# Cycle Extraction for Perfect Reconstruction and Rate Scalability

*Miguel Arjona Ramírez*

Electronic Systems Eng. Dept. - Escola Politécnica
05508-900 University of São Paulo, São Paulo, SP, Brazil
miguel@lps.usp.br

## Abstract

A cycle extractor is presented to be used in a speech coder independently from the coding stage. It samples cycle waveforms (CyWs) of the original prediction residual signal at their natural nonuniform rate. It is shown that perfect reconstruction is possible due to the interplay of these properties for two cycle length normalization and denormalization techniques. The coding stage is coupled to the cycle extractor in the analysis stage by an evolving waveform interpolator that may handle several interpolation methods and sampling rates for a variety of fixed and variable rate coders. The description of extraction, evolution interpolation and synthesis stages is cast in discrete time. The upper performance bound is perfect reconstruction while the lower bound is equivalent to conventional waveform interpolation (WI) speech coding.

## 1. Introduction

Cycle waveforms are pitch cycles from voiced speech segments in a strict interpretation. Even though the notion of pitch cycles is a classical one in speech analysis, being more important for pitch-synchronous coders, the operation of cycle waveform (CyW) extraction has been used successfully even over unvoiced segments of the speech signal. In its application to speech coding, several interpolation techniques have to be used along. This necessity stems from the variable length of cycle waveforms and the importance of processing their shape information independently from their duration and amplitude.

Actually, waveform interpolation provides a flexible excitation signal model for speech coding, usually coupled to linear prediction coding of the spectral model [1]. However, its signal and parameter waveforms have different inherent bandwidths and critical rates that are not generally uniform. These rates include the cycle rate, the prediction (LP) rate, the pitch detection rate, the signal sampling rate and the waveform coding rate. Therefore, it becomes simpler and more appealing to present the model in a continuous time description as was originally done. Actually, continuous-time representations can be implemented by digital descriptions like cubic B-splines [2], but the management of different signal and parameter rates remains a hurdle to coder implementation. It is usually solved by imposing the signal sampling or parameter determination rate by design in compatibility with the coder transmission rate, which usually leads to modeling imperfection or coding inefficiency.

The standpoint presented here considers that signals and parameters should be extracted or determined at their natural rates and evolution interpolators should handle their delivery at the rates required by the coder. Conceptually, the highest sampling rate considered is the speech signal sampling rate, so that discrete-time representations are sufficient for processing and algorithmic description as well.

Therefore, the source side of the coder may be controlled by the source characteristics and the coding side may be matched to the transmission or network requirements, that may require a coder operating at a fixed rate or at variable rate. This approach supports a highly flexible rate scalability range. If quality can be traded for efficiency, the most straightforward rate scalable coder is an embedded coder, using a single encoding model [3].

This cycle extractor also suits pitch-synchronous coders [4] as long as pitch is redefined as the duration of the segment of the signal extending between two adjacent interpeak low-amplitude instants regardless of voicing.

## 2. Waveform interpolation description

In waveform interpolation [1], the surface $u\left(t,\phi(t)\right)$ characterizes the excitation in conjunction with the phase track

$$\phi(t)=\phi\left(t_0\right)+2\pi\int_{t_0}^{t}\frac{1}{p(t)}dt, \tag{1}$$

where $p(t)$ is the pitch track. The *characteristic waveform* (CW) $c_{t_0}\left(\phi\right)=u\left(t,\phi\right)|_{t=t_0}$ for $\phi\in[-\pi,\pi)$ describes the potential pitch cycle waveform at time $t=t_0$, which is only revealed by the sample

$$c_{t_0}\left(\phi\right)=u\left(t_0,\phi\left(t_0\right)\right)=r\left(t_0\right)$$

of the residual signal. Therefore, for the moment, we will require for perfect reconstruction that the characteristic waveform be a warped version of the segment of the residual signal extending from $t_0$ onwards up to the next interpeak midcycle instant $t_0+p\left(t_0\right)$, assuming that $t_0$ itself is an interpeak midcycle instant.

