MULTI-PROTOTYPE WAVEFORM CODING USING FRAME-BY-FRAME ANALYSIS-BY-SYNTHESIS

I.S. Burnett and D.H. Pham

Department of Electrical and Computer Engineering,
University of Wollongong, NSW, Australia
http://www.whisper.elec.uow.edu.au
i.burnett@elec.uow.edu.au

ABSTRACT

A new mechanism for using Analysis-by-Synthesis techniques in low rate Waveform Interpolation based coders is introduced. The algorithm, implemented as part of a Multi-Prototype Waveform coder, exploits the high quality speech produced by interpolating unquantised speech-domain Prototype Waveforms. In the new scheme, a frame of Prototype Waveforms is quantised using two sets of codebook searches, one representing the slowly evolving prototype shape and the other the rapid, noisy components. The scheme offers performance advantages over the previous open-loop Multi-Prototype Waveform coder, particularly when perceptual weighting is incorporated in the search. Reductions in search complexity and the use of the scheme for quantisation at higher rates are also considered. This results in a generalised Analysis-by-Synthesis Waveform Interpolation architecture with closed-loop optimisation of all Prototype Waveform properties.

1. INTRODUCTION

Prototype Waveform based coders and, more generally, coders which use the principles of Waveform Interpolation have been found successful for speech coding at rates as low as 2.4kb/s [1-3]. One of the distinguishing features of WI coders over other 2.4kb/s algorithms is that they offer scalability to higher rates. Primarily this scalability can be achieved by improving the quantisation of the Prototype or Characteristic Waveforms. It has been reported by several authors that high quality near-transparent speech, can be generated from unquantised Characteristic Waveforms using the WI technique [1,2]. This property of the WI technique should be qualified by the requirement that the Characteristic Waveforms need to be extracted accurately, with appropriate pitch. Further, the speech must be reconstructed using an appropriate, interpolated pitch track. However, if these requirements are carefully followed, high quality speech results.

This paper utilises the near-transparent speech property to produce a new Analysis-by-Synthesis quantisation mechanism. This contrasts with previous work whereby the residual prototypes are quantised directly using variable length VQ techniques [4-6]. The latter decompose the successive prototypes into slowly and rapidly evolving components (known as the SEW and REW [1]) which each have distinctive quantisation requirements. It should also be noted that the technique overcomes the difficulty of applying Analysis-by-Synthesis techniques to WI based techniques, namely that the reproduced speech (or residual) need not be and is normally not synchronous with the input waveform. This phenomenon prevents the use of objective measures for WI coded speech, and would, generally, prevent the use of a standard Analysis-by-Synthesis loop. The quantisation of each individual Characteristic Waveform using an A-by-S mechanism, overcomes this problem while retaining the essential elements which make WI suitable for low rate coding.

2. ANALYSIS-BY-SYNTHESIS AT 2.4KB/S

The architecture of the modified 2.4kb/s Multi-Prototype Waveform (MPW) coder is shown in Fig.1 (the decoder is similar to that in previous work [2]). The coder nominally extracts ten prototypes per 25ms frame of 200 samples. Prototypes are extracted using the algorithm detailed in [2], transformed to the DFT domain and normalised to have a mean magnitude of unity. The normalised, extracted prototypes are quantised using two codebook searches. The first search is performed on a codebook of ‘mean prototype’ vectors (effectively the SEW component in [1]). The ‘Mean Prototype’ or SEW codebook search is performed using a standard Mean Squared Error criterion, except that the vector which minimises the error across all ten extracted prototypes from the current frame is chosen.

2.1 Codebook Structures

The ‘Mean Prototype’ or SEW codebook was trained across a large database of speakers. Each vector of DFT coefficients contains magnitude and phase information. The retention of phase information, ensures that realistic phase spectra are attached to the prototype magnitudes during synthesis. It should be noted that all mean prototypes in the training database are aligned to a reference to ensure the correct accumulation of phase spectra information. The mean prototype vectors are normalised (by zero-padding of the DFT coefficients) to a standard length during training, since the length of the training vector will necessarily vary with pitch. This equates to interpolating the time-domain prototypes to a standard pitch length. During codebook searching, vectors of length ‘pitch’ are extracted from

