NON-LINEAR TECHNIQUES FOR PITCH AND WAVEFORM ENHANCEMENT IN PWI CODERS

Hui Li, Gordon B. Lockhart

Dept. of Electronic and Electrical Engineering
The University of Leeds, Leeds, LS2 9JT, UK

ABSTRACT

Two non-linear interpolation techniques are introduced for enhancing speech reproduction in Prototype Waveform Interpolation (PWI) and similar encoders. A Temporal Differential Rate (TDR) vector is used to characterise the non-uniform evolution of pitch cycle temporal structure during interpolation. Experimental results show a clear improvement in the accuracy of decoded pitch cycle lengths and in the reproduction of periodicity in general. It is also shown that waveform reproduction can be significantly improved by vector quantising sets of Optimal Combination Coefficients (OCC) aimed at maximising the similarity between interpolated and target signal segments. Both time domain waveform similarity and frequency domain spectral envelope similarity derived OCC are tested. Subjective assessment suggests a general preference for non-linear interpolation methods and the scheme using frequency domain derived OCC with perceptual weighting provided the best subjective preference.

1. INTRODUCTION

Prototype Waveform Interpolation (PWI) [1,2,3] is a promising low bit rate coding technique applicable to voiced speech. It is characterized by transmitting only one pitch cycle per frame and reconstructing the missing speech between the prototypes by linear interpolation. This characteristic pitch cycle is referred to as a prototype and generally is at or near the end of the updating frame. Waveform Interpolation (WI), extends PWI concepts to both voiced and unvoiced speech[4,5], has been considered an important codec in the 1990s. The reproduction performance of WI and PWI are surprisingly impressive at low bit rates[1,5].

Studies of waveform coders, including CELP, reveal that waveform similarity enhancement invariably leads to improved reconstruction quality[6] and this is also generally true for PWI[7,8]. We show that waveform similarity in PWI systems can be enhanced by introducing two non-linear interpolation techniques. A Temporal Differential Rate (TDR) vector accounts for non-uniform evolution of pitch temporal structure in the interpolated region while a vector of Optimal Combination Coefficients (OCC) provides for optimal interpolation aimed at maximizing waveform shape or spectrum envelope similarity between the original signal and the reproduction. Simulation results are based on the Variable PWI (VPWI) coder[9] using a variable frame length to ensure that each transmission frame covers an integer number of pitch cycles.

2. REPRESENTATION OF PITCH CYCLE TEMPORAL EVOLUTION BY THE TDR VECTOR

The magnitudes of pitch cycle lengths over typical frames are usually less variable than cycle to cycle changes, suggesting the use of differential encoding. Let vector $\boldsymbol{t}_k = [t_{0,k}, t_{1,k}, t_{2,k}, \cdots, t_{M,k}]$ represent pitch cycle lengths in samples for the *k*-th frame of input speech. $t_{0,k}$ denotes the last pitch cycle length in frame *k-1* and $t_{1,k}$ to $t_{M,k}$ are the *M* pitch cycle lengths associated with the current frame *k*. The TDR vector is therefore defined as

$$\boldsymbol{r}_k = [\, r_{1,k}, r_{2,k}, \cdots r_{M,k} \,]$$
$$r_{i,k} = \frac{t_{i,k} - t_{0,k}}{t_{0,k}}, \quad i = 1,2,\cdots M \qquad \textbf{(1)}$$

Given r_k and the initial condition $t_{0,k}$, the $\boldsymbol{t}$ vector can be reconstructed recursively. An error criterion, e_k was devised to emphasize the accuracy of the last cycle length of the current frame:

$$e_k = \frac{\sum_{i=1}^{M-1} \alpha \left| t_{i,k}^m - t_{i,k} \right| + \beta \left| t_{M,k}^m - t_{M,k} \right|}{(M-1)\cdot\alpha + \beta} \qquad t_{i,k}^m = t_{0,k}\left(r_{i,k} + 1\right)$$

Superscript *m* denotes reconstructed cycle lengths. A suitable choice of the weighting coefficients, α and β prevents error propagation in the recursive reconstruction. This error criterion can be replaced by using the decoded previous prototype length for encoding current TDR vector forming a feedback control loop. Similar coding performance results from these two approaches.

