WAVEFORM INTERPOLATION CODING WITH PITCH-SPACED SUBBANDS

W. Bastiaan Kleijn

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
100 44 Stockholm, Sweden

Huimin Yang

Institute of Microelectronics
Tsinghua University
Beijing 10084, China

Ed F. Deprettere

Department of Electrical Engineering
Delft University of Technology
2628 CD Delft, The Netherlands

ABSTRACT

We present new waveform-interpolation coding procedures which allow perfect reconstruction of the speech signal from the unquantized parameter set. Instead of using adaptive parameter extraction methods, we combine a time warping of the original signal with nonadaptive parameter extraction methods. The new coding structure has good performance at low bit rates and provides convergence to the original waveform with increasing rate.

1. INTRODUCTION

Speech coding algorithms are often classified into waveform coders and parametric coders. *Waveform coders* are characterized by the fact that the mean squared difference between the original signal and decoded signal can be made arbitrarily small by decreasing the quantization error, i.e. by increasing the bit rate. The term *parametric coders* traditionally referred to coders which use a more sophisticated signal model with parameters which are motivated by the human speech production mechanism. We define *strictly parametric coders* as coders which do not approach the original signal in a mean squared sense with increasing bit rate. This implies that the speech quality at infinite bit rate is determined by the accuracy of the model.

Despite the constraints in quality due to model inadequacies, strictly parametric coders have been popular at low rates, where decoded speech quality is low in general. The advantage of parametric coders is usually that most of the perceptually relevant information is concentrated in a low number of parameter signals which have relatively low bandwidth and which are relatively independent. However, the constraint of strictly parametric coding is not always inherent to the speech model. For example, linear-prediction based analysis-by-synthesis techniques (exemplified by CELP [1]) benefit from modeling traditionally associated with parametric coding *and* from waveform-coder properties. The parametric modeling used in these coders leads to high coding efficiency at low bit rates, while their waveform coding character means that the original signal can be approached arbitrarily closely by increasing the bit rate.

Whereas the benefits of decreasing waveform error with increasing bit rate have been exploited in linear predictive coding, this has been done neither in sinusoidal coding (e.g., [2]) nor in waveform interpolation coding (e.g., [3]) techniques. In the present paper, we demonstrate how the waveform-interpolation procedure can be modified to include the desirable property of being a waveform coder. From another viewpoint, the new procedure can be seen as a modification of the sinusoidal coding method. The modified waveform-interpolation technique presented here has the potential to provide good coding performance at both low and high rates. Based on earlier results with the waveform-interpolation coder, it is likely that the present paradigm allows coding at state-of-the-art speech quality for rates above about 1.5 kb/s.

2. PARAMETERS AND ESTIMATION

Voicing implies that the signal is nearly periodic. This has several clear advantages in human communication: high efficiency (loudness versus physical effort) and good perception in background noise. It also should allow for efficient coding since periodicity implies redundancy. The coding of unvoiced sounds, on the other hand, is facilitated by the fact that it can be modeled in a perceptually accurate manner as colored noise [4]. Based on this knowledge we can postulate several properties which are desirable for the parameter set (usually sampled at 40 to 200 Hz) associated with a speech model to be used in coding:

1. Periodic signals result in time-invariant parameter signals.
2. Stationary colored noise results in two classes of parameter signals: time-invariant signals and band-limited noise signals with a flat pass-band.
3. Small changes in slowly varying parameter signals result in small perceptual changes.
4. Replacement of noise-like parameter signals by synthetic noise signals results in small perceptual changes.
5. Unquantized parameter signals give perfect reconstruction.

This list is not exhaustive, but it motivates the choices made in this paper. The first four properties are particularly relevant for the coding of stationary signals. The last property is equivalent to membership of the waveform coder class.

Since an efficient representation of voiced speech is difficult to achieve, we will focus on the model structure it requires. Particularly the time dependency of the pitch and the time and frequency dependency of the periodicity level complicate the design of efficient speech coding algorithms. To deal with these factors, *adaptive parameter estimation techniques* such as adaptive window length and peak picking [2] or adaptive window length and placement and circular pitch-cycle alignment [3] have been used. In general, the adaptive procedures used for parameter estimation in low rate coders are not amenable to perfect reconstruction.

The method proposed in this paper can be interpreted as a *modification of the input signal* in combination with *nonadaptive parameter extraction*. The signal is modified so that nonadaptive analysis becomes effective. This concept is similar to generalized analysis-by-synthesis, where the original signal is modified to facilitate analysis-by-synthesis coding [5].

