A PROTOTYPE WAVEFORM INTERPOLATION LOW BIT RATE SPEECH CODEC

Gloria Menegaz [†], *Michele Mazzoleni* [‡]
[†] DE-LTS, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Tel: +39 2 66161267; fax: +39 2 66100448
e-mail: menegaz@mailer.cefriel.it

[‡] CEFRIEL, Via Emanueli 15, I-20126 Milano

ABSTRACT

Voiced speech is characterized by a high level of periodicity. In order to encode voiced speech with a good quality, the correct degree of periodicity must be preserved. The proposed coding algorithm attempts to reproduce the correct level of periodicity even at low bit rates. The method exploits the temporal redundancy of voiced segments in order to achieve high compression rates. Voiced speech is interpreted as a concatenation of slowly evolving pitch-cycle waveforms. The signal is synthesized by waveform interpolation from a downsampled sequence of pitch-cycles with a rate of one prototype waveform per frame (20-30ms). An original method of prototype parametrization and coding based on a proper mixed time-frequency representation allows a high quality prototype reconstruction. The effectiveness of such a parametrization renders it well suited to low bit rate applications, yet maintaining a good quality of the reconstructed signal. The method can be combined with existing LP-based speech coders, such as CELP, for unvoiced segments.

1 Introduction

In low bit rate speech coders, improper reproduction of the signal periodicity may result in various artifacts in the reconstructed speech: a hoarse synthetic speech is due, in general, to a superimposed noisy component which is uncorrelated with adjacent pitch cycles; tonal artifacts are often due to an increase of periodicity in the synthetic signal and reverberation has been associated with the lack of phase coherence during successive pitch cycles. The overall purpose of the proposed coding algorithm is to exploit the regularity features of the voiced segments while preserving the correct level of periodicity to compression purposes. The algorithm is based on the Prototype Waveform Interpolation (PWI) approach, which consists in representing voiced speech on a *pitch cycle* basis. The basic idea of PWI based coders is to extract a representative pitch-cycle, the *prototype* waveform, at regularly spaced intervals, to transmit a description of this prototype, and to reconstruct the voiced segment by simultaneous interpolation of both the shape and the length of the prototypes. This naturally leads to a smooth evolution of the pitch-cycle waveform, which is a common feature of the evolving cycles in voiced speech. The PWI method is specifically taylored on the voiced speech characteristics, and is not suited to the processing of the unvoiced segments; the Code-Excited Linear Predictor (CELP) algorithm can be used to that purpose.

The paper is organized as follows: in Section 2 the principles of the PWI method are discussed. Section 3 describes the analysis by synthesis system. In Section 4 the prototype parametrization and coding process is described. Results are given in Section 5.

2 Principles of Prototype Waveform Interpolation

The PWI method ([1], [2]) is based on the assumption that, for voiced speech, a single pitch cycle adequately represents the information content of the whole frame. A good quality synthetic signal can thus be obtained by waveform interpolation from the subsampled sequence of the representative pitch cycles. In the following, the mathematical fundamentals of the PWI method are briefly reviewed.

The *nearly-periodic* nature of the voiced speech allows the modellization of a given segment as a periodic function with time-varying parameters [1]. At each time instant such parameters can be *frozen*, generating the function associated to that instant. In the continuous-time domain it is thus conceivable to associate to the voiced signal an infinite set of functions $z(t,\tau)$ with period $p(t)$ as a function of τ, t being the time index. A corresponding set of *instantaneous waveform* $u(t,\phi)$ can be derived by normalization of $z(t,\tau)$ with respect to the pitch-period:

$$u(t,\phi) = z\left(t, p(t)\frac{\phi}{2\pi}\right) \qquad (1)$$

With these assumptions, the signal can be interpreted as an infinite sequence of infinitesimal segments belonging to the different istantaneous waveforms. Each segment is obtained by sampling the corresponding instantaneous function, and has to be time-scaled to recover the correct periodicity [1].

