A SPEECH CODER BASED ON DECOMPOSITION OF CHARACTERISTIC WAVEFORMS

W. Bastiaan Kleijn and Jesper Haagen [1]

Information Principles Research Laboratory
AT&T Bell Laboratories, Murray Hill, NJ 07974, USA.

ABSTRACT

For low-rate speech coding it is advantageous to represent the speech signal as an evolving characteristic waveform (CW). The CW evolves slowly when the speech signal is clearly voiced and rapidly when the speech signal is clearly unvoiced. The voiced (periodic) and unvoiced (nonperiodic) components of the speech signal can be separated by a simple nonadaptive filter in the CW domain. Because of perceptual effects, a significant increase in coding efficiency is obtained by coding these two components separately. A 2.4 kb/s coder using these principles was developed. In an independent evaluation, the performance of the 2.4 kb/s WI coder was found to be at least equivalent to the 4.8 kb/s FS1016 standard for all of the many tests.

1. INTRODUCTION

Voiced speech can be interpreted as a sequence of pitch-cycle waveforms. The general shape of these pitch-cycle waveforms evolves slowly as a function of time. Slow evolution facilitates prediction and interpolation, which are commonly used in coding to remove the redundancy in transmitted information. At low bit rates, prediction or interpolation over distances of one pitch period or more has proven difficult for a one-dimensional speech signal, because of nonstationarity of both the pitch period and the level of periodicity.

Greater success was attained by first extracting the pitch-cycle waveforms, describing each extracted waveform with a vector, and then performing interpolation, prediction, and quantization on these vectors [1] [2] [3] [4] [5]. In this method, the one-dimensional voiced-speech signal is transformed into a two-dimensional signal prior to the removal of redundancy. The two-dimensional signal displays the pitch-cycle waveform shape along one axis (with finite support), and the evolution of this pitch-cycle waveform over time on the second axis (see Fig. 1). The two-dimensional signal will be written as $u(t, \phi)$, where t is time, and ϕ is the axis along which the waveform shape is displayed.

The discussion will now be generalized to include unvoiced (nonperiodic) signals. To avoid using the term pitch-cycle waveforms for such signals, the more general term *characteristic waveform* (CW) will be used. The pitch period used is arbitrary for nonperiodic signals. Exact reconstruction of the original signal from a sequence of CWs (each of length one pitch period) is possible if the CW is sampled (extracted) with a sampling rate of at least once per pitch period. In general, the reconstruction accuracy decreases when the CW sampling rate falls below this rate. Whereas good reconstruction of quasi-periodic signals is possible at CW sampling rates significantly lower than once per pitch period, this is not the case for nonperiodic signals. Thus, the sensitivity to CW update rate decreases with increasing periodicity of the original signal. Still, an increase in voiced speech quality is observed with increasing CW sampling rate [3] [5]. However, an increase in CW sampling rate comes at the expense of an increase in bit rate. This increase in bit rate is stronger for less periodic signals.

Recently [6], advantage was taken of the structure of the evolving CW to allow simultaneously a high CW sampling rate and a low bit rate. The evolving CW does not display periodicity, but instead the CW evolves slowly where the original signal is quasi-periodic and rapidly where it is nonperiodic. Low-pass and high-pass filtering of the function $u(t, \phi)$ in the t direction (using a cut-off of about 20 Hz) separates it into two two-dimensional functions, which represent the voiced and the unvoiced components of speech. Separation of the voiced and unvoiced components, prior to quantization, allows a significant increase in coding efficiency, because they are interpreted differently by the human auditory system.

To insure continuity of the reconstructed one-dimensional speech signal, interpolation of the CW signal, which is sampled at a lower rate than the speech signal, is required. Linear interpolation is used to upsample the CW signal to one CW per output speech sample. The one-dimensional output signal is then obtained by defining the phase as a function of time using the pitch track (see section 2). The CW interpolation process led to the name waveform interpolation (WI) for the coding procedure.

A 2.4 kb/s WI coder exploiting the fore-mentioned principles has been created. The high performance of this coder illustrates the effectiveness of the WI paradigm in a practical application. The next two sections describe the basic WI procedure. Sections 4 and 5 provide details of the coder and its performance.

