17-9-2000

Very low rate speech coding using temporal decomposition and waveform interpolation

C. H. Ritz
University of Wollongong, critz@uow.edu.au

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

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

Very low rate speech coding using temporal decomposition and waveform interpolation

Abstract

In very low rate coding the aim is to accurately represent speech characteristics as efficiently as possible. High coding gains for the spectral features can be achieved through the use of temporal decomposition. Waveform interpolation coders accurately represent the excitation using characteristic waveforms (CWs) extracted at a constant rate. In this paper, the two approaches are combined into a very low rate coder operating at around 1 kbps. It is shown that the evolution of the excitation is related to the evolution of the speech spectrum. To minimise bit rates, the transmission of CWs is adapted to the spectral parameter evolution using the parameters derived from temporal decomposition of the spectral parameters.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper originally appeared as: Ritz, CH, Burnett, I & Lukasiak, J, Very low rate speech coding using temporal decomposition and waveform interpolation, Proceedings. 2000 IEEE Workshop on Speech Coding, 17-20 September 2000, 29-31.

VERY LOW RATE SPEECH CODING USING TEMPORAL DECOMPOSITION AND WAVEFORM INTERPOLATION

C.H. Ritz, I.S. Burnett, and J. Lukasiak

Whisper Laboratories, TITR,
University of Wollongong, NSW, Australia

ABSTRACT

In very low rate coding the aim is to accurately represent speech characteristics as efficiently as possible. High coding gains for the spectral features can be achieved through the use of Temporal Decomposition. Waveform Interpolation coders accurately represent the excitation using Characteristic Waveforms (CWs) extracted at a constant rate. In this paper, the two approaches are combined into a very low rate coder operating at around 1 kbps. It is shown that the evolution of the excitation is related to the evolution of the speech spectrum. To minimise bit rates, the transmission of CWs is adapted to the spectral parameter evolution using the parameters derived from Temporal Decomposition of the spectral parameters.

1. INTRODUCTION

Temporal Decomposition [1] is a method for achieving a significant reduction in the information required to represent the spectral characteristics of speech. It derives so called event functions and target vectors, which will be referred to as events and targets, to model the evolution of the spectral parameters. These are then used to reconstruct the spectrum through smooth interpolations.

Smooth interpolation is also used in Waveform Interpolation (WI) [2]. The evolution of the excitation is modeled by characteristic waveforms (CWs), which represent pitch cycles of the excitation waveform, and are extracted at a constant rate. Reconstruction of the excitation waveform proceeds by smoothly interpolating between the CWs. Hence the two forms of modeling are well suited to each other.

To make WI applicable to very low rate coding the bit rate required to represent the CWs must be minimised. Standard WI transmits quantised CWs, known as prototype waveforms, at a constant rate, with typical bit rates around 600 bps [2]. However, the rate of evolution of the excitation is non-constant, and so transmitting the prototypes at a constant rate is inefficient. Observations of the evolution have shown that it is related to the characteristics of the event functions found via temporal decomposition. Here it is proposed to adapt the transmission rate of the prototypes to the event rate. Hence a variable rate WI scheme using temporal decomposition is created.

The next section describes the temporal decomposition of the spectral parameters. Section 3 describes variable rate WI including prototype extraction and quantisation while pitch and gain quantisation techniques are described in Section 4. A summary of results, including bit rates for all parameters is presented in Section 5.

2. TEMPORAL DECOMPOSITION

As described in the introduction, Temporal Decomposition derives events and targets to model the speech spectral parameters, in this paper the Line Spectral Frequencies (LSFs). The method can be described as a weighted sum of event functions (1), where the weights are the target vectors.

$$\hat{y}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \qquad 1 \le n \le N,\ 1 \le i \le p, \tag{1}$$

Here, $\hat{y}_i(n)$ is the approximation of the ith LSF produced by the model, $\phi_k(n)$ is the kth event function, a_{ik} is the kth event function corresponding to LSF i, N is the length of the segment, p is the number of LSFs per frame and m is the number of event functions found for the segment.

The target vectors are modified versions of the LSF vectors located at the stable sections of their trajectories. Hence, the method can be thought of as a form of interpolation of the LSF vectors between stable points.

2.1 Derivation of events and targets

For derivation of events and targets, the restricted temporal decomposition approach [3] was chosen. The LSF vector trajectories are analysed to find event function centres, corresponding to stable points. Between consecutive event centres, events and targets are derived by minimising the mean squared error between the original LSFs and those reconstructed using expression (1). This approach avoids both the use of Singular Value Decomposition (SVD) and an exhaustive search for events as well as the costly process of an iterative refinement technique, commonly used in other approaches [1,4].

2.2 Quantisation of events and targets

The amplitude of the event functions are restricted to be within the range zero to one, and no more than two events can overlap at any one time, as described in [3]. Hence they can easily be described a vector representing the shape between consecutive event centres. In this paper a ninth order vector is used and the event shapes are quantised using vector quantisation. Since the separation of event centres varies, the width of each event is also transmitted.