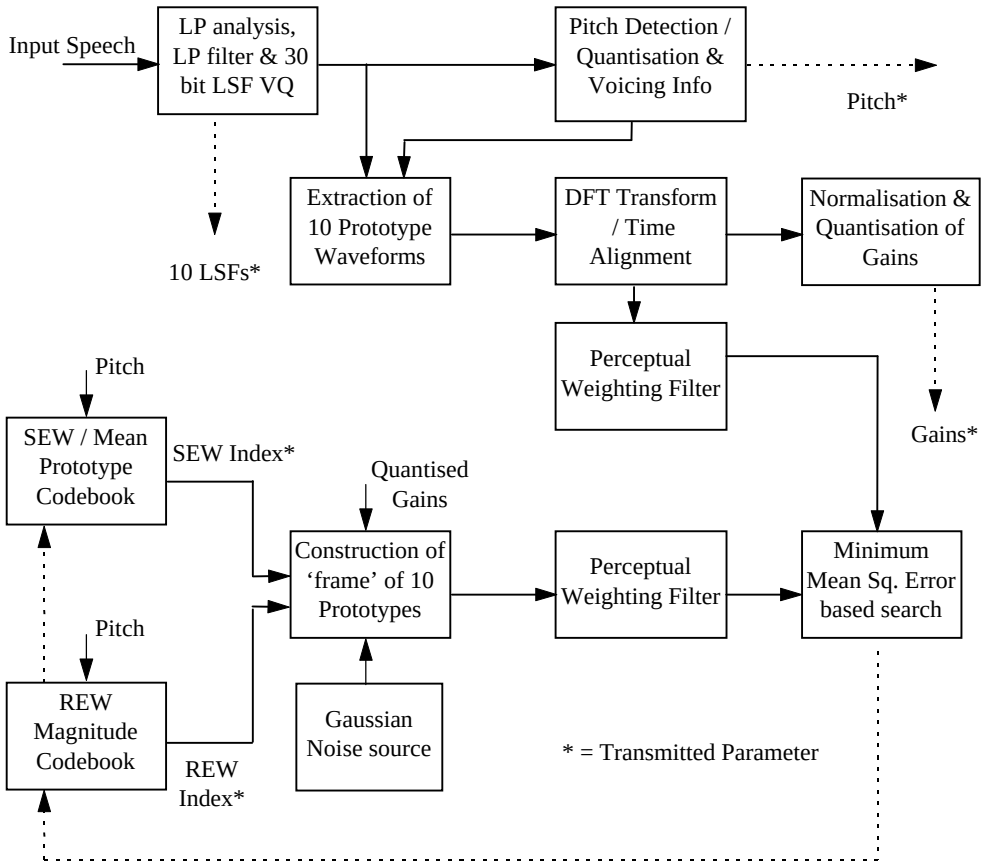

Fig.1: Analysis-by -Synthesis MPW Coder Architecture (in the speech domain)

the codebook, with superfluous vector coefficients being ignored. A SEW codebook length of 256 vectors was used.

The codebook consists of noise-like REW magnitude spectra which is trained, and used in a similar manner to the SEW codebook. The magnitude spectra are searched against the extracted prototype, following removal of the 'mean prototype' contribution. This second search can be performed multiple times if multiple REW vectors are coded per frame. Alternatively, a series of vectors forming an evolutionary pattern for the REW component can be chosen in a single search. The REW effectively adds detail to the selected mean prototype. The complete, quantised prototype is formed as the addition of the 'mean prototype' and REW components.

At 2.4kb/s explicit gain terms related to the REW and SEW cannot be transmitted and have proven unnecessary. The residual prototype DFTs are normalised to have, on average, unity magnitude and thus it is possible to derive the appropriate gain of the REW contribution from each SEW. The REW magnitude is computed as:

$$\left|\mathrm{REW}[k]\right| = 1 - \left|\mathrm{SEW}[k]\right| \quad \textit{for } k = 0,1,....,\tau-1 \qquad ....(1)$$

where τ represents the 'pitch' in samples of the current prototype. An alternative mechanism can also be used, in which the error resulting from the Unity Magnitude assumption for all coefficients is incorporated into the REW quantisation. In this case the adjusted REW magnitude can be simply computed as:

$$\left|\mathrm{REW}[k]\right| = 1 + \mathrm{error}[k] - \left|\mathrm{SEW}[k]\right| \quad \textit{for } k = 0,1,....,\tau-1 \qquad ....(2)$$

where error[k] expresses the error between the Unity Magnitude approximation and the actual residual prototype magnitude for the kth coefficient. This has the advantage of correcting for the approximation error since this error can now be explicitly quantised. Thus the normalisation of the prototype magnitude, and the subsequent transmission of an energy term is the only 'gain information' transmitted at low rates. The choice of the SEW from a codebook of normalised shapes dictates the REW gain.

2.2 Search mechanisms

Two search mechanisms for low-rate coding have been investigated. In the first, the candidate residual prototype is constructed from the SEW codebook contribution and the related

REW magnitude from Equation (2). In this case the error is quantised using a separate codebook search. The result is a candidate prototype which can be directly compared with the input residual prototype during the search process.

