GENERALIZED ANALYSIS-BY-SYNTHESIS CODING AND ITS APPLICATION TO PITCH PREDICTION

W. Bastiaan Kleijn, Ravi P. Ramachandran, and Peter Kroon

AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

Many modifications can be applied to a speech signal without changing its perceptual quality. For a particular speech coder, the coding efficiency will differ for distinct modifications. To exploit this, we introduce a generalized analysis-by-synthesis procedure. In this procedure, a search is performed over a multitude of modified original signals (on a blockwise basis), and the signal which can be encoded with the least distortion is selected for transmission. At the receiver, a quantized version of this modified original signal is constructed.

We discuss the application of generalized analysis-by-synthesis coding to the pitch predictor of a CELP coder. The use of this technique makes it possible to transmit the pitch-predictor parameters at a much lower rate than conventional approaches, without compromising speech quality.

1. INTRODUCTION

In the analysis-by-synthesis method [1], a useful vector of model parameters is obtained by synthesizing a signal for each of a set of such vectors, and selecting the vector for which the synthesized signal resembles a reference signal most closely. This procedure has proven particularly useful in linear-predictive (LP) coding of speech signals. The code-excited linear prediction (CELP) algorithm [2] is the most well-known example of this class of analysis-by-synthesis coders. In these coders, the reference signal is the original speech signal. The resemblance between the original and trial signals is evaluated using a perceptually relevant error criterion. The basic principle of this procedure is shown in Figure 1a.

Here, we propose a generalization of the analysis-by-synthesis procedure, which is illustrated in Figure 1b. A multitude of modified speech signals is generated, with the constraint that each of these signals is perceptually close or identical to the original speech signal. The speech coder performance is evaluated for each of these modified signals, and the modified signal which results in the best coding performance is selected. The model parameters corresponding to this modified signal are transmitted to the receiver.

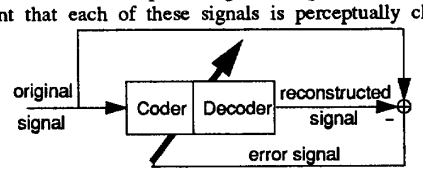

Figure 1a. Conventional analysis-by-synthesis coder.

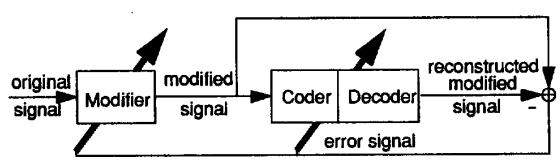

Figure 1b. Generalized analysis-by-synthesis coder. The modified signal is constrained to be perceptually similar to the original signal.

In principle, this generalized analysis-by-synthesis procedure will lead to improved performance for any speech coder. However, it is advantageous to modify the speech coder to exploit the generalized analysis-by-synthesis structure. In particular, coder parameters can be constrained to be the interpolated values from open-loop estimates. The speech can then be modified such that interpolated model parameters provide an accurate match for the modified speech signal. In this paper, we will show with an application how the generalized analysis-by-synthesis procedure can be used to obtain a significant improvement in coding efficiency.

We will apply the generalized analysis-by-synthesis procedure to the pitch predictor (PP). The PP provides a major contribution to the coding efficiency of many LP-based analysis-by-synthesis coders. However, in most of these coders, the PP requires a large proportion of the overall bit rate. This makes a reduction of the bit rate of the PP through generalized analysis-by-synthesis an attractive proposition for low-bit-rate coding.

2. INTERPOLATION OF PITCH-PREDICTOR PARAMETERS

In LP coding, first a set of prediction coefficients is computed. The associated prediction filter is used to obtain the LP-residual signal, $x(t)$. In LP-based analysis-by-synthesis coding, the residual signal is encoded on a blockwise basis. Following standard convention, we will call these blocks *subframes*. For each subframe, the quantized LP-excitation signal, $v(t)$, is obtained by selecting from a collection of candidate excitation signals the one that results in the most accurate reconstruction of the original signal, according to a perceptually based error criterion [3].

The pitch-predictor parameters are an integral part of the analysis-by-synthesis structure [4]. This method of obtaining model parameters is often referred to as *closed-loop* estimation. Figure 2 illustrates the basic synthesis structure of an LP-based analysis-by-synthesis algorithm with a PP. The PP-excitation signal, $e(t)$, is added to a delayed (and scaled) version of the past reconstructed LP-excitation signal, $\lambda v(t-d)$. The resulting LP-excitation signal, $v(t)$, is used as the input for the LP-synthesis filter which adds the formant structure of the speech signal. Speech is synthesized for a multitude of allowed delay values, d, without adding the PP-excitation signal. From this procedure the delay and gain values are selected. The best PP-excitation signal, $e(t)$, is determined after the selection of the PP parameters. In CELP coders, $e(t)$ is selected from a fixed codebook.