2. TRANSFORM TO AND FROM THE CW

As a first step, the pitch period is determined on a regular basis, and linearly interpolated. To account for pitch doubling and halving, and to prevent chirps, the pitch period is changed discontinuously at the interpolation midpoint by an integer (or its inverse) factor when appropriate.

A CW which is extracted from the speech signal will be called a *prototype* waveform. The best trade-off between time and frequency resolution for coding is obtained by using a square window of length one pitch period. To facilitate this window choice, it is convenient to perform the extraction on the linear-prediction residual signal. The window location is not tightly constrained, but instead the location is determined so that the window boundaries are located in regions of relatively low power.

As mentioned before, the CW is of finite length in ϕ. It is convenient to consider this finite segment as one cycle of a periodic function of ϕ, and to normalize the pitch period of this periodic function to 2π. Because the original speech

[1] J. Haagen is now with Tele Danmark Research, Horsholm, Denmark.

0-7803-2431-5/95 $4.00 © 1995 IEEE

signal is bandlimited, the periodic function of ϕ can be described by a finite (time-dependent) Fourier series (FS):

$$u(t,\phi) = \sum_{k=1}^{k=K} a_k(t)cos(k\phi) + b_k(t)sin(k\phi). \quad (1)$$

(Although this FS description with normalized pitch period is convenient, a time-domain description without pitch-period normalization can also be used, e.g. [2]. For the procedures presented here, these two methods result in slight numerical differences, which are usually not perceptually significant.) Upon extraction, the prototypes are aligned. In the alignment procedure, the phase ϕ of the newly extracted prototype is offset so as to maximize the cross correlation of it and the previous prototype waveform. Figure 1 shows the aligned prototype waveforms and the linear-prediction residual from which they are obtained, for a fixed sampling interval of 2.5 ms. The box around a segment of the residual signal outlines the segment which corresponds to a prototype waveform, *prior* to alignment. The residual-domain CW can be converted into a speech-domain CW by means of circular convolution, which can be performed directly on the FS representation [7].

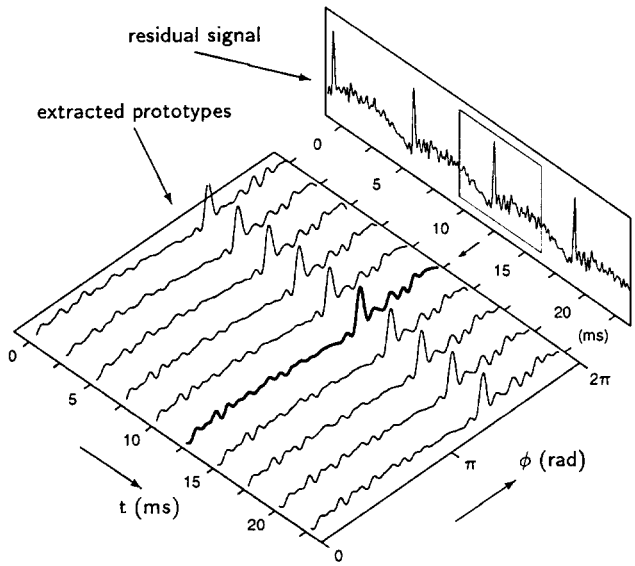

Figure 1. Extraction of the prototype waveforms.

The conversion from the two-dimensional to the one-dimensional domain can be performed in the speech domain or in the linear-prediction residual/excitation domain. In practice, only small numerical differences are found between the two methods. These differences are caused by the differences between circular and linear convolution. The principle of the transform from the two-dimensional CW domain back towards the one-dimensional domain is illustrated in Fig. 2 for the excitation-domain case. The contours in planes parallel to the ϕ and vertical axes represent the prototype waveforms. The CW is upsampled by linear interpolation between the prototype waveforms. To obtain a one-dimensional function from the evolving CW, the relation between the phase ϕ and time t must be specified. This relation is given in Fig. 2 by the diagonal curves, which have coordinates of the form $[t, \phi(t), u(t, \phi(t))]$. The slope $d\phi(t)/dt$ is the fundamental frequency. If $p(t)$ is the pitch period, the relation between ϕ and t is given by

$$\phi(t) = \int_{t_0}^{t} \frac{2\pi}{p(t')} dt'. \quad (2)$$

The time-domain signal is $e(t) = u(t, \phi(t))$.