The adaptation of the input signal consists of time warping the signal such that the pitch period is constant. This has significant advantages. A fixed fundamental frequency facilitates the design of subband coders with bands lined up with the speech harmonics, providing slowly evolving parameters for voiced speech. Furthermore, the constant pitch track removes confusion between noise and the effects of time-varying pitch. Without time warping, conventional analysis methods often lead to a misinterpretation of signal components as noise-like for signals with time-varying pitch. This effect is strongest at high frequencies.

2.1. Pitch Normalization

We perform the time warping on the linear prediction residual of speech, consistent with independent evolution of the vocal tract and its excitation signal (particularly the pitch). While the equivalence of linear prediction residual and vocal-tract excitation is is strictly speaking incorrect, it appears to suffice for our purpose. Moreover, this approach allows the usage of established techniques for the encoding of the prediction filter specification.

Good adaptation of the signal, i.e. good time warping is crucial for good performance of our paradigm. A first step towards this goal is a reliable pitch estimation procedure (we use the method described in [6], updated on 20 ms intervals). It is beneficial to refine this track for analysis purposes. However, the synthesis pitch does not require such refinement and can, in general, be transmitted at a 50 Hz rate. An accurate time warp from time t to a modified time scale τ of constant pitch period P must satisfy

$$\tau(t) = \tau(t - p(t)) + P, \qquad (1)$$

where $p(t)$ is the maximum average cross-correlation $s(t)s(t - p(t))$ over a neighborhood of t. Equation 1 does not specify the warping *within* the pitch-cycle waveform. The recursive application of equation 1 may result in undesirable warping within the pitch cycle because of numerical inaccuracies and/or inaccurate initialization. Thus, we used a discrete version of the nonoptimal warping $\tau'(t - p(t)/2)) = P/p(t)$ with $\tau'(t) = \frac{d\tau(t)}{dt}$.

We use an approximation of band-limited interpolation to obtain a regular sampling on the τ scale and the same procedure to invert this process. These mappings are inconsistent but in practice performance is good. Using a 12-coefficient approximation of the sinc function and $P = 128$, we obtained a segmental signal-to-noise ratio of 60 dB for our mappings from and to τ over a speech data base 13.9 s in length (6.8 s female and 7.1 s male).

2.2. Representation of the Signal

Pitch normalization of the signal is relevant for voiced speech, which is nearly periodic (time-varying spectral content and pitch). A natural representation for the pitch normalized signal is

$$s(\tau) = \sum_{k=0}^{k=P-1} a_k(\tau)\exp(\frac{j2\pi k\tau}{P}). \qquad (2)$$

As is illustrated in figure 1, it is convenient to consider each of the terms of this summation to be the output of a filter $h_k(\tau)$ of a filterbank. $a_k(\tau)$ is then simply the demodulated band k signal:

$$a_k(\tau) = \exp(\frac{-j2\pi k\tau}{P})\int_{-\infty}^{\infty} s(\tau-\tau')h_k(\tau')d\tau'. \qquad (3)$$

This representation is good if the $h_k(\tau)$ are ideal bandpass filters. The $a_k(\tau)$ are constant for periodic $s(\tau)$ and nearly flat lowpass signals for colored stationary noise. Furthermore, the perceptual sensitivity of the modulation amplitudes $a_k(\tau)$ to an incorrect pitch normalization is low; the low frequency modulations (k small), to which the human ear is most sensitive, will be least affected, and thus still easy to code. This contrasts with sampled time domain (whether t or τ) representations where the inefficient quantization resulting from inaccurate pitch normalization affects the full frequency band.

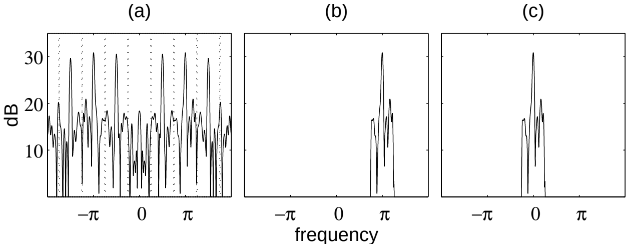

Figure 1: (a) Pitch-synchronous power spectrum.
(b) Power spectrum of one term.
(c) Demodulated power spectrum.

If the analog filters are not ideal, a periodic signal of unity period will result in spectral peaks at a frequency spacing of $2\pi/P$, which is undesirable for coding. A white noise input signal will result in a power spectrum of $a_k(\tau)$ with its shape determined by the filter transfer function.