Viewing the characteristic waveform surface along the time axis is important for sampling and interpolation of CWs. For a normalized phase $\phi=\phi_0$, the corresponding *evolving waveform* (EW) is $e_{\phi_0}\left(t\right)=u\left(\phi,t\right)|_{\phi=\phi_0}$. A smoother characteristic waveform evolution may be obtained by interpolating the extracted waveforms after length normalization.

The standard waveform extraction procedure applies uniform sampling [5] but critical pitch cycle extraction has been used to lower coding complexity [6] as well as to enable perfect waveform reconstruction [7].

## 3. Cycle extraction

The waveform selected for processing is the linear prediction residual signal $r(n)$, which is usually chosen due to its enhanced periodic characteristics over the original speech signal. The periodicity of the residual signal is further analyzed by a robust pitch detector based on its autocorrelation function, which follows the guidelines for pitch detection set forth in [8] and [9].

A pitch period value estimate is delivered per pitch analysis interval even if the signal should be unvoiced over the interval so that a voicing detector is required along with the pitch detector. As shown in Fig. 1, an autocorrelation voicing detector is used, providing a decision $v(n_i)$ per interval as well.

The pitch period estimates ease the task of the waveform demarcator which looks for the endpoints of pitch cycles (see Fig. 3). For the sake of perfect reconstruction, the starting endpoint of cycle $n_c$ is the sample at time $n = d(n_c - 1) + 1$ that follows the end of the previous extracted cycle while the terminating endpoint $n = d(n_c)$ is placed at a low-amplitude position between the next two pitch peaks. The cycle demarcator in Fig. 3 searches for the next two peaks within an amplitude tolerance from the current peak for a length of time determined by the current pitch estimate within a set tolerance. Next, the cycle picker will search a region around the midpoint between the next two peaks for a low-amplitude sample. Some demarcations of cycle waveforms are shown in Fig. 2.

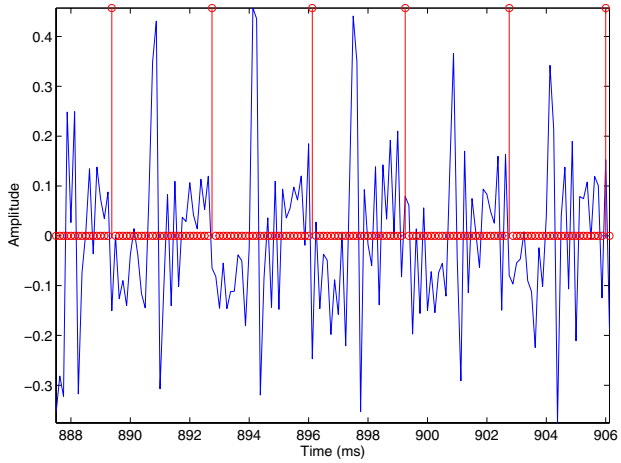


Figure 2: Cycle waveforms demarcated over a section of a residual signal.

As an important result of the waveform picking process for scalability, the pitch period $p_0(n_i)$ determined by the pitch detector for interval $n_i$ where the pitch cycle lies is replaced by the cycle length $p(n_c) = d(n_c) - d(n_c - 1)$.

## 4. Cycle length normalization

Extracted pitch cycle waveforms undergo a sampling rate expansion to a constant period or phase cycle. Considering the periodic nature of pitch cycles, a Fourier series was the original representation used to perform the transformation to a constant cycle length domain. Usually, the evolving Fourier-series coefficients

$$a_t(k) = \frac{1}{p(t)} \int_t^{t+p(t)} r(t) e^{-j\frac{2\pi k}{p(t)}t} dt \qquad (2)$$

are used for $k = -K, -K+1, \ldots, K$ with $K = \lfloor f_{\text{Ny}} p(t) \rfloor$, where $f_{\text{Ny}}$ is the signal's bandwidth or Nyquist frequency. These Fourier coefficients may be used in their raw form for analysis. However, a new normalized time scale $\phi$ is more efficient for coding because it avoids the birth and death of harmonic tracks. It is normally referred to as the phase axis and the CW along this axis becomes

