A WAVEFORM EXTRACTOR FOR SCALABLE SPEECH CODING

Miguel Arjona Ramírez[*]

Electronic Systems Eng. Dept. - Escola Politécnica
05508-900 University of São Paulo, São Paulo, SP, Brazil

ABSTRACT

Waveform interpolation (WI) models for speech coding contain many parameters whose sampling rates may not be simply related so that most implementations tend to fix their rates right from the waveform extraction stage, thereby compromising quality by departing from perfect reconstruction. A waveform extractor is proposed which samples waveform cycles of the original prediction residual signal at their natural variable rate so that it can perfectly reconstruct the signal. The speech coder, which may operate at a uniform sampling rate, is coupled to the waveform extractor by means of an evolving waveform interpolator that may handle several interpolation methods and sampling rates for a variety of fixed and variable rate coders, including conventional WI coders.

1. INTRODUCTION

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 waveform 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 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].

[*]E-mail: miguel@lps.usp.br

This waveform 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 FEATURES

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 time-warped extracted waveforms.

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. WAVEFORM 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 by [8] and [9]. A pitch period value $p_0(n_i)$ 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.

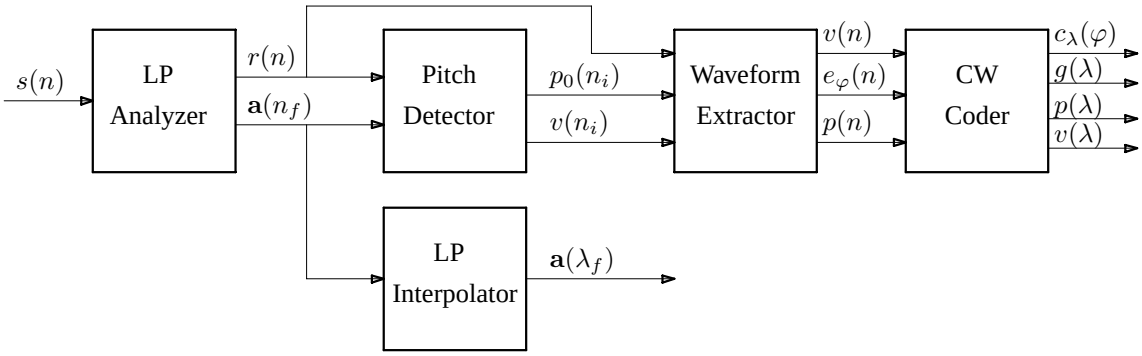

Fig. 1. Block diagram of generic speech coder that uses the waveform extractor.

The pitch period values ease the task of the waveform demarcator which looks for the endpoints of pitch cycles (see Fig. 2). 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 waveform while the terminating endpoint $n = d(n_c)$ is placed at a low-amplitude position between the next two pitch peaks, which in their turn are constrained to lie around their corresponding integer pitch cycle distances from the previous peak as determined by the pitch detector within a set tolerance margin. As an important result of the waveform picking process for scalibility, 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. TIME WARPING

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 time warping to a constant cycle length domain. Usually, the evolving Fourier series coefficients

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

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

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

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 waveform cycle

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

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

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

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, $K_l = K_u = p(n_c)/2$. The CW is obtained by the extended Fourier series

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

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

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

whereas

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

for p_{n_c} odd, and

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

for p_{n_c} even. 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. One waveform cycle is sampled at the nonuniform rate signaled by time index n_c. It is further interpolated to the constant length of P samples by

$$c_{n_c}(\varphi) = M \sum_{m=0}^{p(n_c)-1} c_{n_c}(m) h(\varphi - Mm) \tag{7}$$

for an upsampling factor $M = P/p(n_c)$, that is generally not an integer. In principle, the scaled sinc function is

$$\begin{aligned} h(n) &= 2f_c \frac{\sin(\omega_c n)}{\omega_c n} \\ &= 2f_c \text{sinc}(2f_c n), \end{aligned} \tag{8}$$