In the second scheme, two codebook searches are performed. The first search finds the optimum SEW from the aligned SEW codebook. The REW codebook search is then performed on the 'remainder' of the input prototype from the first search. The gains of the REW vectors are derived using the Unity Magnitude assumption of Equation (1). This mechanism is directly analogous to a CELP search.

One of the distinct problems of low-rate coding of prototypes using an Analysis-by-Synthesis mechanism is the nature of the REW and SEW. In general, it proves unnecessary [6], and low rate coders cannot afford, to send information regarding the phase of the REW; this implies that the synthesis process used at the coder must be performed entirely on the magnitude spectrum of the candidate prototype. This, however, ignores the interaction between the detailed REW component and the underlying prototype shape. In this work it was found that the introduction of random REW phase (unrelated to that used during synthesis at the decoder) during analysis, enhances the quantisation performance.

In prototype waveform quantisation, the mix of REW and SEW is crucial if buzzy, unvoiced and noisy, voiced speech are to be avoided. In general this has not proved to be a significant problem in the A-by-S MPW coder, however for some speakers buzzy, unvoiced sections can be generated. This has been overcome by constraining the choice of SEW vectors during clearly unvoiced input speech. During these segments, spectrally flat SEW vectors are chosen from a constrained section of the SEW codebook.

2.3 Synthesis and Perceptual Weighting

The A-by-S search procedures can be performed either in the residual or speech domain. It was, however, found beneficial to search directly in the speech domain (as illustrated in Fig. 1). The LPC synthesis procedure is performed by DFT domain multiplication using the spectral colouring technique suggested by Kleijn et al.[4]. A further advantage of performing the MSE computation in the speech domain is that perceptual weighting can be explicitly brought into the encoding process. In an approach reminiscent of DFT domain searched CELP [7,8] the perceptual weighting can be performed as:

$$\mathbf{P}_\gamma[\mathrm{k}] = \frac{\mathbf{A}^*_\gamma[\mathrm{k}]}{\mathbf{A}_\gamma[\mathrm{k}]\mathbf{A}^*_\gamma[\mathrm{k}]}\mathbf{P}[\mathrm{k}] \qquad \textit{for } \mathrm{k} = 0,1,\ldots,\tau-1 \qquad \ldots.(3)$$

where $\mathbf{P}_\gamma[\mathrm{k}]$ is the weighted, speech-domain, DFT prototype vector, $\mathbf{P}[\mathrm{k}]$ is the DFT of the residual prototype (or SEW/REW component), and $\mathbf{A}_\gamma[\mathrm{k}]$ is the DFT of the weighted LP coefficients. It is then a simple matter to incorporate this weighting into the codebook search procedures. More complex, threshold-based masking models are also possible (such as those used in the MPEG and AC3 audio coders) and these are currently under investigation.

In our base MPW coder, the Analysis-by-Synthesis computation results in ten pitch length prototypes being involved in the A-by-S search procedure. However, the complexity of the search will increase significantly, as noted by Kleijn et al [4], with increasing pitch. Thus an alternative search may be performed in which the number of prototypes included is limited to the number of 'unique' prototypes that exist in the frame. This can be viewed as removing the oversampling associated with extracting prototypes in the standard scheme. As an example, with prototypes extracted every 2.5ms and a pitch length of 40 samples, five unnecessary prototypes and the associated computations can be excluded from the search. This reduction in search complexity can be used to, as far as possible, normalise the search computational complexity from frame-to-frame. Our results show that the reduced complexity scheme generates no degradation in performance. The reduction in oversampling emphasises the view of this scheme as an overlapping pitchlength subframe version of CELP.

3. QUANTISATION AT HIGHER BIT RATES

If Analysis-by-Synthesis WI techniques are to be used at higher rates (e.g. 4kb/s) the search procedure and codebook techniques can be adapted in a number of ways.

3.1 Codebook Structures

At 2.4kb/s, the REW phase is not transmitted, however at higher rates exact REW phase information can be included in the search. In this case the quantisation mechanism can be divided into two, clear codebook searches, the first of a 'pitch-length' codebook of SEW pulse shapes, and the second a full search of an REW codebook containing both magnitude and phase information. The transmission of this information was found to improve the naturalness of speech in some cases. It is also possible to introduce larger codebooks for both the SEW and REW, but complexity of the coder is likely to make this infeasible. It is also worth noting that the transmission of REW phase information, further illustrates the view, expressed previously, of the proposed A-by-S quantisation scheme as an overlapping, pitch-length subframe, version of CELP.