$$c_t(\phi) = \sum_{k=-\frac{P}{2}}^{\frac{P}{2}} a_t(k) e^{j\frac{2\pi k}{P}\phi} \qquad (3)$$

where $P/f_s$ is the constant pitch period for signal sampling frequency $f_s$. This time warping delivers perfect reconstruction as long as the constant pitch period is not smaller than the longest pitch period. Additionally, in Eq. (3) the Fourier series has been extended by the terms with coefficients $a_t(k) = 0$ for $k = \pm(K+1), \pm(K+2), \ldots, \pm\frac{P}{2}$. Conversely, the original Fourier series may be obtained by truncation.

A discrete-time representation is more convenient here. Instead of the running signal $r(t)$ in Eq. (2), the extracted cycle waveform

$$c_{n_c}(m) = r(d(n_c - 1) + m + 1) \qquad (4)$$

is used for $m = 0, 1, \ldots, p(n_c) - 1$. Now the discrete Fourier series of the waveform cycle beginning at time $n_c$ is

$$a_{n_c}(k) = \frac{1}{p(n_c)} \sum_{m=0}^{p(n_c)-1} c_{n_c}(m) e^{-j\frac{2\pi k}{P}m} \qquad (5)$$

for $k = -K_l, -K_l + 1, \ldots, K_u$. For $p(n_c)$ even, $K_l = p(n_c)/2$ and $K_u = K_l - 1$. Otherwise, for $p(n_c)$ odd, $K_l = K_u = \frac{p(n_c)-1}{2}$. The CW is obtained by the extended Fourier series

$$c_{n_c}(\varphi) = \sum_{\varphi=-\frac{P}{2}}^{\frac{P}{2}-1} a'_{n_c}(k) e^{j\frac{2\pi k}{P}\varphi}, \qquad (6)$$

where constant pitch period $P$ is assumed to be even, without loss of generality, and the extended coefficients are

$$a'_{n_c}(k) = 0 \text{ for } K_l + 1 \le |k| \le \frac{P}{2} - 1 \text{ and } k = -\frac{P}{2}$$

along with

$$a'_{n_c}(-K_l) = a'_{n_c}(K_l) = \frac{1}{2} a_{n_c}(-K_l)$$

for $p_{n_c}$ even, or

$$a'_{n_c}(-K_l) = a_{n_c}(-K_l) \text{ and } a'_{n_c}(K_u) = a_{n_c}(K_u)$$

for $p_{n_c}$ odd. Conversely, the original Fourier series may be obtained by truncation and the inverse of the endpoint operation outlined above for the extended coefficients.

For efficient coding, band-limited interpolation with truncated sinc functions may be used instead to normalize the length of the extracted cycle waveforms [10].

## 5. Characteristic waveform composition

The sequence $\left\{\{c_{n_c}(\varphi)\}_{\varphi=0}^{P-1}\right\}_{n_c}$ of characteristic waveforms subsumes a characteristic surface when the waveforms are aligned and properly layed out along time axis $n$, provided that a high enough cycle sampling rate has been used. The cycle extraction process outlined in Section 3 guarantees a great degree of alignment between consecutive extracted cycles due to the placement of the peak in the middle region of $c_{n_c}(m)$. However, a residual misalignment still remains, caused by the variable

