1-10-2002

Spanning the 4 kbps divide using pulse modeled residual

J Lukasiak
University of Wollongong, jl01@ouw.edu.au

I. Burnett
University of Wollongong, ianb@uow.edu.au

Spanning the 4 kbps divide using pulse modeled residual

Abstract

This paper reports a scalable method for coding the LP residual. The scalable method is capable of increasing the accuracy of the reconstructed speech from a parametric representation at low rates to a more accurate waveform matched representation at higher rates. The method entails pitch length segmentation, decomposition into pulsed and noise components and modeling of the pulsed components using a fixed shape pulse model in a closed-loop, analysis by synthesis system.

Publication Details

This article was published as: Lukasiak, J & Burnett, I, Spanning the 4 kbps divide using pulse modeled residual, IEEE Workshop Proceedingson Speech Coding, 6-9 October 2002, 20-22. Copyright IEEE 2002.

SPANNING THE 4 KBPS DIVIDE USING PULSE MODELED RESIDUAL

J. Lukasiak, I.S. Burnett

Whisper Laboratories, TITR

University of Wollongong

Wollongong, NSW, Australia, 2522

1. ABSTRACT

This paper reports a scalable method for coding the LP residual. The scalable method is capable of increasing the accuracy of the reconstructed speech from a parametric representation at low rates to a more accurate waveform matched representation at higher rates. The method entails pitch length segmentation, decomposition into pulsed and noise components and modeling of the pulsed components using a fixed shape pulse model in a closed-loop, Analysis by Synthesis system.

2. INTRODUCTION

Current speech coders exhibit a 'bit-rate barrier' at approximately 4kbps. Below the barrier parametric coders dominate, while above, waveform coders give preferable results. To increase the throughput over variable bit-rate transmission infrastructures such as shared medium networks, it is desirable to design a scalable coder spanning this barrier. As standardised speech compression algorithms are predominantly based on Linear Prediction (LP), developing scalable compression algorithms within this paradigm has been a research focus. Some examples of this research are hybrid parametric/waveform coders that switch at predetermined rates [1] and perfect reconstruction parametric coders that attempt to code the LP residual very accurately [2][6].

The first of these techniques, dynamic switching between waveform and parametric coders, has some serious drawbacks; firstly, oscillatory switching can cause artifacts in the speech and secondly, both extra complexity and storage are required to run two separate algorithms. The second set of techniques require complex mechanisms to modify or warp the pitch track. They have proven to lack robustness and scalability to higher bit rates (particularly within delay constraints).

At high rates, linear predictive coders using waveform matching, produce higher quality speech than parametric coders which directly model (open-loop) the LP residual. The waveform matching is achieved by minimising the error in the speech domain using an Analysis by Synthesis (AbyS) structure such as that used in [3]. At low rates, this 'exact waveform approach' fails to exploit the perceptual redundancy utilised by open loop parametric coders. In particular, low-rate parametric coders will tend to smooth, and reduce the detail of the coded residual. There are thus two contradictory approaches on either side of the artificial bit-rate boundary; precise matching at higher rates versus 'perceptually acceptable parameterization' at low rates. In this paper we proposes a solution to the non scalable characteristics of waveform-matching coders so as to breach the divide.

Our scalable method of LP residual coding is detailed in the following section, with practical results presented in Section 4.

3. METHOD

The key point in our approach is the assumption that we must exploit AbyS modeling at high bit rates and thus it is the scalability of that technique to lower rates that needs to be addressed. However, at low bit rates the quality of speech produced by AbyS based speech coders tends to deteriorate rapidly due to the coder wasting bits modelling perceptually unimportant information [4]. Thus we focus here on a mechanism that avoids this bit wastage by identifying the key elements required in residual representation at low rates. For unvoiced speech, [5] suggests that the signal can be represented in a perceptually transparent manner by replacing the unvoiced LP residual with gain shaped Gaussian noise. Our own results and that work suggest that the low-rate perceptual scalability of speech signals is to be found in the representation of the voiced speech sections. Thus, for high quality low-rate reconstruction of speech signals, we concentrate on the problem of restricting the allocation of AbyS bits such that pitch pulses (and their surrounding details) are adequately represented in synthesised speech.