0-7803-0532-9/92 $3.00 © 1992 IEEE

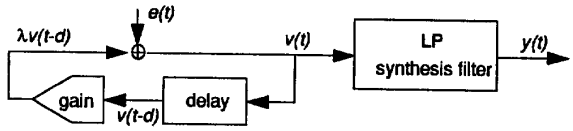

Figure 2. Synthesis structure of an LP-based analysis-by-synthesis speech-coding algorithm, using a pitch predictor.

The PP generally requires a bit rate of 1500 to 2500 b/s, taking a significant share of the overall bit rate. The high bit rate results from the frequent update of the PP parameters. However, the pitch of the speech signal is known to vary slowly, and small deviations in the pitch contour of the reconstructed signal from that of the original signal are not objectionable. This is exploited in LP-based vocoders in which the pitch is interpolated. The close relation between the PP delay and the pitch period suggests that the PP delay is a natural candidate for interpolation.

However, naive interpolation of the PP delay results in a significant decrease in performance because it leads to suboptimal delay values in the individual subframes. The PP maps the past LP-excitation signal into the present subframe. In a conventional closed-loop PP, this mapping tries to put LP-excitation features (such as the pitch pulses) at locations where similar features occur in the LP-residual signal. However, if the delay is suboptimal in this subframe (because of interpolation), then a time mismatch between the features in the LP-residual signal and the LP-excitation signal occurs as is shown in Figure 3. Pitch pulses may be lost or repeated near subframe boundaries. When time mismatches occur, the PP-excitation, $e(t)$, attempts to undo the effects of this time mismatch, rather than to refine the pitch-pulse shape. As a result, time mismatches, as present in naive interpolation of the PP delay, cause significant audible distortion.

The time mismatches result in changes in the dynamics of the pitch-cycle waveform. From coding techniques based on the interpolation of pitch-cycle waveforms [5] we know that a smooth evolution of the pitch-cycle waveform is essential for obtaining natural-sounding voiced speech. However, the same techniques show that small, gradual errors in the pitch period are not disturbing. Thus, it seems natural to modify the original speech signal such that the optimal PP parameters become known functions of time, allowing straightforward interpolation without degrading performance. The proposed modifications of the original signal are minor time warps and amplitude scalings, which do not affect the natural quality of the speech signal.

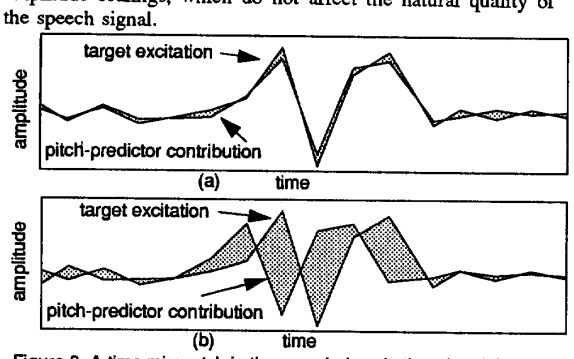

Figure 3. A time mismatch in the sampled excitation signal. In (a) a pitch pulse, and the PP contribution with the correct time alignment are shown. In (b) the PP contribution is moved to the left by one sample. In both cases, the shaded area indicates the error signal.

2.1 Pitch Predictor with Continuous Delay Contour

For a conventional PP, the delay is constant within each subframe, and changes stepwise at the subframe boundaries. We call this a *stepped* delay contour. In this type of delay contour, the changes in delay at the subframe boundaries lead to discontinuities in the PP mapping of the past LP-excitation signal into the present. These discontinuities are an obstacle in our goal of signal reconstruction with a smoothly evolving pitch-cycle waveform, and interfere with the interpolation mechanism. Therefore, we first reformulate the PP to eliminate the discontinuities in the LP-excitation signal. For this new PP, the delay contour has no discontinuities, and we refer to it as a *continuous* delay contour.

The approach proposed here for the determination of the continuous delay contour for the PP is straightforward: select the best of a set of feasible delay contours over the current subframe, all starting at the end value of the delay contour in the previous subframe. Consider now the delay as a continuous function of time: $d(t)$. We denote by t_j, the starting time of subframe j, and restrict the delay contours to be linear within a subframe. For the subframe j, which ranges over $t_j < t \le t_{j+1}$, the instantaneous delay $d(t)$ is then:

$$d(t) = d(t_j) + \alpha\,(t - t_j), \qquad t_j < t \le t_{j+1}, \tag{1}$$

where α is a constant. Denoting a continuous LP-excitation signal by $v(t)$, then, for a given $d(t)$, the unscaled PP contribution to the present subframe LP-excitation signal is:

$$u(t) = v(t - d(t)), \qquad t_j < t \le t_{j+1}. \tag{2}$$

In a digital implementation, the continuous function $u(t)$ is approximated by a sampled version obtained by band-limited interpolation of the past excitation signal.