In order to translate the theoretic foundations into an automated procedure, some semplifications must be introduced. If the pitch period markers of the signal can be identified accurately, the functional set can be sampled at time instants corresponding to such markers. Furthermore, only one pitch cycle per frame is transmitted, the in-between cycles being reconstructed by interpolation.

3 The PWI Analysis/Synthesis System

The critical issues are extraction, interpolation and coding of prototype waveforms.

Prototype waveform extraction is achieved by segmentation of the current frame into its constituent cycles. The last pitch cycle is then chosen as the current frame prototype. Pitch cycles identification is performed by a robust pitch estimation algorithm. It has to be emphasized that the correct pitch tracking is determinant for the quality of the reconstructed signal since it allows to preserve the correct phase relationship between successively extracted prototypes. This will avoid the introduction of discontinuities in the junction point of successive cycles of the reconstructed signal during the interpolation process. Since the PWI method is specifically tailored for voiced speech coding, a robust voiced/unvoiced segmentation tool is needed, suitable to work on a cycle by cycle basis. Such a system is currently under development, so a hand-made phonetic classification was performed to test the algorithm.

3.1 Prototype Waveform Extraction

Pitch estimation is performed by a robust algorithm based on the maximization of a cross-correlation function. A double search procedure is followed in order to exploit the voiced segment regularity features referring both to the past and to the future evolution of the signal. With reference to the past, the pitch of the last identified pitch-cycle is used as a first estimate of the pitch of the current cycle. The current pitch period is searched for in a given surrounding of the reference one by maximizing the correlation between the corresponding waveforms.

The tracking of the future signal evolution is based on the interpretation of the pitch period as the common length which maximizes the cross-correlation of two adjacent signal segments. An exaustive scanning of the pitch variability interval is carried out in order to identify such a length. Some thresholds and different kinds of control procedures are used to avoid estimation errors. The procedure is iterated until the frame boundary is reached. Since the constraint is the identification of an integer number of pitch cycles around a predefined *nominal frame length*, the system is intrinsecally a variable rate one.

Fig.1 shows the estimated pitch profile for a female speaker with a pitch frequency of about 330 Hz pronouncing the sentence *Why were you away a year Roy?* (file f1 of the test set). The smoothness of the profile in the central voiced segment gives an example of the algorithm robustness.

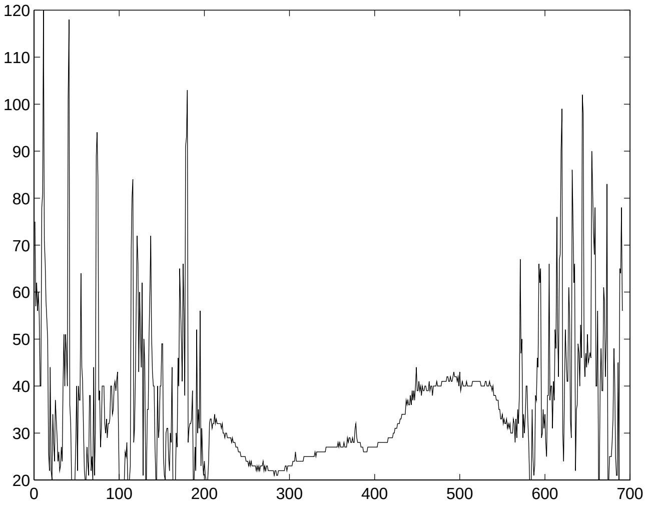

Figure 1: Pitch profile for file f1 of the test set.

3.2 Waveform Interpolation

The description and interpolation of the prototype waveforms must preserve the properties of the voiced signal. The istantaneuos waveform and the pitch period evolve slowly over time, allowing the interpolation over time intervals. The interpolation method should not generate discontinuities in the synthesized signals and the pitch profile must be smooth.