The articulation rate of the reconstructed speech signal can be changed by changing the sampling interval between the prototype waveforms from that used during analysis. To prevent an increase in the level of periodicity (buzziness) for a decrease in the articulation rate, the decomposition described in the next section must be used. Small changes of the pitch track can be accommodated by changing $p(t)$ in equation 2. For large changes, this must be combined with band-limited interpolation of the discrete complex spectrum specified by the FS coefficients prior to pitch scaling.

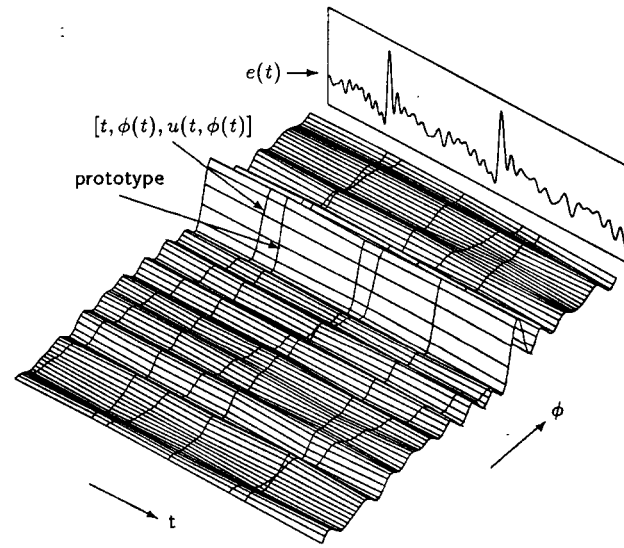

Figure 2. Construction of the excitation signal.

To be useful in coding, performing a forward transform followed by a backward transform should not result in significant perceptual distortion. If forward and backward transforms are not the exact inverse of each other, the performance of the transform must be evaluated by subjective testing. Thus, the transform to and from the CW was subjected to formal Mean Opinion Score (MOS) testing, with results shown in Table 1. The deteriorating effect of undersampling of the CW is clearly shown. Signals which are undersampled generally sound "buzzy" for regions containing a significant nonperiodic component. In formal comparison tests, it was found that, at a 400 Hz update rate, the speech quality is near that of the original signal and significantly better than the ITU G.726 (32 kb/s ADPCM) standard.

Table 1. MOS as a function of the update rate in WI.

CW Sampling Rate (Hz)	50	100	200	400
Mean Opinion Score	2.3	2.8	3.6	4.0

3. DECOMPOSITION OF THE CW

Since the nonperiodic (noise) component of a speech signal contains little redundancy, it requires a high bit rate for accurate transmission. However, unvoiced speech can be replaced by a signal with roughly identical magnitude spectrum and an almost similar signal-power contour without a decrease in the perceived naturalness of the speech signal [8]. Conversely, the human-auditory system is very sensitive to small changes in the spectrum of the quasi-periodic component of the speech signal. An example of such sensitivity is the perception of reverberation. These differences in perception suggest separate coding of these components would be beneficial.