Table 1. Quantization performance

Entries	Partition	Avg. SNR(dB)	No Error (times)	Min. SNR (dB)	No. of Outliers
512	No	40.45	5900	7.48	30
256	No	39.56	3946	7.78	31
128	No	39.04	2837	7.37	45
512	480+32	40.46	5525	14.00	0
256	224+32	39.43	3625	14.02	0
128	104+24	38.77	1936	13.57	1

Because the number of pitch cycles, M is a frame-dependent parameter, variable dimension vector quantization (VD-VQ)[10] was used to encode the vector $\boldsymbol{r}$. Pitch cycle lengths can vary very rapidly in transitional and quasi-periodic regions and although average SNR can be very high, these regions produce in codebook training, a number of codewords with low associated SNR. Some codebook entries were reserved for such regions by employing a partitioned codebook structure. The VD-VQ scheme uses a single codebook with "overlapping" codevector dimensions (rather than the multiple sub-codebook structure[10]) with the advantage that codevector dimensions need not be individually specified. If $\boldsymbol{N}$ is the codebook size and $M_{\max}$ is the maximum possible dimension, a $\boldsymbol{N}\times M_{\max}$ codebook is required. The dimensions of any L dimensional codevector are in common with the first L dimensions of higher dimensional codevectors and the first L dimensions of every codevector are used to generate a match for a L dimensional input $\boldsymbol{r}$. The codevector dimension is determined as an integer as every coding frame encompasses an integer number of pitch cycles[9].

28 minutes of speech sampled at 8 kHz was used as a training source. The pitch cycle lengths for TDR encoding were produced by a pitch marker[11] operating with 200 sample frames and output cycle lengths constrained from 16 to 120 samples. We tested a variety of non-partitioned and partitioned codebook structures as detailed in Table-1. Outliers are defined for mappings with distortion exceeding 4% input energy. Frames for which the magnitude of accumulated error in cycle lengths over a frame is less than 1 sample are designated "no error" and the SNR is set to 40 dB.

Because of error propagation, it is necessary to control minimum SNR and outliers at the expense of average SNR. The maximum average SNR difference between all codebooks in Table 1 is only 1.69 dB suggesting that an extra 2 bits per frame required for 512 rather than 128 entries is hardly justifiable.

Two types of errors can arise in using the TDR scheme: errors in the reproduction of individual cycle lengths and errors in total length given by the sumation of individual lengths over an entire frame. Both were measured objectively by mean square error. Two male and female speech signals, independent of the codebook

Table 2. Cycle length distortion (MSE: samples×samples)

Codebook	(VPWI)	512 Non-Part.	256 Part.	128 Part.
signal A	167.03	7.03	7.05	8.17
signal B	197.40	3.48	3.24	5.21

Table 3. Frame size distortion (MSE: samples×samples)

codebook	(VPWI)	512 Non-Part.	256 Part.	128 Part.
signal A	180.13	71.16	85.93	76.08
signal B	218.93	37.55	31.50	60.53

training set and of approximately 10 seconds duration, was used as a test set. Experiments were devised using the TDR codebooks detailed in Tables 2 and 3 and VPWI as a reference system with a linear interpolation. Table 2 shows clear improvements in cycle and frame length reproduction using TDR encoding. In particular, the performance of the 128 entry partitioned codebook suggests that a codebook requiring 7 bits per frame is adequate for TDR quantisation**.**

In SPE-CELP[11], temporal pitch information is encoded using 34 bits per 200 sample frame (8 Khz sampling rate) with an upper limit of 400 Hz on pitch frequency. Such a bit allocation scheme gives error-free coding of pitch cycle lengths in contrast to the approximately accurate representation provided by the TDR scheme. However, we found that the error introduced by TDR is hardly perceptible and for a low bit rate speech coder such as 4.0 kbps SPE-CELP[11], a saving of 25-27 bits per frame (i.e., 1.0 - 1.08 kbps) is very significant.