The basic assumption for interpolation is that the pitch and the formant structure evolve separately over time. Pitch period and waveform are thus interpolated independently. The pitch period is linearly interpolated, resulting in a smooth monotonically varying pitch period dynamics:

$$\tilde{p}_{j,N} = \frac{p_N - p_{N-1}}{n_N}(j+1) + p_{N-1} \qquad (2)$$

where p_N and p_{N-1} are respectively the pitch periods of the prototypes of the current (N-th) and the previous (($N-1$)-th) frames , $\tilde{p}_{j,N}$ is the length attributed to the j-th cycle of the N-th frame, and n_N is the number of pitch cycles of the current frame. It is important to emphasize that, in general, the original and the reconstructed signal will be asynchronous because the pitch track is different.

The waveform characteristic features are the shape factor and the dynamics. They are interpolated separately in order to preserve the respective information content. The first step of waveform interpolation is the normalization with respect to the pitch period. The prototypes are thus time-warped on an abstract temporal axis ϕ by an time-varying interpolation procedure. The waveform interpolation is then carried out in the abstract time domain ϕ, where the length of each cycle is constant.

The separate handling of the shape factor and the dynamics implies the normalization of the prototypes $v_N(i,\phi)$ with respect to the energy content:

$$\overline{v_N}(i,\phi) = \frac{v_N(i,\phi)}{\sqrt{\sum_i v_N^2(i,\phi)}} \qquad 0 \le i < \phi_m \qquad (3)$$

where $\overline{v_N}(i,\phi)$ represents the resulting function. The shape factor of each cycle $\tilde{\overline{u}}_{j,N}(i)$ is obtained by linear combination of the normalized prototypes:

$$\tilde{\overline{u}}_{j,N}(i) = \alpha_{j,N} \cdot \overline{v}_{N-1}(i) + (1-\alpha_{j,N}) \cdot \overline{v}_N(i) \qquad (4)$$

the coefficient $\alpha_{j,N}$ depending on the position of the cycle within the frame:

$$\alpha_{j,N} = \frac{\sum_{k=0}^{j-1} p_{j,N}(k)}{\sum_{k=0}^{n_N-1} p_{j,N}(k)} \qquad 0 \le j < n_N \qquad (5)$$

In general, the reconstructed cycle $\tilde{\overline{u}}_{j,N}$ is not energy normalized. In order to recover the correct dynamic range, the root mean square (RMS) of the prototypes is linearly interpolated along the frame. The RMS value $\tilde{R}_{j,N}$ associated to j-th cycle is thus obtained as the linear combination of the RMS values R_N and R_{N-1} of the prototypes:

$$\tilde{R}_{j,N} = \frac{R_N - R_{N-1}}{n_N}(j+1) + R_{N-1} \qquad 0 \le j < n_N \quad (6)$$

The last step of the cycle reconstruction procedure is denormalization. The correct energy content is restored by multiplying each sample value for a proper correction factor:

$$\tilde{u}_{j,N}(i,\phi) = \tilde{\overline{u}}_{j,N}(i,\phi)\frac{\tilde{R}_{j,N}}{\sqrt{\sum_i \tilde{\overline{u}}^2_{j,N}(i,\phi)}} \qquad (7)$$

Finally, each interpolated pitch cycle is time-warped according to its interpolated pitch length.

4 Prototype Parametrization and Coding

The adopted coding scheme for the prototype is based on a properly mixed time-frequency description: an LPC-based prototype parametrization and an LPC excitation based on a single pulse waveform.

In the classical LPC coefficient extraction methods, signal is assumed to be zero outside a predefined fixed-length time window. This approach may cause undesidered edge effects leading to distortions in the LPC representation of the signal spectrum. The adopted LPC-based prototype parametrization exploits the nature of the prototype as the representative cycle of a nearly-periodic waveform. This leads to the *modified-autocorrelation* method in which the length of the time window for the LPC coefficient computation is exactely equal to the prototype length. The aforementioned edge effects are avoided by employing a periodic signal extention outside the time window. In this way, *modified* correlation coefficients can be computed by including also signal samples which fall outside of the time window.