To ensure that the AbyS modeling at low rates is concerned only with reproducing the pitch pulse, the proposed method firstly critically samples fixed length frames of LP residual (25 ms) into pitch length sub-frames. This segmentation can be achieved in real time using the critical sampling method detailed in [6] or any alternate method that generates non-overlapped pitch length subframes. The non-overlapping/critically sampled nature of the subframes is important as it provides for the use of AbyS modeling. This contrasts with early WI coders that use overlapped (and over-sampled) pitch length subframes.

The extracted pitch length subframes are then decomposed into pulsed and noise components. The decomposition process is analogous to the SEW/REW decomposition performed in WI [7] however, due to the variable number of subframes per frame, fixed length linear filtering (as used in WI) of the subframe evolution requires interpolation of the subframes to produce a fixed number of subframes per frame. An alternative is to use the decomposition method proposed in [8]. This method achieves a scalable decomposition of the subframes into pulsed and noise components using a SVD based approach.

0-7803-7549-1/02/$17.00 ©2002 IEEE.

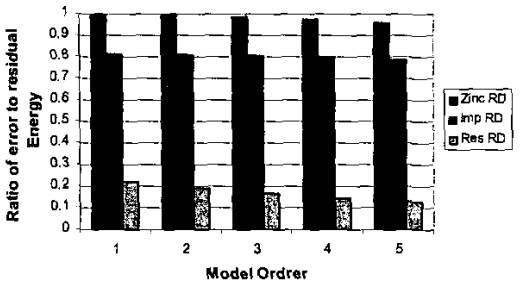

Figure 1: Comparison of residual domain MER

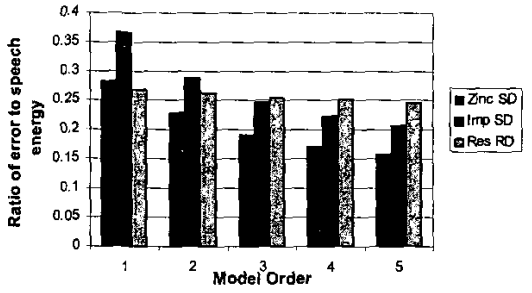

Figure 2: Comparison of speech domain MER

The net result of these operations is that the residual signal is reduced to a parametric representation (i.e. pulse and noise). However, in contrast to traditional parametric coding algorithms where time asynchrony is introduced (such as WI and MELP), the critical sampling of the residual signal maintains time synchrony with the input signal and thus preserves the possibility of using AbyS to model the parameters. If AbyS is now used to model the pulsed component, at low bit rates this operation is concerned only with reproducing a pulse. Further, if a pulse model that naturally represents the shape of the residual pulse (such as a zinc pulse [9]) is used in the AbyS operation, a scalable representation of the residual can be achieved. AbyS coding using a zinc model is detailed in [9], but the basis used in our work involves representing each pitch length pulsed component by minimising:

$$\begin{aligned} e(n) &= X(n) - Z(n) \\ &= X(n) - \sum_{i=1}^{P} z_i(n) * h(n) \end{aligned} \quad (1)$$

where $h(n)$ is the impulse response of the LP synthesis filter, $X(n)$ is the input pulsed component in the speech domain, $Z(n)$ is the representation of the pulsed component in the speech domain, $z(n)$ is a zinc pulse and P is the order of the zinc model (number of pulses).

4. PRACTICAL RESULTS

This section concentrates on the scalable representation of the pulsed component of the pitch length subframes, and depends on the technique proposed in [5] for representation of the noise component as gain shaped Gaussian noise. Our reference point is residual synthesized from a limited direct PCM coding of each residual pulsed sub-frame (using a limited set of samples centred on the residual domain pulse); we refer to this approach as 'Direct Modeling' as it simulates direct representation of the residual domain signal with varying degrees of accuracy. We then compare the error of such an approach with AbyS modelling of the pulsed sub-frames using both impulse and zinc [9] pulse models. We performed the comparisons on a cross-section of sentences from the TIMIT database.