3.2 Search Mechanisms

The relative magnitude of the REW and SEW components can also be transmitted more accurately at higher rates. This results in the A-by-S search resembling CELP more closely, in that individual gain terms for both the 'mean prototype'/SEW and REW components can be introduced. This scheme is a generalisation of the proposed 2.4kb/s low-rate scheme and is illustrated in Fig.2. In particular this brings to the forefront comparisons between the proposed techniques and PSI-CELP [9]. In drawing such comparisons, it should, however, be noted

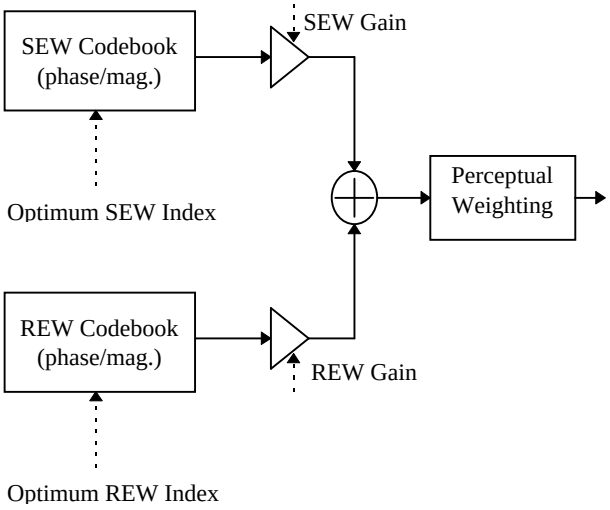

Fig.2: Generalised A-by-S Prototype Construction

that the proposed scheme retains the significant advantages of WI techniques (explicitly, the high quality reproduction at low bit rates) while using analysis-by-synthesis schemes purely for improved quantisation. At the encoder, there are, however, clear relationships between the overall search process and those used in CELP coding.

Higher bit rates also allow an increased frequency of transmission of both the SEW and REW components. This is of particular benefit to the rapidly evolving REW surface description. In this case, the combined prototype comparison mechanisms described in Section 2 can still be utilised, but with a reduced number of input prototypes per search.

4. RESULTS

The A-by-S MPW coder has been tested at 2.4kb/s in limited MOS tests. The tests used a listening base of fifty and examined sixteen mixed male/female sentences from the TIMIT database. The results indicate that the A-by-S technique improves the performance of the 2.4kb/s MPW coder [2]. Further, the coder was preferred to the Federal Standard 1016 coder in some 90% of tests. Listeners comment that the A-by-S algorithm improves the clarity of speech, particularly when perceptual weighting is included in the search.

5. CONCLUSIONS

This paper has discussed a new Analysis-by-Synthesis mechanism for Multi-Prototype Waveform coding. The method exploits the fact that interpolated, accurately extracted, unquantised prototype waveforms produce near-transparent speech quality [1,2]. The algorithm has shown improved performance over previous open-loop schemes [2] and can be used at rates of 2.4kb/s and above. At higher rates, minor modifications lead to the optimisation of other parameters in the closed loop.

6. ACKNOWLEDGEMENTS

This work was supported under grants from ATERB, and the Australian Research Council's Small Grants scheme (1995 and 1996). Pham Duong's period of study at the University of Wollongong was supported by AusAid. The authors would also like to acknowledge Mr Jun Ni for his continuing work on this project.

7. REFERENCES

[1] W.B. Kleijn, J. Haagen, "Waveform Interpolation for Coding and Synthesis", *in Speech Coding and Synthesis*, edited by W.B. Kleijn and K.K. Paliwal, Elsevier 1995.

[2] I.S. Burnett, G. J. Bradley, "New techniques for Multi-Prototype Waveform Coding at 2.84kb/s", *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Detroit, pp. 261-264, 1995.

[3] 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.*, Detroit, pp. 173-176, 1995.

[4] W.B. Kleijn, Y. Shoham, D. Sen, R. Hagen, "A Low Complexity Waveform Interpolation Coder", *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Atlanta, pp. 212-215, 1996.

[5] D.H. Pham, I.S. Burnett, "Quantisation Techniques for Prototype Waveforms", *IEEE International Symposium on Signal Processing and its Applications*, Gold Coast, Australia, 1996.

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

[7] I.M. Trancoso, B.S. Atal, "Efficient Search Procedures for Selecting the Optimum Innovation in Stochastic Coders", *IEEE Trans. On Acoust., Speech and Signal Proc.*, Vol. 38, No. 3, pp. 385-396, March 1990.

[8] I.S. Burnett and R.J. Holbeche, "The Application of the DFT to CELP Architectures," *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Whistler, B.C., Canada, Sept. 1991.

[9] S. Miki, K. Mano, T. Moriya, K. Oguchi, "A Pitch Synchronous Innovation CELP (PSI-CELP) Coder for 2-4kbit/s", *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Adelaide, Australia, pp II-113 - II-116, 1994.