The use of the TDR codebook for PWI requires at most two extra bits per frame in comparison with conventional PWI or WI systems[1-5] that require 7 bits to transmit prototype length using linear interpolation to recover pitch cycle lengths from neighbouring prototypes. To initialize the recursive TDR procedure, the true pitch cycle length is transmitted in the very first frame of each voiced speech segment instead of the TDR codebook index. It is also remains the case as in PWI when TDR scheme is applied to vocoders.

3. SIMILARITY ENHANCEMENT USING OPTIMAL COMBINATION COEFFICIENTS

In the original PWI systems missing speech is linearly interpolated at the decoder using forward and backward interpolation coefficients, normally on the basis of two given prototypes. If no constraints are placed on the interpolating coefficients a standard optimisation approach can be applied to obtain a least distortion solution. If the i

Table 4. Mean cross-correlation coefficients using VPWI to compare OCC with linear interpolation

	speech 1	speech 2	speech 3	speech 4
Linear	0.44	0.54	0.45	0.62
OCC	0.84	0.94	0.84	0.86

th missing pitch cycle $\boldsymbol{u}_i$, is to be interpolated on the basis of given previous and current prototypes, $\boldsymbol{u}_0$ and $\boldsymbol{u}_M$ respectively, then the interpolated segment, $\hat{\boldsymbol{u}}_i$, may be expressed as

$$\hat{\boldsymbol{u}}_i = \lambda \cdot \boldsymbol{u}_0 + \varsigma \cdot \boldsymbol{u}_M \qquad (2)$$

where λ and ζ are the forward and backward interpolation coefficients respectively and assuming all waveforms involved in Equ.(2) are of the same length. The *i* th error vector $\boldsymbol{e}_i$ is then

$$\boldsymbol{e}_i = \boldsymbol{u}_i - \hat{\boldsymbol{u}}_i = \boldsymbol{u}_i - \lambda \cdot \boldsymbol{u}_0 - \varsigma \cdot \boldsymbol{u}_M \qquad (3)$$

Taking partial differentials of $\boldsymbol{e}_i \boldsymbol{e}_i^t$ ($1 \le i \le M-1$) with respect to λ and ζ respectively and setting each to zero leads to

$$\lambda_{opt}(i) = \frac{c_{i,0}c_{M,M} - c_{0,M}c_{i,M}}{c_{0,0}c_{M,M} - c_{0,M}c_{0,M}}$$

$$\varsigma_{opt}(i) = \frac{c_{0,0}c_{i,M} - c_{i,0}c_{0,M}}{c_{0,0}c_{M,M} - c_{0,M}c_{0,M}} \qquad (4)$$

$$c_{j,k} = c_{k,j} = \boldsymbol{u}_j \boldsymbol{u}_k^t$$

$\lambda_{opt}(i)$ and $\varsigma_{opt}(i)$ together are referred to as the *Optimal Combination Coefficients* (OCC). Note that when $\boldsymbol{u}_0 = \boldsymbol{u}_M$ Equ.(4) has no solution but when the previous prototype is identical to the current, waveform changes in the interpolated region are likely to be moderate and simply repeating the protoype waveform over the entire interpolated region is unlikely to introduce significant distortion. (In fact, this never occurred in an experiment involving 28 minutes of input speech.)

In the case of voiced to unvoiced and unvoiced to voiced transitions only one prototype is available in which case

$$\hat{\boldsymbol{u}}_i' = \rho \cdot \boldsymbol{u}_a \quad \text{and} \quad \rho_{opt}(i) = \frac{\boldsymbol{u}_i \cdot \boldsymbol{u}_a^t}{\boldsymbol{u}_a \cdot \boldsymbol{u}_a^t} \qquad (5)$$

Efficient tranmission of the OCC can be achieved using vector quantisation techniques to exploit correlations between vectors of forward and backward combination coefficients $\lambda_{opt}(i)$ and $\varsigma_{opt}(i)$ and between the dimensions of each vector.