Next, we consider the case of sampled signals $s(n)$ and digital filters $h_k(n)$ ($n \in \mathcal{Z}$ are samples of τ). If the filters are ideal bandpass filters, the situation is identical to that in the case of analog filters. However, if the filters are not ideal, the sampling rate affects the nature of the signal representation. If the filterbank outputs are downsampled, a periodic input will result in aliasing of harmonics. However, for the specific case where the filter outputs are *critically sampled*, this aliasing is not detrimental: in this case the aliasing *folds all spectral peaks exactly onto the dc (zero) frequency*. The filter outputs, $a(mP)$, $m \in \mathcal{Z}$, are then constant for a periodic input, as desired.

The behavior of the output of the critically sampled filterbank has a simple time-domain interpretation. For a steady-state periodic input, each bandpass filter has an output which is either periodic or constant. By sampling the output signal pitch-synchronously, we are guaranteed to obtain a constant output.

Thus, the frequency resolution of the subbands of the filterbank is not important when the pitch-period contour of the signal is normalized accurately and the filterbank is critically sampled. If the pitch normalization is not perfect, it is advantageous to have a filterbank with good suppression of the aliased components even for critical sampling. In summary, it is desirable both to have high-frequency resolution filters and critical sampling.

Whereas the sampled τ domain is redundant in terms of samples, the combined set of modulation amplitudes may not be so. If the synthesis filterbank is constructed with expansion vectors with finite time support, then it is seen that *the number of significant modulation amplitudes is approximately $p(t)$ at $\tau = \tau(t)$* (where $p(t)$ is measured in samples at the original sampling rate and “significant” means not approximately zero). The total number of significant samples in all bands of a critically sampled FIR filter bank operating on τ over an interval approximates the number of samples of the original signal over that interval.

2.3. The Pitch-Synchronous Spectrum

It is useful to consider the relation between the spectra of the Fourier-series coefficients and the time-warped signal spectrum,

$$
\begin{aligned}
S(\omega) &= \int_{-\infty}^{\infty} s(t(\tau)) \exp(-j\omega\tau) d\tau \\
&= \sum_{n=0}^{n=N-1} A_n(\omega - 2\pi n/P), \qquad (4)
\end{aligned}
$$

where $A_k(\omega)$ is the Fourier transform of $a_k(\tau)$. Equation 4 shows a simple relation between the pitch-synchronous spectrum and the spectra of the evolving Fourier-series coefficients. It shows that, for high frequency resolution filters, zooming in on harmonic k of the pitch-synchronous spectrum essentially provides the spectrum of the corresponding modulation amplitude $a_k(\tau)$.

3. PRACTICAL METHODS

3.1. The Block DFT

From the previous section we conclude that we want to have a critically-sampled perfect-reconstruction filterbank with regular frequency spacing of the subbands. A simple filterbank satisfying these conditions is the block discrete Fourier transform (DFT):

$$
\begin{aligned}
a_k(mP) &= \sum_n s(n) w_r(n - mP) W^{-kn}, \qquad (5) \\
s(n) &= \frac{1}{P} \sum_m \sum_{k=0}^{P-1} a_k(mP) w_r(n - mP) W^{kn}, \ (6)
\end{aligned}
$$

where $w_r(n)$ is a length P rectangular window and $W = \mathrm{e}^{\mathrm{j}2\pi/\mathrm{P}}$.

Interpreting the filterbank as an expansion into basis vectors (discrete functions) $w_r(n)W^{kn}$, we note that these basis vectors have significant discontinuities at their support boundaries. As a result, the perceived reconstruction accuracy of the speech signal is sensitive to small changes in the parameter signals $a_k(mP)$. The low frequency resolution of the block DFT also results in perceptual sensitivity since this implies that pitch normalization errors result in strongly deleterious aliasing effects.

We note in passing that the original waveform interpolation coder [3] uses an analysis procedure which is similar to the block DFT, but which operates on the unwarped signal. Straight application of the method implies that mismatches in the analysis and synthesis pitch tracks result in discontinuities of the signal at the block boundaries. These problems are avoided in [3] by adaptive window placement, alignment, and triangular windowing during synthesis. This results in good performance at low rates but prevents perfect reconstruction.

3.2. The Gabor Transform

The critically sampled block DFT has significant disadvantages resulting from the rectangular windowing. We would like to define a critically sampled filterbank with a basis which consists of a smoothly windowed exponentials. The Balian-Low theorem theorem (e.g., [7]) shows that such a basis cannot have both good frequency resolution and good time resolution. Thus, we relax the critical-sampling requirement. As penalty we obtain redundant sampling and detrimental effects due to aliasing even when the pitch normalization is accurate. A tradeoff between aliasing and the oversampling rate exists.