2.2 Interpolation of the Continuous Delay Contour

It is convenient to perform the time warping on the LP-residual signal rather than the original signal. Consider a particular interpolation interval. The goal is to time warp the LP-residual signal such that the linear delay contour is optimal within this interval.

Instead of determining the delay contour with a search procedure, an *a-priori* delay contour, $d(\tau)$, is defined for the entire interpolation interval, in the warped time domain τ. Because we are dealing with a continuous-delay PP we can use simple, linear a-priori delay contours. To maintain continuity at the interpolation-interval boundaries, the endpoint of the a-priori delay contour of the previous interpolation interval must be the starting point of the a-priori delay contour in the present interpolation interval. To prevent perceptual distortions due to the time warping, it is essential that the a-priori delay contour is a close approximation to the delay contour obtained with a conventional PP. A reasonable a-priori delay contour can be obtained from pitch estimates obtained directly from the original signal (so-called open-loop delay estimates). The a-priori delay-contour can be obtained by linear interpolation of open-loop delay estimates.

The following procedure can be used sequentially for each subframe within an interpolation interval. Because the delay contour is known, the (as yet unscaled) contribution $u(\tau)$ of the PP to the LP-excitation signal in this subframe can be computed without further reference to the original signal, by using the equivalent of equation (2) in the warped time domain:

$$u(\tau) = v(\tau - d(\tau)), \qquad \tau_j < \tau \le \tau_{j+1}, \tag{3}$$

where τ_j is the starting time of subframe j in this domain.

We define the time-warping function $\zeta(t)$:

$$\zeta(t) = \frac{d\tau}{dt} \tag{4}$$

The relationship between the time-warped and original LP-residual signals is then given by:

$$x_\zeta(\tau) = x_\zeta(\tau_j + \int_{t_j}^{t} \zeta(t)dt) = x(t), \qquad t_j < t \le t_{j+1}. \tag{5}$$

In a conventional PP, the PP contribution to the LP-excitation signal is computed by selecting the best delay contour from a codebook of delay contours. In the present method, we search through a codebook of different time-warping functions to obtain the best match of the time-warped LP residual signal to the PP contribution. Any time-warping function $\zeta(t)$ which does not affect the perceptual quality of the original signal can be a codebook entry. To facilitate this requirement we restricted ourselves to continuous time-warping functions. Note that no information about this codebook is transmitted; its size is limited only by the computational requirements.

Because the gain of the PP contribution in the current subframe is not yet known, the criterion used for the selection of a particular time-warping function $\zeta(t)$ must be based on shape only. The normalized correlation between the perceptually-weighted time-warped LP-residual signal, and the similarly-weighted PP contribution can be used for this purpose. Once the proper time warp is determined, the optimal scaling factor for the PP contribution can be computed. The PP-excitation contribution for this subframe can then be obtained as in a conventional analysis-by-synthesis coder, using the time-warped residual signal as reference. Thus, for a CELP coder, the procedure is equivalent to matching the PP and fixed-codebook contributions to the time-warped LP-residual signal.

A proper choice of the codebook with time-warping functions is critical for performance. In particular, the choice must result in nonoscillatory time warping. The subframe boundaries must be located with the same considerations in mind. (Note that when the PP parameters are interpolated the subframe length is independent of the bit rate.) To this purpose, let us consider PP subframes where the pitch pulse is located near the end of the subframe. A good qualitative measure of the local time warp is the ratio of the distances between the pitch pulse in the warped time domain, τ, and in the original time domain, t. Thus, if the pitch pulses are at the boundary points, it is desirable to have time-warping functions which satisfy:

$$\zeta(t_{j+1}) \approx \frac{\tau_{j+1} - \tau_j}{t_{j+1} - t_j} = \frac{\int_{t_j}^{t_{j+1}} \zeta(t)\, dt}{t_{j+1} - t_j}. \tag{6}$$

If the pitch pulses are somewhat before the PP-subframe boundaries, $\zeta(t)$ should maintain its end value in this neighborhood of the PP-subframe boundary. If equation (6) is not satisfied, the time warps tend to oscillate.

