speech and Signal Processing.* San Diego, CA, pages 10.4.1-10.4.4. April, 1984.

[5] Ramamoorthy V. and Jayant N.S. "Enhancement of ADPCM speech by adaptive postfiltering". *Bell System Technical Journal.* 63: 1465-1475. 1984

[6] Salami R.A., Hanzo L., Steele R.,Wong K.H.J and Wassell I., *Speech Coding* in *Mobile Radio Communications,* Steele R. editor, Pentech Press, London, IEEE Press, N.Y., 1992, pages 186-336.