For each of the pulse models used in AbyS, the analysis order was varied, and in the Direct modeling, for comparison, the number of adjacent positions transmitted was altered. For each modeling approach the Mean Error Ratio (MER), defined as the ratio of MSE to mean input energy for each pitch length sub frame was calculated according to:

$$MER = \left(\frac{1}{N} \sum_{x=0}^{N-1} \left(Input(x) - Estimate(x) \right)^2 \right) \Big/ \left(\frac{1}{N} \sum_{x=0}^{N-1} Input(x)^2 \right) \quad (2)$$

where N is the number of samples in the sub frame. The MER was computed for both the residual and speech waveforms and the resultant MERs for each model averaged for all sentences. Figures 1 and 2 show residual and speech domain MER results respectively.

The model orders in Figures 1 and 2, represent the number of pulses per sub-frame for the zinc and impulse methods and, for direct residual modeling, the number of transmitted samples centred around the residual pulse according to the following key:

Order	Transmitted Samples
1	7 (pulse centred)
2	9
3	11
4	13
5	15

These sample numbers were chosen such that an order of 1 indicates three samples on each side of the pulse, order 2 four samples etc. They provide a comparable waveform-matching reference point for the pulsed models. Comparing Figures 1 and 2 it is evident that, for pulsed models (as with waveform matching), minimizing the MSE in the residual domain is not analogous to minimizing the MSE in the speech domain. In fact, the pulse models consistently reduce the speech domain error as the order of the model is increased, whilst the residual domain error for the same pulse models remains almost constant. For direct modelling of the residual the opposite is true. The residual domain error (which is quite small even for the lowest model order - indicating that the method is capturing the majority of the residual domain pulse) is consistently reduced as the model order is increased, however, a corresponding reduction in the speech domain error is not achieved. Moreover, for some individual sentences, increasing the order of the direct residual modelling achieved a reduction in the residual domain MER but resulted in a worsening in the speech domain error. This never occurred in our test set for the pulse models minimized in the speech domain; increasing the model order always reduced the overall speech domain error results.

Comparing the error values for the different methods in Figure 2 shows that zinc and impulse models using 2 and 3 pulses per sub-frame respectively, achieved a lower error value than the highest order of direct modelling which uses 15 adjacent pulses.

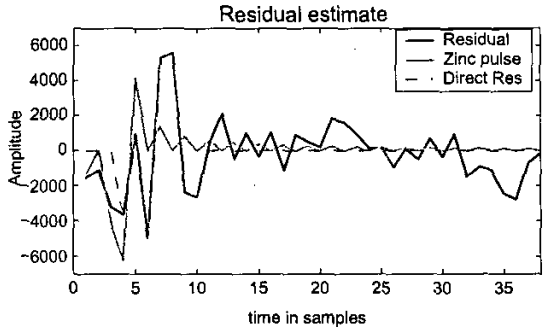

Figure 3: Residual domain pulse Comparison

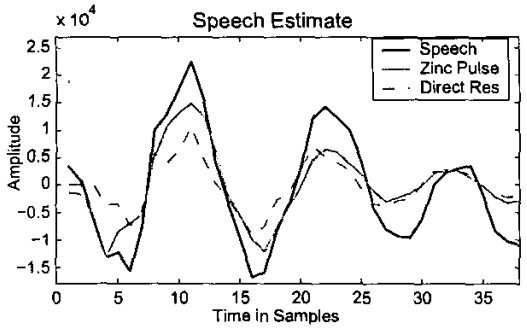

Figure 4: Speech domain pulse comparison

Figure 2 also indicates that the zinc pulse model using only a single pulse per sub frame almost matched the error achieved using 7 adjacent pulses for direct modelling.

The results in Figure 2 show a clear scalability with order, in terms of error minimisation for the pulse models calculated in the speech domain. However, at low rates it is the parametric representation of the pulse shape (and hence the perceptually important smoothness etc) that is perceptually important. Figures 3 and 4 compare residual and speech domain waveform modelling using both a single zinc pulse and direct residual modelling of 7 adjacent samples.