Transforms into frame coefficients (we use the mathematical meaning of "frame", e.g., [7]), i.e. into coefficients of an overcomplete expansion, where the frame consists of windowed exponentials, are called Gabor transforms. The Gabor transform is

$$
\begin{aligned}
b_k(mN) &= \sum_n s(n) w(n - mN) W^{-kn}, \qquad (7) \\
s(n) &= \frac{1}{N} \sum_m \sum_{k=0}^{P-1} b_k(mN) g(n - mN) W^{kn}, \ (8)
\end{aligned}
$$

where $w(n)$ and $g(n)$ are the analysis and synthesis windows and P/N is the oversampling factor. (The definition includes the block DFT.) For the experimental results reported below, $g(n)$ is a Hamming window (this choice means the frame is not tight) of length $2P$ and $P/N = 2$. Thus, quantization of a $b_k(mN)$ results in a smooth speech distortion over an interval of $2P$. This smoothness is desirable, but can cause smearing at speech onsets. We compute the dual frame (inverse transform) using a method which renders it maximally similar to the original frame [8].

It is interesting to note that, for a given oversampled analysis filterbank, it is possible to determine an optimal synthesis filterbank which uses noise shaping to minimize the effect of the quantization error on the subband signals [9]. At least in principle, this should allow for improved performance.

3.3. The MLT

The Balian-Low theorem shows that the design of a critically sampled filterbank based on windowed exponentials is impossible with satisfactory time and frequency resolution. However, it is possible to define a critically sampled filterbank based on smoothly windowed cosines. We have considered the commonly used modulated lapped transform (MLT) of the following form (e.g., [7]):

$$
\begin{aligned}
c_k(mP) = & \sum_n s(n) w(n - mP) f_k(n - mP), \qquad (9) \\
s(n) &= \sum_m \sum_{k=0}^{P-1} c_k(mP) w(n - mP) f_k(n - mP), (10)
\end{aligned}
$$

where the window $w(n)$ is symmetric, has a support of two pitch cycles, and satisfies $w(n)^2 + w(P-n-1)^2 = 1, n = 0, \cdots, P-1$ (we used a half sine wave) and where

$$
f_k(n) = \sqrt{\frac{2}{P}} \cos(\frac{(2n - P + 1)(2k + 1)\pi}{4P}). \qquad (11)
$$

Ignoring the windowing effects, the coefficients $c_k(mP), k = 0, \cdots, P-1$ describe the signal component with even symmetry of pitch cycle m and the component with odd symmetry of pitch cycle $m + 1$ (or vice versa). As in our Gabor-transform implementation, quantization of the subband signals results in speech distortions spread over a two pitch-cycle interval.

Figure 2 shows the frequency domain interpretation of the MLT filterbank for an artificial signal. Each term in the summation of equation 9 corresponds to the sum of the outputs of two symmetrically located band-pass filters. Demodulation and addition of these terms gives the modulation amplitude of each cosine term.

While it generally performs well for coding purposes, the MLT has some disadvantages. First, while the orthogonal basis functions of equation 10 are smoothly windowed, they have odd symmetry in one half. This can have a similar effect to a discontinuity of the basis functions and result in discontinuities in the speech due to the quantization process. The second disadvantage of the MLT is that there is no notion of phase. This is disadvantageous since it is well-known that the human auditory system is relatively insensitive to the phase of the pitch cycle waveform. In the original waveform interpolation coder the phase of the rapidly evolving component of the waveform is replaced by a random phase without introducing significant distortion [3]. We have not found a computationally simple manner to implement phase randomization in the MLT representation, both since the subband samples provide information about successive pitch cycles and since there is no simple relation between these samples and the corresponding magnitude and phase spectra.

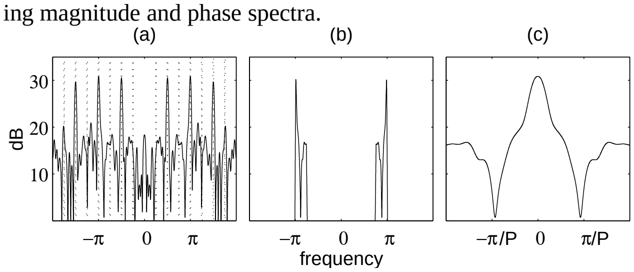

Figure 2: (a) Pitch-synchronous power spectrum.
(b) Power spectrum of odd term.
(c) Demodulated power spectrum of odd term.