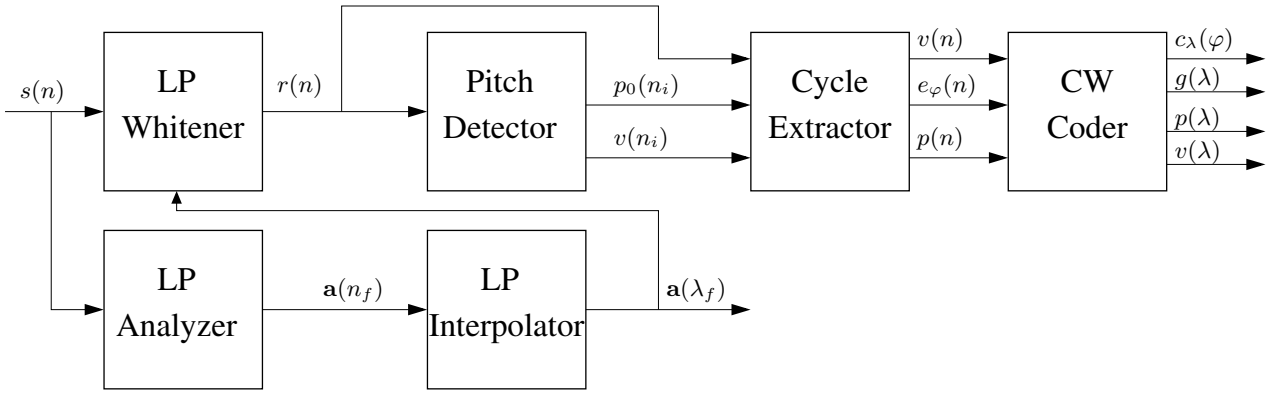


Figure 1: Block diagram of generic speech coder that uses the cycle extractor.

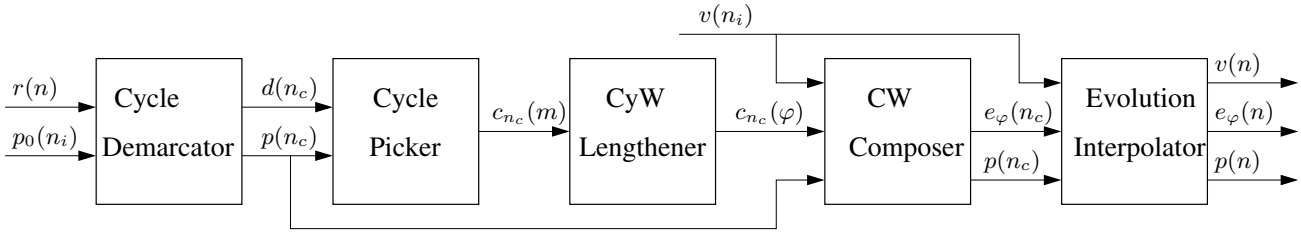


Figure 3: Block diagram of the cycle extractor.

pitch period, which the CW composer included in Fig. 3 corrects by means of cyclic shifting when the waveform happens to be voiced. As a consequence, the pitch track has to be adjusted for the alignment offset so that the synchrony may be recovered. Peak alignment has been found to be more effective than alignment by maximum autocorrelation in agreement with [8].

Concerning the placement of CWs along the time axis constrained by alignment, the best strategy for a first approximation is to hold the same CW for the duration of the corresponding cycle waveform as

$$c_n(\varphi) = c_{n_c}(\varphi) \tag{7}$$

for $\varphi = 0, 1, \ldots, P-1$ and $n = d(n_c - 1) + 1, d(n_c - 1) + 2, \ldots, d(n_c)$.

Further, various kinds of interpolation may be used to allow sampling the CWs at uniform rates. Applying band-limited sinc interpolation, as was done for cycle lengthening in [10], a smoother evolving surface may be obtained, which may be downsampled to the lower rates used for uniform sampling. Employing $2D+1$ original samples for interpolation, the warped signal is generated as

$$e_\varphi(\lambda) = Q \sum_{n=\frac{\lambda}{Q}-D}^{\frac{\lambda}{Q}+D} e_\varphi(n)\, h(\lambda - Qn), \tag{8}$$

for $\varphi = 0, 1, \ldots, P-1$, where $Q = f_s/f_{\mathrm{CW}}$ is the CW downsampling factor from the signal sampling rate $f_s$ to the final CW sampling rate $f_{\mathrm{CW}}$. The CW surface obtained for $f_{\mathrm{CW}} = 400$ Hz is illustrated in Fig. 4.