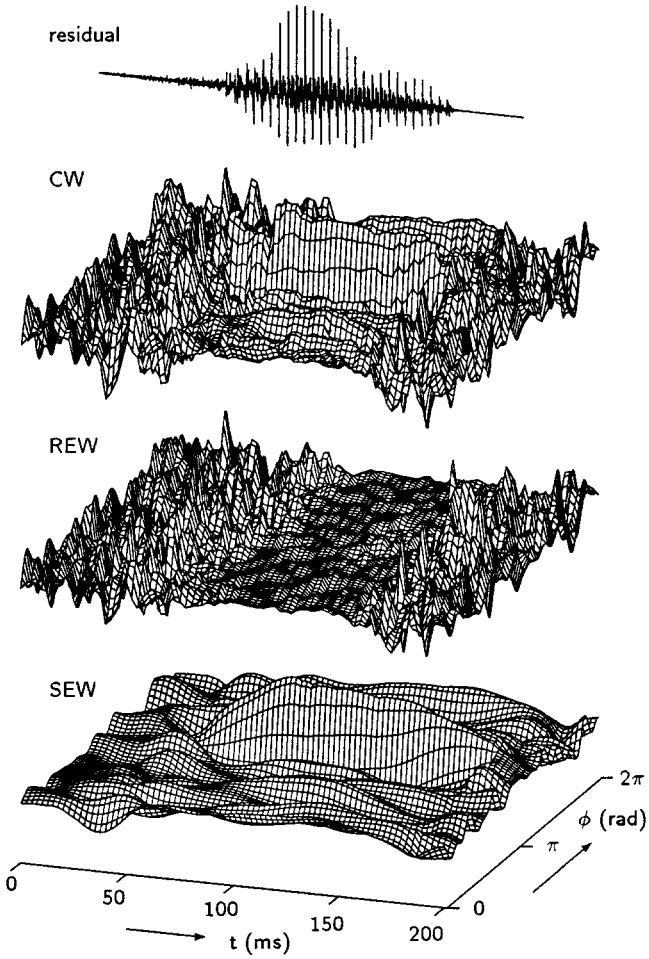

Figure 3. Decomposition of the residual signal.

When the original signal is quasi-periodic, the CW evolves slowly. In other words, the correlation between CWs decays slowly with increasing separation. For a nonperiodic signal, the correlation between CWs vanishes when their separation is one pitch period or more (at shorter separations they overlap). As a result, simple linear filtering can be used to separate the quasi-periodic and nonperiodic components of speech. By high-pass filtering $u(t,\phi)$ (i.e. the FS coefficients $a_k(t)$ and $b_k(t)$) in the t direction, a rapidly-evolving waveform (REW) is obtained. This waveform represents the nonperiodic component of speech and dominates during unvoiced speech. Similarly, low-pass filtering leads to a slowly-evolving waveform (SEW), which represents the quasi-periodic (voiced) component of the speech signal.

Figure 3 illustrates the decomposition for the word "help" spoken by a female. At the top the linear prediction residual signal is shown and below that the evolving CW, normalized to unit power. Low-pass and high-pass filtering with a cut-off rate of 20 Hz results in the SEW and REW which are shown below the CW. Note that the SEW can be downsampled to 40 Hz (assuming an ideal low-pass filter).

The decomposition of the evolving CW into a SEW and a REW allows efficient coding of the CW despite a high CW sampling rate. To obtain the same perceptual quality by direct quantization of the CW itself would require a significantly higher bit rate. The SEW can be described accurately at a low bit rate because it changes slowly as a function of time, allowing downsampling and/or differential quantization. The REW can be described accurately at a low bit rate because perceptually accurate reconstruction requires only a rough description of the REW magnitude spectrum. Earlier experiments [9] suggest that it will be useful to specify the power contour of the REW within a CW (modulate the power along the ϕ axis). Informal trials indicate that such fine-resolution of the REW power contour leads to a small improvement in quality. However, in the implementations discussed below, the REW phase is random.

4. CODER IMPLEMENTATION

Two principles were used in the design of a WI coder implementation. First, the parameters to be quantized were selected so as to be maximally independent. For example, the pitch period and the energy of the speech signal are independent parameters. Using such "orthogonality" between model parameters allows the quantizers to be designed independently, and it minimizes the effect of channel errors. Second, each transmitted index is to affect a minimal number (preferably only one) of the model parameters. In other words, a classification which selects the speech model or the quantizers is not used. The avoidance of classification facilitates robustness to background noise (particularly previously unknown background noise), and increases the robustness to channel errors.

Table 2. Bit allocation for the 2.4 kb/s WI coder.

parameter	update rate (Hz)	bit allocation
LP coefficients	40	30
pitch period	40	7
signal power	80	2
REW magnitude	240	3 or 1
REW phase	480	0
SEW	40	7

In the 2.4 kb/s WI coder, the prototype waveform is extracted from the original signal at a rate of 480 Hz and represented by FS coefficients. At the receiver, these prototype waveforms are reconstructed from quantized parameters transmitted at a number of different rates. These parameters, their update rates, and their bit allocation are summarized in Table 2.