A variety of time-warping functions can satisfy the discussed boundary conditions. Good results were obtained with the following family of time-warping functions:

$$\zeta(t) = A + B\exp(-\frac{(t-t_j)}{\sigma_B}) + C(t-t_j)\exp(-\frac{(t-t_j)}{\sigma_C}), \quad t_j < t \le t_{j+1}, \tag{7}$$

where A, B, C, σ_B, and σ_C are constants. Note that the time-warping function converges towards A with increasing t. At t_j the value of the time-warping function is just $A+B$. To ensure continuity, this is set to the value of $\zeta(t)$ at the endpoint of the previous subframe. The value of C can be used to satisfy equation (6) exactly. An entry of a codebook of continuous time warps can be generated by 1) choosing a value for A, 2) using B to satisfy the boundary condition at t_j, and 3) choose C to satisfy the boundary condition of equation (6) at t_{j+1}.

For small values of σ_B and σ_C the time-warping function of equation (7) converges to the case where $\zeta(t)$ is discontinuous. However, the mapping from original to warped residual signal remains a 1-to-1 mapping. We found that small discontinuities in $\zeta(t)$ do not affect performance.

A complication of the present method is that the time warping results in asynchrony between the original and the reconstructed speech signal. In nonreal-time applications, this is of no consequence, but in real-time applications an approximate synchrony is desired. This means that the integral of the time-warping function (from some initial time t_0) should be small compared to some time interval L:

$$\int_{t_0}^{t} \zeta(t)dt - (t-t_0) \ll L. \tag{8}$$

For each interpolation interval, a bias can be added to the a-priori delay contour, $d(\tau)$, to counter the asynchrony measured at the beginning of this interval.

2.3 Pitch Doubling and Halving

Although the present method attempts to maintain a continuous delay contour, doubling, or halving of the delay is difficult to prevent in practical applications. However, these cases can be accommodated as follows. As a first step, the open-loop delay estimate for the endpoint in the present interpolation interval is compared with the last delay in the previous interpolation interval. If the former is close to a multiple or submultiple of the latter, then delay multiplication or division is assumed to have occurred.

As an example, we consider the case of delay doubling in an interpolation interval $(\tau_A, \tau_B]$. Let the open-loop estimate of the end value delay be denoted as $d_2(\tau_B)$, where the subscript 2 indicates that the delay corresponds to two pitch cycles. For clarity we also write $d_1(\tau_A)$ for a delay corresponding to one pitch cycle. In general, the two-cycle delay and the single-cycle delay are related by:

$$d_2(\tau) = d_1(\tau) + d_1(\tau - d_1(\tau)). \tag{9}$$

Equation (9) describes two sequential mappings by the PP. Note that, in general, a simple multiplication of the delay by two does not result in a correct mapping.

Now consider the case where $d_1(\tau)$ is linear within the present interpolation interval:

$$d_1(\tau) = d_1(\tau_A) + \beta(\tau - \tau_A). \tag{10}$$

Then combination of equations (9) and (10) gives:

$$d_2(\tau) = (2-\beta)\, d_1(\tau_A) + (2-\beta)\beta\, (\tau - \tau_A), \quad \tau - d_1(\tau) > \tau_A. \tag{11}$$

Equation (11) shows, that, within a restricted range, $d_2(\tau)$ is linear in τ. However, in general, $d_2(\tau)$ is not linear in the range where $\tau_A < \tau < \tau_A + d_1(\tau)$. The following procedure can be used for delay doubling. At the outset $d_1(\tau_A)$ and $d_2(\tau_B)$ are known. By using $\tau = \tau_B$ in equation (11), β can be obtained. Then both $d_1(\tau)$ and $d_2(\tau)$ are known within the interpolation interval. The single-cycle delay, $d_1(\tau)$, satisfies equation (10) within the entire interpolation interval. For

$d_2(\tau)$, note that equation (9) is valid over the entire interpolation interval, while equation (11) is valid over only a restricted part. The PP contribution to the LP-excitation signal within the interpolation interval is obtained by a smooth transition from the single-cycle to the two-cycle delay:

$$u(\tau) = \psi(\tau)\, v(\tau - d_2(\tau)) + (1-\psi(\tau))\, v(\tau - d_1(\tau)), \quad \tau_A < \tau \le \tau_B \qquad (12)$$

where $\psi(\tau)$ is a smooth function increasing from 0 to 1 over the interpolation interval. The entire procedure assumed that the interpolation interval is larger than the two-cycle delay. For delay halving, the same procedure is used in the opposite direction.