For band-limited sinc evolution interpolation, the evolving waveforms are upsampled for synthesis by

$$\tilde{e}_\varphi(n) = \sum_{\lambda=Q(n-D)}^{Q(n+D)} e_\varphi(\lambda) h(Qn - \lambda). \tag{9}$$

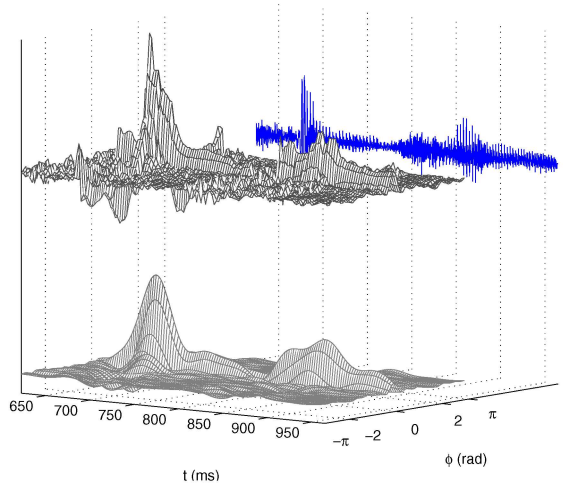


Figure 4: Filtered CW surface in the foreground for a section of the residual signal in the background. Below a smoothed slowly evolving surface is shown.

## 6. Cycle waveform synthesis

In the synthesizer, the evolving waveforms are upsampled as exemplified at the end of last section in order to obtain the reconstructed characteristic waveforms.

Besides, the decoded pitch track $\tilde{p}(\lambda)$ is upsampled to the signal sampling rate producing the interpolated track $\tilde{p}(n)$. The received voicing track $\tilde{v}(\lambda)$ is upsampled as well and both tracks are used to shrink the characteristic waveforms back to cycle waveforms as depicted in Fig. 5.

Further, the pitch track drives a cycle waveform sampler

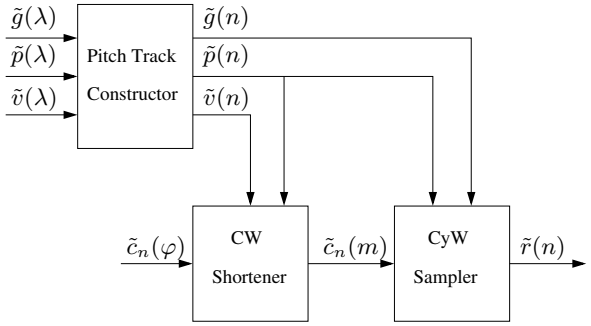


Figure 5: Block diagram of cycle waveform composer for signal synthesis.

through a derived phase track

$$\tilde{m}(n) = \left( \tilde{m}\,(0) + \sum_{i=1}^{n} 1 \right) \mod \tilde{p}(n) \qquad (10)$$

to regenerate the residual signal

$$\tilde{r}(n) = \tilde{g}(n)\tilde{c}_n\,(\tilde{m}(n)) \qquad (11)$$

after scaling the cycle samples by the interpolated gain $\tilde{g}(n)$.

For upsampling the extracted cycles, other types of interpolation may be used besides Fourier-series extension and windowed sinc interpolation [10], such as cubic B-spline interpolation [2].

## 7. Experiments

The cycle extractor has been tested at the natural cycle rate upsampled to the signal sampling rate $f_s = 8$ kHz as an intermediate characteristic waveform sampling rate and it has also been used to emulate the uniform CW sampling rate $f_{\mathrm{CW}} = 400\,\mathrm{Hz}$ established by [5]. Cycle waveforms are represented in the normalized phase domain where they are lenghthened to by Fourier-series extension and the corresponding CWs are shortened back by Fourier-series truncation. But band-limited sinc time-warping has been applied for comparison as well. In all cases, the accurate pitch track extracted has been used throughout. As test signals, eight sentences from the TIMIT speech database have been used, equally distributed between male and female speakers, for a total recording time of 14.5 s of female speech and 12.4 s of male speech.