A recent approach makes use of a single-pulse excitation to represent the prototype waveform [3]. However, such an excitation model may not be suited to obtain a high quality reconstructed speech because it implies a minimum phase representation. As the phase content of the reconstructed signal is so important for its quality, a special stress about it is needed [4]. In order to improve the synthetic prototype quality, a phase-adapted single pulse (*PASP*) scheme has been developed. This consists in properly shaping the classical single pulse phase spectrum in order to minimize a pre-defined error criterion. To this purpose, the phase spectrum of the classical single-pulse is varied by successively attributing to its lowest N_A harmonic componenents a set of phase values defined by a grid of N_P different samples. The mininization of the mean square error between the original and the reconstructed waveforms leads to the determination of the phase sample values retained for each harmonic component. The LPC excitation is thus modeled by a waveform obtained by proper variation of a single-pulse phase spectrum, and is parametrized in terms of a position, an amplitude and a phase spectrum.

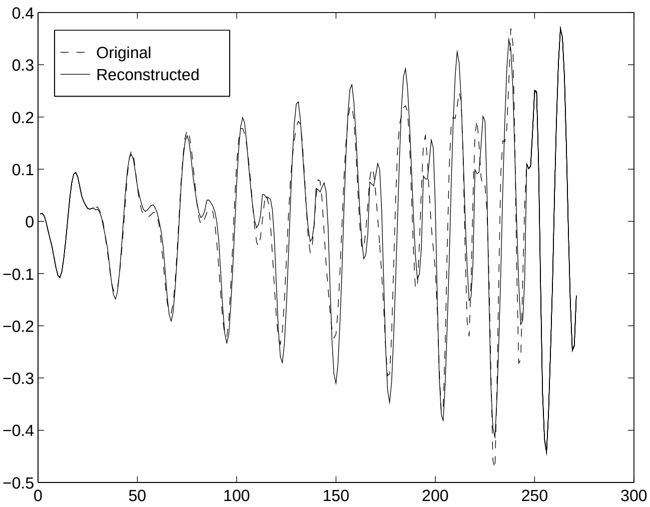

Figure 2: Original and PWI-reconstructed frames.

5 Results

The algorithm performance was evaluated with an informal listening test. Due to the pitch interpolation process, in general the original and the reconstructed signals will be asynchronous. This prevents the use of the signal to noise ratio (SNR) as a quality factor for the synthesized signal. One way to evaluate the per-

formance of the interpolation process is to overlap the original and the reconstructed signals as shown in fig.2, concerning file f1 of the test set. The cycle waveform smoothly evolves between the two reference prototypes, well reproducing the correct signal dynamics.
On the other end, the performance of the coding algorithm can be quantified. In fig.3 the SNR between the original prototypes and the reconstructed ones is depicted; the dashed line refers to a single-pulse excitation waveform, while the solid line refers to a *PASP* excitation waveform. These are respectively shown in fig.4 together with the original prototype. Different values of N_A and N_P have been tested in order to find out the best compromise between SNR improvement and bit overhead. Table1 shows the SNR between the original prototypes and the synthetic counterparts. The measures of the average, minimum and maximun SNR(dB) were obtained by varying the phase samples at the lowest N_A harmonic frequencies according to a phase grid of N_P samples. An overall bit number:

$$Bit = N_A log_2(N_P) \qquad (8)$$

was allocated for phase sample description. The results show that the SNR increases as the number of the phase-adapted frequency samples increases, and as a finer phase grid resolution is adopted. Simulations also show that the SNR tends to saturation beyond the values $N_A = 3$ and $N_P = 8$ (Bit=9) for all speech files of the test set. A considerable improvement of the recontructed prototype quality can thus be achieved with a very little bit overhead.