2.4 Interpolation of the Pitch-Predictor Gain

Direct interpolation of the gain of the pitch predictor reduces its performance. However, a similar generalization of the analysis-by-synthesis procedure as that used for the delay can be used for the gain. Thus, the pitch predictor and fixed-codebook contributions need only match the shape of the original speech signal, but not its exact energy contour. We allow the original signal to be scaled up or down in amplitude, to maximize the fit with the reconstructed signal. This additional degree of freedom serves the same purpose for the PP gain as the time warping for the PP delay. Of course, the range of amplitude scaling should be limited such that no significant audible distortion occurs. In this way, the accuracy of the magnitude of the signal can be decreased in exchange for increased accuracy of the waveform shape.

Direct linear interpolation of the PP gain is not advisable as is illustrated by the simple case where the PP gain is constant in an interpolation interval, and has value υ. Furthermore, we assume that the contribution of the fixed-codebook to the overall signal energy can be neglected. Then, if there are P pitch cycles within the interval, the energy of the signal will change by a factor of υ^P. From this simple example, it is seen that, especially for speakers with a short pitch period, small errors in the estimation and time-dependence of the PP gain result in large errors in the energy contour of the reconstructed signal.

To prevent this sensitivity to mismatches between the optimal and interpolated gain contour, it is better to interpolate a less sensitive parameter which can be converted into the PP gain for each subframe. For this purpose, the rms energy of the PP contribution to the excitation signal can be used. (The rms energy must be computed in an approximately pitch-synchronous fashion.) Its values or their logarithms can be interpolated linearly to obtain the rms energy of the PP contribution to the excitation signal for each subframe.

3. PRACTICAL IMPLEMENTATION AND RESULTS

The coding techniques presented in the previous subsections were aimed at improving the performance of the pitch predictor during voiced segments of the speech signal. In comparison with the conventional PP, the new techniques have a bias towards generating periodicity. The resulting impairments can be reduced by using two different modes for the coder, one aimed at voiced speech and one aimed at unvoiced speech. During the unvoiced speech segments, the pitch predictor is not used.

The pitch gain and delay interpolation procedure was incorporated in a two-mode CELP coder. The decision between the voicing modes was made using a threshold on the maximum of the normalized autocorrelation function. The speech produced by this coder was compared to speech produced by a two-mode CELP coder that did not use interpolation of the pitch predictor parameters. From informal listening tests it was found that the coder with interpolation and an update rate of 20 ms provides similar quality as the coder without interpolation and an update rate of 5 ms.

4. CONCLUSION

The interpolation of the pitch-predictor delay and gain were improved by a generalization of the analysis-by-synthesis procedure. Conventional speech coders using the analysis-by-synthesis principle attempt to match the speech waveform of the original signal. In our new paradigm, the coding algorithm is allowed to select one signal from a set of modified original signals. Each of these modified signals represents excellent speech quality. By selecting from these modified signals the one for which the speech-coding algorithm performs best, a significant increase in coding efficiency is obtained.

We described a generalized analysis-by-synthesis procedure which uses time-warping for a pitch predictor using continuous delay contours. When used without interpolation, such a pitch predictor has the tendency to produce oscillating delay contours, and provides only a small speech-quality improvement over the conventional, stepped delay contour. Oscillating delay contours are not a problem when the delay is interpolated. In fact, the continuous mapping of the past signal into the present is an advantage when interpolation is used, because it ensures that pitch pulses are not lost or repeated near subframe boundaries.

In the generalized analysis-by-synthesis technique described in the present paper, the original signal is modified by continuous time warping. The modified signal is accurately described by the interpolated pitch-predictor parameters. To reduce the computational complexity with a minimal loss of speech quality, a blockwise time shift of the LP-residual signal can also be used. This procedure does not maintain the 1-to-1 mapping between the original and the modified signal. Using such "discrete" time warping, we have implemented a CELP coder which requires similar computational effort as existing fast procedures for a conventional CELP coder, but with a significant increase in coding efficiency.

References

[1] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J. Acoust. Soc. Am.* **33** pp. 1725-1736 (1961).

[2] B. S. Atal and M. R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates," *Proc. Int. Conf. Comm., Amsterdam*, pp. 1610-1613 (1984).

[3] P. Kroon and Ed. F. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates between 4.8 and 16 kbit/s.," *IEEE J. Selected Areas Comm.* **6** pp. 353-363 (1988).

[4] S. Singhal and B. S. Atal, "Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates," *Proc. Int. Conf. Acoust. Speech and Sign. Process., San Diego*, pp. 1.3.1-1.3.4 (1984).

[5] W. B. Kleijn and W. Granzow, "Methods for Waveform Interpolation in Speech Coding," *Digital Signal Processing* **1**(4) pp. 215-230 (1991).