Signal reconstruction performance has been evaluated at the residual signal level by measuring the segmental signal-to-noise ratio (SNRSEG) with 16 ms segments between the residual signal $r(n)$ in Fig. 1 and its reconstruction $\tilde{r}(n)$ by Eq. (11) and located in Fig. 5.

First of all, cycle waveform extraction at the natural cycle rate with Fourier-series stretching to length $P = 256$ along the phase axis attains virtually perfect reconstruction. Upsampling the evolving waveforms from the natural cycle rate to the signal sampling rate by zero-order hold interpolation maintains the perfect reconstruction situation. However, when sinc interpolation based on $2D + 1 = 11$ samples of the naturally extracted waveform is used instead for time warping, the SNRSEG drops to around 50 dB.

When the evolving waveforms are lowpass filtered and sampled at the rate $f_{\mathrm{CW}} = 400$ Hz by means of sinc interpolation based on $2D + 1 = 11$ samples of the evolving waveforms at the signal sampling rate as illustrated in Fig. 4, the average SNRSEG is about 30 dB, matching the performance of the conventional waveform extraction process [2].

Besides, preliminary experiments with multiple cycle extraction have shown an SNRSEG increase of more than 7 dB over conventional WI extraction. Multiple cycle representation has already been used for a pitch-synchronous transform representation [4].

## 8. Conclusion

The algorithmic description of cycle waveform extraction and interpolation has been cast in discrete time. This representation matches the processing steps of the analysis stage of a speech encoder. This stage incorporates a proposed cycle extractor which operates at the natural nonuniform cycle rate of the prediction residual signal. This tuning of the analysis stage to the signal features makes it possible to attain perfect reconstruction. The coding stage may operate at its intrinsic rates or at network transmission rates, being linked to the analysis stage by an evolution waveform interpolator. Several interpolators have been tried, providing results between perfect reconstruction and conventional WI extraction and representation.

## 9. References

[1] W. Bastiaan Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, W. Bastiaan Kleijn and K. K. Paliwal, Eds., pp. 175–207. Elsevier Science, Amsterdam, 1995.

[2] V. T. Ruoppila, M. Tammi, and J. Saarinen, "Waveform extraction for perfect reconstruction in WI coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, 2000, vol. 3, pp. 1359–1362.

[3] Hong-Goo Kang and D. Sen, "Embedded WI coding between 2.0 and 4.8 kbit/s," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, 1999, vol. 1, pp. 87–89.

[4] Huimin Yang, W. Bastiaan Kleijn, E. Deprettere, and Hongyi Chen, "Pitch synchronous modulated lapped transform of the linear prediction residual of speech," in *Proc. of IEEE Int. Conf. on Signal Processing*, Beijing, 1998, vol. 1, pp. 591–594.

[5] W. Bastiaan Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Detroit, 1995, vol. 1, pp. 508–511.

[6] W. Bastiaan Kleijn, Y. Shoham, D. Sen, and R. Hagen, "A low-complexity waveform interpolation coder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, 1996, vol. 1, pp. 212–215.

[7] N. R. Chong, I. S. Burnett, and J. F. Chicharo, "Adapting waveform interpolation (with pitch-spaced subbands) for quantisation," in *Proc. IEEE Workshop on Speech Coding*, Porvoo, 1999, pp. 96–98.

[8] N. R. Chong-White and I. S. Burnett, "Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition," *IEE Electronics Letters*, vol. 36, no. 14, pp. 1245–1247, Jul. 2000.

[9] W. Bastiaan Kleijn, P. Kroon, L. Cellario, and D. Sereno, "A 5.85 kbit/s CELP algorithm for cellular applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, 1993, vol. 2, pp. 596–599.

[10] M. Arjona Ramírez, "A waveform extractor for scalable speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Hong Kong, 2003, vol. 2, pp. 169–172.