

2003

Scalable speech coding spanning the 4 Kbps divide

J. Lukasiak
University of Wollongong

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

Publication Details

This article was published as: Lukasiak, J & Burnett, I, Scalable speech coding spanning the 4 Kbps divide, Proceedings Seventh International Symposium on Signal Processing and Its Applications, 1-4 July 2003, vol 1, 397-400. Copyright IEEE 2003.

Scalable speech coding spanning the 4 Kbps divide

Abstract

This paper examines 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. Subjective testing is presented that indicates that in addition to the AbyS modeling, the pulse parameter evolution must be constrained in synthesis. Results indicate that this proposed method is capable of producing perceptually scalable speech quality as the bit rate is increased through 4 kbps.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was published as: Lukiasiak, J & Burnett, I, Scalable speech coding spanning the 4 Kbps divide, Proceedings Seventh International Symposium on Signal Processing and Its Applications, 1-4 July 2003, vol 1, 397-400. Copyright IEEE 2003.

SCALABLE SPEECH CODING SPANNING THE 4 KBPS DIVIDE

J. Lukasiak, I.S. Burnett

Whisper Laboratories, TITR

University of Wollongong

Wollongong, NSW, Australia, 2522

ABSTRACT

This paper examines 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. Subjective testing is presented that indicates that in addition to the AbyS modeling, the pulse parameter evolution must be constrained in synthesis. Results indicate that this proposed method is capable of producing perceptually scalable speech quality as the bit rate is increased through 4 kbps.

1. 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 propose a solution to the non scalable characteristics of LP based coders so as to breach this divide.

Our initial scalable method of LP residual coding is detailed in the following section. Practical results characterizing this method are presented in Section 3. Section 4 details subjective analysis of the proposed method and modifications that are necessary to provide good subjective performance. The major findings are summarized in Section 5

2. METHOD

The key point in our approach is the assumption that a single scalable algorithm capable of bridging 4 kbps must provide a parametric representation at low rates and smoothly migrate to AbyS modeling at high bit rates. As the objective is to achieve AbyS modeling at high rates, our approach identifies that 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 and also limits the look ahead required for the decomposition method.

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).

3. PRACTICAL RESULTS FOR PULSED SUB-FRAMES

This section concentrates on the scalable representation of the pulsed component of the pitch length sub-frames. Our

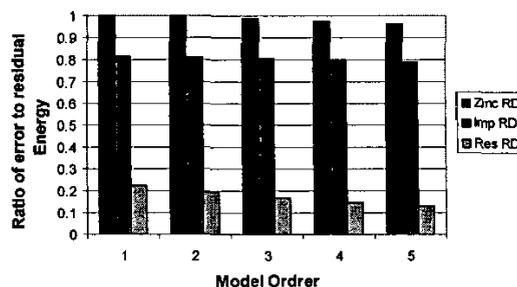


Figure 1: Comparison of residual domain MER

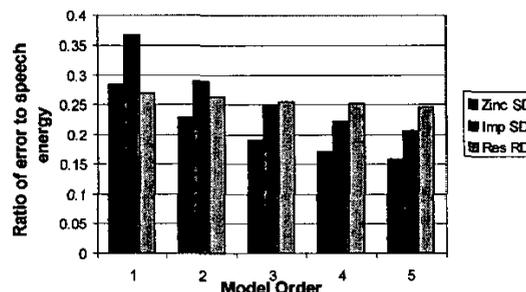


Figure 2: Comparison of speech domain MER

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:

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

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 (Res in Figures 1 & 2),

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 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.

4. SUBJECTIVE RESULTS FOR ENTIRE SCALABLE CODER

The results presented in Section 3 give a useful insight into the scalability of the proposed method in a largely objective sense. However, when incorporated into an entire coding structure and tested subjectively, it was found that the high-rate representation generated using multiple pulses per sub-frame had a noisy and harsh feel. This was in opposition to a low rate representation that used only a single pulse per sub frame, the magnitude of which was generated from linearly interpolating a single magnitude per frame (a parametric representation), and sounded

smooth and full. The cause of this noisy feel at high rates was found to be due to the change between adjacent pitch pulse shapes being unconstrained in synthesis.

The noisy effect was apparent despite the fact that the pulse parameters had been calculated in a closed loop AbyS method, and the quantization scheme for the parameters was achieving a SNR between the original and synthesized pulsed components in excess of 9 dB. This result is in direct conflict with conventional multi-pulse CELP waveform modelling techniques [3,9], which use fixed size sub-frames. In these coders increasing the number of pulses used per sub-frame and hence increasing the SNR, increases the subjective quality of the synthesised speech.