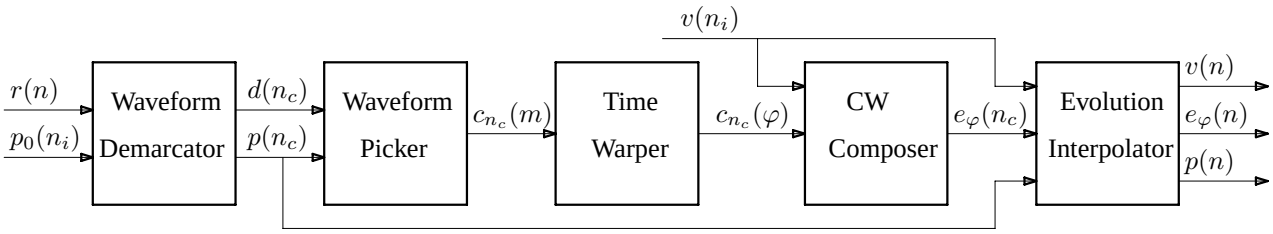

Fig. 2. Block diagram of the waveform extractor.

where $f_c = \frac{1}{2M}$ is the lowpass cutoff frequency of the interpolator. In the decoder the characteristic waveform is unwarped by

$$\tilde{c}_n(m) = \sum_{\varphi=0}^{P-1} c_n(\varphi) h(Mm - \varphi). \tag{9}$$

For a rectangular window, using $2D+1$ samples for interpolation, the warped signal is generated as

$$c_{n_c}(\varphi) = M \sum_{m=\frac{\varphi}{M}-D}^{\frac{\varphi}{M}+D} c_{n_c}(m) h(\varphi - Mm) \tag{10}$$

and it may be recovered by

$$\tilde{c}_n(m) = \sum_{\varphi=M(m-D)}^{M(m+D)} c_n(\varphi) h(Mm - \varphi). \tag{11}$$

Assuming that the CW is shrunk back to its original length $p(n)$, the decoder constructs its phase track

$$\tilde{m}(n) = \left(m(0) + \sum_{i=1}^{n} 1 \right) \mod p(n) \tag{12}$$

out of the pitch track $p(n)$ and uses it for sampling the CW for the reconstructed residual signal

$$\tilde{r}(n) = \tilde{c}_n(\tilde{m}(n)). \tag{13}$$

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

5. WAVEFORM BINDING

The sequence $\left\{ \{c_{n_c}(\varphi)\}_{\varphi=0}^{P-1} \right\}_{n_c}$ of time-warped characteristic waveforms makes up a characteristic surface when the waveforms are aligned and properly layed out along time axis n. The waveform extraction process outlined in Section 3 guarantees a great degree of alignment between consecutive extracted waveforms 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 pitch period, which the CW composer included in Fig. 2 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 along its cycle of occurrence as

$$c_n(\varphi) = c_{n_c}(\varphi) \tag{14}$$

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 time warping in Section 4, 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

$$e_\varphi(\lambda) = Q \sum_{n=\frac{\lambda}{Q}-D}^{\frac{\lambda}{Q}+D} e_\varphi(n)\, h(\lambda - Qn), \tag{15}$$

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_{CW}. The evolving waveforms may be upsampled for synthesis by

$$\tilde{e}_\varphi(n) = \sum_{\lambda=Q(n-D)}^{Q(n+D)} e_\varphi(\lambda) h(Qn - \lambda). \tag{16}$$

6. EXPERIMENTS

The waveform extractor has been tested at the natural cycle rate upsampled to the signal sampling rate $f_s = 8$ kHz and it has also been used to emulate the uniform CW sampling rate $f_{\mathrm{CW}} = 400$ Hz established by [5]. Characteristic waveforms are represented in the normalized phase domain where they are time warped to by Fourier-series extension and unwarped from 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. (13).

First of all, characteristic waveform extraction at the natural cycle rate with Fourier-series time warping to length $P = 256$ along the phase axis attains virtually perfect reconstruction with

SNRSEG values in excess of 270 dB. 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, the average SNRSEG is about 30 dB, matching the performance of the conventional waveform extraction process [2].

7. CONCLUSION

The process of waveform cycle extraction, usually described in continuous time, has been recast in discrete time by upsampling the various signals and parameters to the common signal sampling rate. A waveform extractor was proposed which extracts the waveform cycles at their natural rate while preserving their perfect reconstruction capability, thereby decoupling the extractor from the coder, which may operate at a uniform sampling rate or at variable rate as required. Interpolation and downsampling of the resultant evolving waveforms reproduces the uniform waveform sampling usually performed by WI coders. Besides, several interpolation methods and sampling rates provide results lying between perfect reconstruction and conventional WI performance.

8. 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 kbits celp algorithm for cellular applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Minneapolis, 1993, vol. 2, pp. 596–599.