Direct use of the OCC at the decoder for waveform recovery results in a blockwise nonlinear interpolation function leading to waveform discontinuities when the coefficients change abruptly at the pitch cycle boundaries, clearly affecting reproduction quality. A smoothing process was therefore used to generate appropriate bi-directional interpolation coefficients at every sampling instant in the interpolated region. We tested a variety of methods such as low-pass filtering the blockwise OCC, polynomial interpolation and Lagrange interpolation. Objective differences between these methods were not significant although an informal listening test slightly favoured Lagrange interpolation.

The cross-correlation coefficient provides a satisfactory objective measure of the performance of the OCC schemes because of its energy invariant property. 4 segments of adult male and female speech of about 40 seconds total duration served as a test signal. The results summarised in Table 4 indicate that a significant objective improvement is achieved in comparison with linear VPWI.

OCC may be applied in both time domain and frequency domains. If u_i, u_0 and u_M now refer to the input spectral envelope, the OCC can be used to maximise the reproduction fidelity of an interpolated segment with respect to the spectral envelope of the original signal segment. Because the Fourier transform follows the distributive law, frequency domain OCC can be used directly for the time domain interpolation or vice versa. Auditory masking effects can be exploited more easily in the frequency domain by employing a LPC-based perceptual weighting technique[12] so that

$$\lambda_{opt}^w(i) = \frac{\hat{c}_{i,0}\hat{c}_{M,M} - \hat{c}_{0,M}\hat{c}_{i,M}}{\hat{c}_{0,0}\hat{c}_{M,M} - \hat{c}_{0,M}\hat{c}_{0,M}}$$

$$\varsigma_{opt}^w(i) = \frac{\hat{c}_{0,0}\hat{c}_{i,M} - \hat{c}_{i,0}\hat{c}_{0,M}}{\hat{c}_{0,0}\hat{c}_{M,M} - \hat{c}_{0,M}\hat{c}_{0,M}} \qquad (6)$$

$$\hat{c}_{j,k} = \hat{c}_{k,j} = \boldsymbol{u}_j \boldsymbol{W}\boldsymbol{W}^t \boldsymbol{u}_k^t$$

where $\boldsymbol{W}$ denotes the diagonal weighting matrix and $\lambda_{opt}^w(i)$ and $\varsigma_{opt}^w(i)$ are perceptually weighted, frequency domain derived OCC.

Fig.1 illustrates OCC non-linear interpolating functions derived in time and frequency domains from the same input speech. We devised an informal listening test involving 4 individuals (all working on speech compression but not on PWI). The results showed that the linear VPWI scheme was the least preferred. VPWI using time-domain derived OCC produced noise, typical of waveform encoders. The differences between time and frquency domain OCC methods were not significant although the distortion introduced was different in nature. The perceptually weighted frequency domain derived OCC method was most preferred confirming the effectiveness of auditory masking.

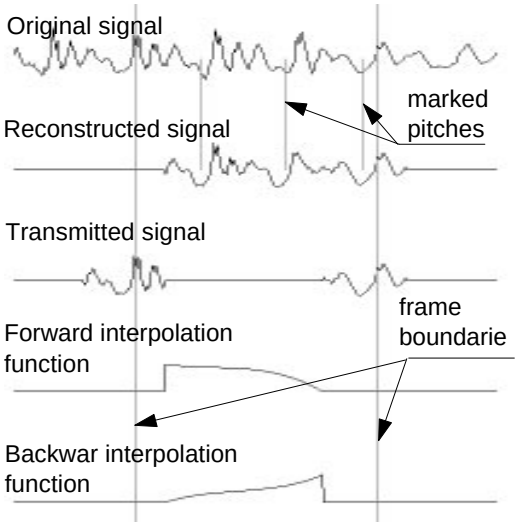

Fig. 1.1 Interpolation functions using frequency domain derived OCC (without perceptual weighting) with the same input as in Fig. 1.2