Kleijn [10] reported the problem of constraining the pitch pulse evolution in a parametric WI coder (that makes no attempt to minimise the perceptually weighted speech domain error), where the accuracy of the reconstructed speech was sacrificed in order to constrain the rate of change of the pitch pulses. This had the effect of improving subjective quality. However, constraining the pulsed component amplitude evolution is not appropriate for our high rate representation, as this would reduce the ability to represent quickly changing or transient sections of speech. It was determined that for our proposed scalable coder the best subjective results could be achieved by constraining only the individual pulse positions within each synthesised sub-frame to a restricted set of positions. Full details of this constraint can be found in [11].

Despite having to constrain the pulse evolution in synthesis, the high rate method still converges to high perceptual quality synthesised speech. This occurs because the analysis loop still operates in an AbyS structure and captures the perceptually important parameters of quickly changing sections of the input speech in the pulsed parameters. Having this very accurate paramatisation available allows the coder to produce high perceptual speech quality, even in quickly changing sections. This contrasts with purely parametric coding structures such as WI, which smear the quickly changing transitional sections in the analysis stage, and as such these sections cannot be reproduced in synthesis regardless of the bit rate available for transmission.

A consequence of constraining the synthesis pulse shapes is that for accurate high rate reconstruction extra bits have to be used better representing the noise sub-frame component. These extra bits are required to modulate the temporal envelope of the original speech back onto the synthesised noise sub-frames.

Taking the stated modifications to the method proposed in Section 3 into consideration, an entire scalable speech coding structure was generated. A detailed description of this coder can be found in [11]. This coder had the added

constraint that the overall algorithmic delay had to be comparable to standardised coders at rates above 4 kbps. This resulted in a coder that uses no look ahead beyond the current frame, with a total algorithmic delay of 30 ms. The bit allocation for the coder parameters operating at 2.4 kbps and 6 kbps are shown in tables 1 and 2 respectively. The frame size for the coder is 200 samples or 25 ms.

Parameter	LSF	Pitch	Pulsed sub-frames	Noise sub-frames	Total
Bits/frame	30	7	13	10	60

Table 1: Bit allocation for scalable coder at 2.4 kbps

No. sub frames	LSF	Pitch	Pulsed sub-frames	Noise sub-frames	Total
1	30	20	12	88	150
2	30	20	24	76	150
3	30	18	36	64	148
4	30	18	48	54	150
5	30	16	60	41	147
6	30	16	66	34	146
7	30	16	66	34	146
8	30	14	78	26	148
9	30	14	78	26	148
>=10	30	14	78	26	148

Table 2: Bit allocation for scalable coder at 6 kbps

It should be noted that the bit allocation for the 6 kbps scalable coder is dependent on the number of pitch length sub-frames/frame. As this places significant emphasis on correct reception of this parameter (it is included in the pitch parameter in Table 2), the spare bits available when the number of sub-frames is greater than 5 are used to protect this parameter.

Mean Opinion Score (MOS) testing for the scalable coder configurations shown in tables 1 and 2 were conducted using 25 listeners each. The MOS test also included standardized coders operating at comparable rates. The results of the testing are shown in Tables 3 and 4 respectively.

Coder	MOS	95% Conf
Scalable	3.01	.095
LPC10	2.44	.085
MELP	3.2	.1

Table 3: 2.4 kbps MOS test Results

Coder	MOS	95% Conf
Scalable	3.45	.07
FS1016	3.28	.08
G729	3.95	.09

Table 4: 6 kbps MOS test Results

The results in Tables 3 and 4 indicate that the subjective quality of the scalable coder clearly scales with an increase

in bit rate. This is despite the fact that 4 kbps has been spanned. The results also indicate that at each rate, the performance is comparable to fixed rate standardized coders operating at similar rates. This is a particularly encouraging result considering the fact that the scalable coder has been restricted to use no look ahead in the coding structure. If added delay can be tolerated it is felt that the subjective quality of the scalable coder could be significantly improved.

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. However, opposed to traditional multi-pulse AbyS techniques, employing AbyS in this structure requires the synthesized pulse evolution to be constrained. This constraint is required to produce high perceptual quality. Despite adding this constraint to the synthesis, the proposed method still converges to a very accurate representation at high rates and subjective results indicate that perceptual scalability is produced as the 4 kbps bit rate barrier is bridged.

6. REFERENCES

- [1] J. Stachurski and A. McCree, iA 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, iOn waveform-interpolation coding with asymptotically perfect reconstruction, Proc. of IEEE Workshop on Speech Coding, pp. 93-95, 1999.
- [3] B.S. Atal, iPredictive 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., iA candidate for the IUT-T 4KBIT/S speech coding standard, Proceedings 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, iPerformance 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, iA 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 iLow 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, iDecomposition of the LPC excitation using the zinc basis functions, IEEE trans on Signal Processing, Vol.379, pp. 1329-1341, Sept. 1989.
- [10] W.B. Kleijn, iEncoding speech using prototype waveforms, IEEE Trans. On speech and Audio Proc., Vol. 1, No.4, pp. 386-399, Oct. 1993.
- [11] J. Lukasiak, Techniques for low-rate Scalable Compression of Speech Signals, PhD. Thesis, University of Wollongong, 2002.