At the analyzer, the signal power is computed for each prototype waveform, transformed to the logarithm of the speech-domain power, low-pass filtered, and then downsampled to a rate of 80 Hz. This sampled signal is quantized using a differential quantizer. At the receiver, linear or stepwise interpolation is used for upsampling of the power to 480 Hz. The linear-prediction coefficients are transmitted at a rate of 40 Hz and quantized using split-VQ of line-spectral frequencies [10], designed with a method which maximizes robustness to channel errors [11]. At the receiver, the line-spectral frequencies are upsampled to 480 Hz by linear interpolation.

The prototype waveform power is normalized to have unity magnitude spectrum before quantization of the REW and the SEW. Separate gain quantizers are not used for the REW and SEW. At the receiver, each REW is the result of combining the quantized magnitude spectrum with a new random phase spectrum. The magnitude spectrum is transmitted at 240 Hz. The quantization index for the shape of the magnitude-spectrum (3 bits) is transmitted at 120 Hz, and the intermediate updates specify the selection of the previous or the next transmitted shape (1 bit).

For the 2.4 kb/s WI coder, the low-frequency band of the magnitude spectrum of the SEW is vector quantized, and transmitted at a rate of 40 Hz. For the remaining frequency band, the overall prototype-waveform magnitude spectrum

Table 3. MOS, DMOS, DAM, and DRT test results for 2.4 kb/s WI coder and 4.8 kb/s FS1016.

test type	MOS	MOS	MOS	DMOS	DMOS	DMOS	DAM	DAM	DAM	DAM	DRT	DRT
condition	quiet	office	MCE	car	Humvee	M2	quiet	quiet	Humvee	2%	quiet	2%
WI 2.4	3.77	2.93	2.73	3.74	3.50	3.11	66.8	68.9	38.3	42.7	92.0	87.2
FS1016	3.59	2.94	2.68	3.78	2.85	3.10	63.1	67.5	37.9	42.2	92.8	87.7

is approximated as exactly flat. Since the prototype waveforms are defined in the LP-residual domain, this is a reasonable approximation. Thus, for this remaining band, the SEW magnitude spectrum can be computed from the (flat) overall magnitude spectrum and the REW magnitude spectrum. One of four phase spectra is selected for the SEW on the basis of the REW spectrum transmitted.

The prototype waveforms are formed by adding the FS coefficients representing the REW and the SEW. A postfilter is used to shape the prototype waveforms [12]. Application of the postfilter to the prototypes is advantageous, because the signal power is set afterward.

The pitch period is transmitted at a rate of 40 Hz. At the receiver the pitch-period signal is upsampled to a rate of 8 kHz by interpolation as described in section 2, and a phase track is computed. An 8 kHz CW signal is obtained by linear interpolation of the prototype waveforms, and the phase-track is used to convert this two-dimensional signal to the one-dimensional linear-prediction excitation signal. FS terms which cross the Nyquist frequency are interpolated from or to zero amplitude. The LP contribution to the spectral shape is added either to the prototype waveforms, or to the one-dimensional signal by conventional filtering.

Using the unquantized SEW in the 2.4 kb/s coder indicates that more accurate quantization of the SEW results in significant performance gains. This may be used in an embedded coding scheme with the 2.4 kb/s system as the lowest rate. In general, it can be said that, at 2.4 kb/s, the reconstructed speech quality is limited by the quantizers, and not by the model.

5. RESULTS

The 2.4 kb/s WI coder was part of a survey of 2.4 kb/s coders, which was organized by the U.S. Department of Defense in the third quarter of 1994 [13]. Eight different coders from various organizations participated in this test. The WI coder obtained the highest score for half of the 22 test conditions. Six of these wins were statistically significant. In 21 of the 22 conditions the WI score was either the highest score or not significantly different from the highest score. The remaining condition was a DRT with very high background noise level.