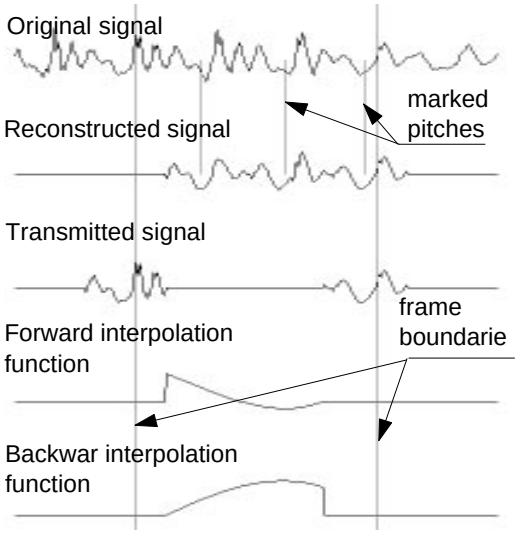

Fig. 1.2 Interpolation functions using time domain derived OCC with the same input as in Fig. 1.1

4. CONCLUSIONS

The use of non-linear interpolation methods in PWI can significantly enhance reproduction quality. Time or frequency domain similarity may be enhanced depending on how the OCC are derived. The use of TDR and OCC is not restricted to PWI and, at least in principle, may also be applied to WI, SPE-CELP and some high bit rate modern vocoders.

The non-linear interpolation methods that have been discussed are aimed at improving waveform similarity and perceptual quality in general. However, such techniques may reduce periodicity or regularity of waveform evolution in comparison with linear PWI and WI coders. It is not clear what degree of periodicity enhancement is subjectively preferred but certain listeners evidently do prefer enhancement of periodicity rather than a general improvement in waveform similarity. In conventional PWI systems, a special measure based on cross-correlation named the Signal-to-Change Radio (SCR) is used to prevent excessive periodicity leading to reverberation and/or buzziness[1,5]. This function is no longer required in non-linear PWI schemes but could prove useful in controlling periodicity to suit particular sujective requirements.

REFERENCES

1. W.B.Kleijn and W.Granzow, “Methods for Waveform Interpolation In Speech Coding”, Digital Signal Processing 1, (1991), pp215-230
2. W.B.Kleijn, “Continuous Representations in Linear Predictive Coding”, ICASSP 1991, pp201-204
3. W.B.Kleijn “Encoding Speech Using Prototype Waveform”, IEEE Trans. on Speech and Audio Processing Vol.1 No.4 Oct 1993, pp386-399
4. W.B.Kleijn and J.Haagen, “Transform and Decomposition of the Speech Signal for Coding”, IEEE Signal Processing Letters, Vol.1, No.9, Sept.1994, pp136-138
5. W.B.Kleijn and J.Haagen, “A Speech Coder Based On Decomposition of Characteristic Waveform”, ICASSP 1995, Vol.I, pp508-511
6. A.Gersho, “Advances in Speech and Audio Compression”, Proceedings of the IEEE, Vol.82 No.6 June 1994, pp900-918
7. H.Li and G.B.Lockhart, “Non-linear Interpolation in Prototype Waveform Interpolation (PWI) Encoders”, IEE Colloquium, Digest No.1994/138, June 1994, London
8. H.Li and G.B.Lockhart, “Non-linear Prototype Waveform Interpolation for Voiced Speech Encoding”, Fifth IEE International Conference on Telecommunication, April 1995, Brighton, U.K.
9. K.W.Tang and B.M.G.Cheetham, “Variable Frame Length Prototype Waveform Interpolation for Low Bit Rate Speech Coding”, IEE Colloquium, Digest No.1993/234, Dec. 1993, London
10. A.Gersho and R.M.Gray, “Vector Quantization and Signal Compression”, London Boston : Kluwer Academic publishers, 1992
11. W.Granzow, B.S.Atal, K.K.Paliwal and J.Schroeter, “Speech Coding at 4 Kb/s and lower Using Single-Pulse and Stochastic Models of LPC Excitation”, ICASSP 1991, pp217-220
12. M.R.Schroeder and B.S.Atal, “Code-Excited Linear Prediction (CELP): High-quality Speech at Very Low Bit Rates”, ICASSP-85, pp937-940