3.4. Experiments

In our experiments, we evaluate the convenience of the representation for quantization. It is easier to create an effective quantizer when the subband signals are of low bandwidth and centered around dc. We define the rms mean bandwidth measure [10] for the subband signal $a_k(t)$ (note: we use the regular time scale) to be

$$B = \sqrt{(\int |A_k(\omega)|^2 \omega^2 d\omega)/(\int |A(\omega)|^2 d\omega)}. \qquad (12)$$

The bandwidths for each subband channel, averaged over 11 subbands in low frequency and high frequency ranges for 13.9 s of speech (6.8 s female, 7.1 s male), are shown in Table 1. For comparison, we note that the mean bandwidth of the characteristic waveforms of the conventional waveform interpolation coder [3] with adaptive analysis, was 42.9 Hz, for the 11 subband channels in the low frequency range.

Table 1: The mean bandwidth of the subbands.

	GT	MLT
B_{low}(Hz)	23	25
B_{high}(Hz)	86	41

Our results show that in terms of bandwidth the Gabor transform and MLT perform similarly. The frequency resolution of the Gabor transform is sufficient to suppress most of the undesired aliasing effects. We conclude that the MLT has as its main advantage critical sampling, whereas the Gabor transform represents the signal in terms of windowed exponentials facilitating a perceptually useful magnitude-and-phase interpretation.

4. CONCLUSION

In the period 1975-1985, subband filters were frequently used for speech coding (e.g., [11]). Research along these lines ultimately led to perfect reconstruction filterbanks [12]. However, while the sophistication of the filterbanks increased, subband coders were eclipsed by linear-prediction based coders in speech applications. This can largely be explained by the fact that straight application of filterbanks cannot exploit speech periodicity. In this paper we have shown that this disadvantage of subband coders can be eliminated by time warping the speech signal.

Application of a perfect reconstruction filter bank to the pitch normalized speech signal results in a representation suitable for coding at both high and low rates. The overlapping of the frame (or basis) vectors of the synthesis filterbank is equivalent to the interpolation of modulation amplitudes of earlier waveform-interpolation coders. However, in the present perfect-reconstruction schemes, the analysis has been changed significantly from that of the earlier waveform-interpolation coders.

5. REFERENCES

1. B. S. Atal and M. R. Schroeder, “Stochastic coding of speech at very low bit rates,” in *Proc. Int. Conf. Comm.*, (Amsterdam), pp. 1610–1613, 1984.
2. R. J. McAulay and T. F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 121–173, Amsterdam: Elsevier Science, 1995.
3. W. B. Kleijn and J. Haagen, “Waveform interpolation for speech coding and synthesis,” in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 175–208, Elsevier Science Publishers, 1995.
4. G. Kubin, B. S. Atal, and W. B. Kleijn, “Performance of noise excitation for unvoiced speech,” in *Proc. IEEE Workshop on Speech Coding for Telecomm.*, (Sainte-Adele, Quebec), pp. 35–36, 1993.
5. W. B. Kleijn, R. P. Ramachandran, and P. Kroon, “Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders,” *IEEE Trans. Speech and Audio Process.*, vol. 2, no. 1, pp. 42–54, 1994.
6. J. Haagen and W. B. Kleijn, “Waveform interpolation,” in *Modern Methods of Speech Processing* (R. P. Ramachandran and R. J. Mammone, eds.), pp. 75–99, Dordrecht, Holland: Kluwer Academic Publishers, 1995.
7. M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Signal Processing, Englewood Cliffs, NJ: Prentice Hall, 1995.
8. S. Qian and D. Chen, “Discrete Gabor Transform,” *IEEE Trans. Signal Process.*, vol. 41, no. 7, pp. 2429–2438, 1993.
9. H. Bölcskei and F. Hlawatsch, “Oversampled filter banks: optimal noise shaping, design freedom, and noise analysis,” in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, (Munich), pp. 2453–2456, 1997.
10. R. Bracewell, *The Fourier Transform and Its Applications*. New York: McGraw Hill, 1986.
11. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, “Digital coding speech in sub-bands,” *Bell Syst. Techn. J.*, vol. 55, no. 8, pp. 1069–1085, 1976.
12. M. Smith and T. P. Barnwell III, “Exact reconstruction for tree-structured subband coders,” *IEEE Trans. Acoust Speech Signal Process.*, vol. 34, no. 3, pp. 431–441, 1986.