The survey also contained the 4.8 kb/s FS1016 standard [14] as a reference. When 2.4 kb/s WI is compared to FS1016, WI scored highest for the majority of conditions. WI outscored FS1016 four times by a statistically significant margin, whereas FS1016 was never better than WI by such a margin. Table 3 shows some representative test results for the 2.4 kb/s WI coder and FS1016. Most test labels are self-explanatory; MCE indicates a mobile command environment, Humvee and M2 indicate the background noise for an all-terrain and fighing vehicle, respectively, and 2% indicates 2% random channel errors. The two DAM tests for quiet background have a different set of speakers.

The 2.4 kb/s coder currently requires 60 ms look-ahead. Of this, 25 ms is used for the filtering operation required for the decomposition. Work is in progress to minimize the complexity of the 2.4 kb/s WI. Without shortcuts affecting the numerical output, the current complexity is about four times that of FS1016 [15].

6. CONCLUSION

The waveform interpolation (WI) method facilitates efficient coding of speech. It provides a speech quality which converges to that of the original signal with increasing bit rate. It can also be used for time-scaling and pitch-scaling of speech. The 2.4 kb/s WI coder implementation demonstrates the potential of the method by providing a speech quality at least equivalent to that of the 4.8 kb/s FS1016 standard under all of 22 tests. The WI method is expected to be effective for bit rates between 2 and 8 kb/s.

Acknowledgement

The authors thank P. Knagenhjelm for his robust vector quantizers and D. Sen for his work on complexity reduction.

REFERENCES

[1] W. B. Kleijn, "Continuous Representations in Linear Predictive Coding," Proc. Int. Conf. Acoust. Speech Sign. Process., pp. 201-204, IEEE, 1991.

[2] J. Haagen, H. Nielsen, and S. Duus Hansen, "Improvements in 2.4 kbps High-Quality Speech Coding," Proc. Int. Conf. Acoust. Speech Sign. Process., pp. II 145-148, IEEE, 1992.

[3] Y. Shoham, "High-Quality Speech Coding at 2.4 to 4.0 kbps based on Time-Frequency Interpolation", Proc. Int. Conf. Acoust. Speech Sign. Process., pp. II 167-170, IEEE, 1993.

[4] I. S. Burnett, and R. J. Holbeche, "A Mixed Prototype Waveform/CELP Coder for Sub 3 kb/s," Proc. Int. Conf. Acoust. Speech Sign. Process., pp. II 175-178, IEEE, 1993.

[5] Y. Tanaka and H. Kimura, "Low-Bit-Rate Speech Coding Using a Two-Dimensional Transform of Residual Signals and Waveform Interpolation," Proc. Int. Conf. Acoust. Speech Sign. Process., pp. I 173-176, IEEE, 1994.

[6] W. B. Kleijn and J. Haagen, "Transformation and Decomposition of the Speech Signal for Coding,", IEEE Signal Processing Letters, Vol. 1, September 1994, pp. 136-138.

[7] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms", IEEE Trans. Speech Audio Processing, Vol 1, No. 4, pp. 386-399, 1993.

[8] G. Kubin, B. S. Atal, and W. B. Kleijn, "Performance of Noise Excitation for Unvoiced Speech," Proc. IEEE Workshop on Speech Coding for Telecommunications, pp. 35-36, 1993.

[9] D. J. Hermes, "Synthesis of Breathy Vowels: Some Research Methods," Speech Communication, Vol. 10, pp. 497-502, 1991.

[10] K. K. Paliwal, and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," IEEE Trans. Speech Audio Process., Vol. 1, No. 1, pp. 3-14, 1993.

[11] P. Knagenhjelm, "How Good is your Index Assignment?," Proc. Int. Conf. Acoust. Speech Sign. Process., II 423-426, IEEE, 1993.

[12] J. H. Chen and A. Gersho, "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering," Proc. Int. Conf. Acoust. Speech Sign. Process., pp. 2185-2188, 1987.

[13] M. A. Kohler and L. M. Supplee, "Progress Towards a new Government Standard 2400 bps Voice Coder," Proc. Int. Conf. Acoust. Speech Sign. Process., IEEE, 1995.

[14] J. P. Campbell, V. C. Welch, and T. E. Tremain, "The DOD 4.8 kbps Standard (Proposed Federal Standard 1016)," In B. S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, pp. 121-133, Kluwer Academic Publishers, Dordrecht, Holland, 1991.

[15] D. Sen. Unpublished work. 1994