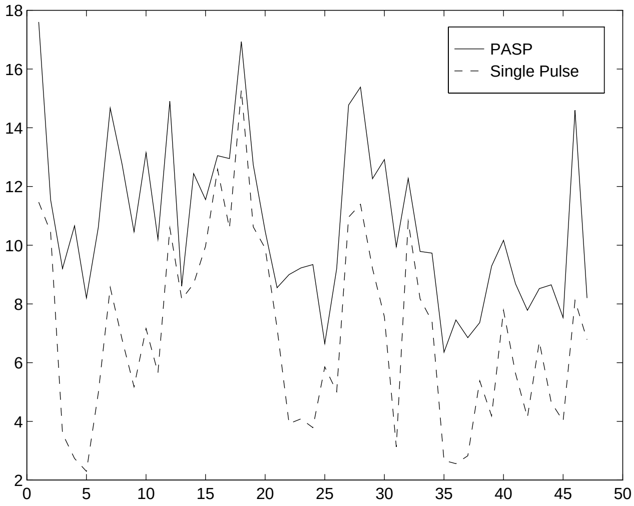

Figure 3: SNR between the original prototype and the single-pulse and *PASP* reconstructed ones.

6 Conclusions

A PWI based codec has been presented, which allows good quality voiced speech reproduction at low bit rates. A robust pitch estimation algorithm which works on a cycle by cycle basis has been conceived and implemented. The information content of each prototype is described in terms of waveform pitch period, shape factor and dynamics. All of these parameters are interpolated separately with a linear interpolation scheme. A modified LPC representation is adopted for prototype parametrization and coding. A time-frequency version of the classical single pulse model is introduced to code the LPC excitation residual, leading to a phase-adaptive single pulse excitation.

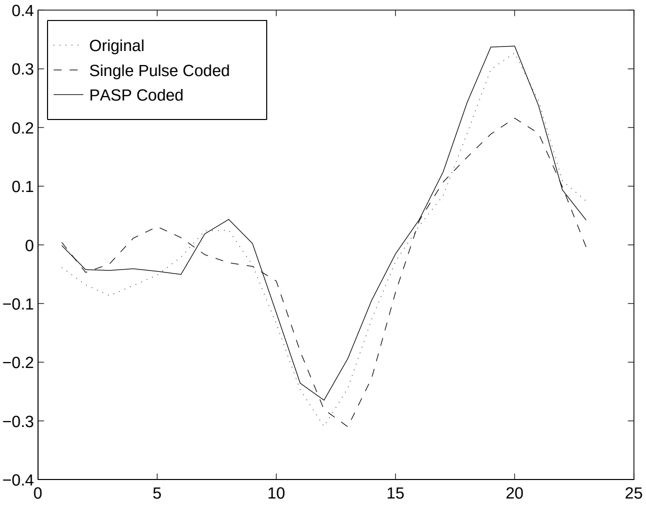

Figure 4: Original prototype, SP and *PASP* reconstructed ones.

N_A	N_P	Bit	$Average$	min	Max
0	0	0	6.83	1.42	17.52
1	4	2	8.28	2.57	18.25
2	8	6	9.95	3.07	19.76
3	4	6	9.75	3.11	19.76
3	8	9	10.58	3.24	21.36

Table 1: Average, miminum and maximum SNR.

References

[1] W.B. Kleijn "Encoding Speech using Prototype Waveforms," *IEEE Transactions on Speech and Audio Processing*, Vol.1, No.4, October 1993.

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

[3] W. Granzow and B.S. Atal "High-quality digital speech at 4kb/s," *Proceedings of GLOBECOM-90*, San Diego, Vol. 2, December 1990, 507B.1.1.

[4] P. Lupini and V. Cuperman "Excitation modeling based on speech residual information," *ICASSP-1992*, San Francisco, March 1992, Vol. 1, I-333.