Figure 4 indicates that a better representation of the speech pulse shape is achieved by the zinc pulse model. This is in spite of there being only a single pulse used in the model. Further, this suggests that, even in a parametric sense (where the MSE is less relevant), pulse modelling of the pitch length segments by minimising the error in the speech domain produces a very good reproduction of the pulse shape. To investigate this further, a single zinc parameter per 25 ms frame was quantised using 10 bits and interpolated for each pitch length sub-frame. The position of the pulse in each sub frame was fixed. This amounts to a 400 bps representation of the voiced speech. Informal listening tests indicated that the synthesized speech sounded clear and natural.

Figure 3 gives a useful insight into the fact that minimising the error in the speech domain using fixed order pulse models does not necessarily minimise the residual domain energy. The zinc pulse in Figure 3 is positioned before the main residual pulse and thus has a large MSE in that domain. In contrast, the zinc speech domain pulse in Figure 4 is a good approximation of the original waveform.

The results indicate that using pitch length sub-frames and pulse models with parameters calculated in a closed loop AbyS system, generates a scalable method for reproducing voiced speech. This contrasts with attempting to achieve scalability through increasing the accuracy of residual domain modeling; a process that may, in practice, offer very little improvement in the speech representation.

5. CONCLUSION

The results indicate that employing parametric pulse models in a AbyS structure, which is restricted to modeling pulsed, pitch length subframes does provide scalability across the artificial 'bit-rate' divide between parametric and waveform coders. The scalability of the representation is achieved by varying the order of the pulse model used (the number of pulses per subframe) in synthesizing the pulsed subframes. We suggest that these results call into question the approach, advocated in [2] and [6], of deriving scalability from pushing parametric coding techniques (such as WI) to higher rates. Instead, we propose that adaptation of higher-rate AbyS algorithms to the use of pulse model parameter optimization and then scaling the quantization of those models is more appropriate. However, while the modeling approaches may differ, it is worth noting that the pitch-synchronous, critical sampling approach of techniques intended to span the 'bit-rate' divide is a common factor.

6. REFERENCES

[1] J. Stachurski and A. McCree, "A 4 kb/s hybrid MELP/CELP coder with alignment phase encoding and zero-phase equalization", Proc. of ICASSP 2000, Vol.3, pp.1379 -1382, 2000.

[2] T. Eriksson and W.B. Kleijn, " On waveform-interpolation coding with asymptotically perfect reconstruction", Proc. of IEEE Workshop on Speech Coding, pp. 93 -95, 1999.

[3] B.S. Atal, "Predictive coding of speech at low bit rates", IEEE Trans. On Comm., vol. COM-30, pp.600-614, April 1982.

[4] J. Thyssen, G. Yang, et al., "A candidate for the IUT-T 4KBIT/S speech coding standard", Processings of IEEE International Conference on Acoustics, Speech, and Signal Proc., Vol.2, pp.681-684, 2001.

[5] G. Kubin, B.S. Atal and W.B. Kleijn, "Performance of noise excitation for unvoiced speech", Proc. of IEEE w/shop on Speech Coding for Telecommunications, pp.35-36, 1993.

[6] N.R. Chong-White, *Novel Analysis, Decomposition and Reconstruction Techniques for Waveform Interpolation Speech Coding*, Phd. Thesis, University of Wollongong, 2000.

[7] W.B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms", Proc of IEEE Conf. On Acoustics, speech and signal processing, Vol. 1, pp.508-511, 1995.

[8] J. Lukasiak and I.S. Burnett "Low Delay Scalable Decomposition of speech waveforms", Proc. of the 6th International Sym on Digital signal Processing for Communications DSPDC 2002, pp. 12-15, January 2002.

[9] R.A. Sukkar, J.L. LoCicero and J.W. Picone, "Decomposition of the LPC excitation using the zinc basis functions", IEEE trans on Signal Processing, Vol.379, pp. 1329-1341, Sept. 1989.