The target vectors are modified versions of the original LSF vectors located at the event centres. Hence, they can be quantised using similar techniques to those used for quantising LSFs. Here, the target vectors are quantised using multistage vector quantisation.

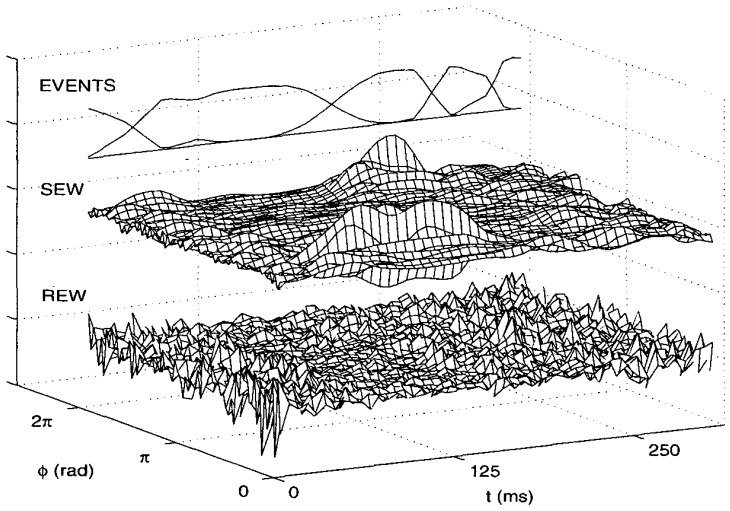

Figure 1. Event functions along with the SEW and REW surfaces for a section of speech. Note how areas of high SEW energy correspond to the two longer duration event functions, while areas of low SEW energy and high REW energy correspond to shorter duration events.

3. VARIABLE RATE WI

In standard WI, each characteristic waveform is extracted at a constant rate and modeled by the decomposition of the prototype waveform into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW). The SEW is obtained by low pass filtering the evolution of the prototype at an appropriate frequency (in this case 20 Hz) to remove the low frequency characteristics of the excitation. The REW is obtained by effectively subtracting the SEW from the original prototype. By normalising their lengths (in the DFT domain), successive SEWs and REWs can be plotted as a function of time to provide the description of the characteristic waveform surface shown in Figure 1. Also shown are the event functions derived for this section of speech.

Areas of high SEW energy correspond to longer duration events, where the SEW waveform is quite stable. Conversely, events of shorter duration correspond to higher REW energy where the SEW is less stable. Hence, prototype information should be sent less often for longer duration events and more often for shorter duration events.

One implementation of this idea is to send a fixed number of prototypes to represent each event, where the prototypes are spaced equally in time. Since event durations vary, prototype extraction rates will then vary and thus a variable rate WI scheme is created. In the decoder, linear interpolation of the prototypes within each event is used to reconstruct the excitation, as done in standard WI.

3.1 REW Quantisation

Two approaches were used to quantise the REW. In both cases, only the magnitude was quantised and the phase replaced by a random phase in the decoder (since phase information has previously be found to be perceptually insignificant [2]) for the first approach we vector quantised the shape of the magnitude spectrum, using eight vectors of polynomial coefficients. In the second approach, the average magnitude of the REW is quantised using eight levels. The magnitude spectrum was reconstructed in the decoder using a random value from a gaussian codebook, which is scaled to the quantised average magnitude.

3.2 SEW Quantisation

Two approaches were used to quantise the SEW. Again, only the magnitude was quantised in each case, with the phase being set to zero in the decoder. In the first approach, the magnitude below 800 Hz is vector quantised and the remainder of the SEW is calculated in the decoder from the REW. In the second approach, no SEW magnitude is transmitted. Instead, the magnitude is replaced by a pulse model, which is inferred in the decoder from the REW magnitude. The second approach requires no bits.

4. PITCH AND GAIN QUANTISATION

Both variable and fixed transmission rates for the pitch and gain information were implemented. Variable transmission sends a fixed number of pitch and gain values for each event. When transmitting less pitches and gains than prototypes, the intermediate values were interpolated. In fixed transmission, pitches and gains were sent every 40 Hz.

In the current implementation, all pitch values can be represented using 7 bits. A non-uniform 6-bit quantiser was also tested for the pitch. In this approach, shorter pitch values were given more weight than longer pitch values, as the quantisation error becomes more significant with decreasing pitch. The gain was quantised using a 5-bit differential quantiser.

5. RESULTS

The coder described in this paper performs a 10^{th} order Linear Prediction (LP) analysis on 25 ms frames of speech sampled at 8 kHz with a 5 ms overlap. The resulting LPC's are converted to LSFs, which are linearly interpolated to two LSFs per frame. A summary of bit rates for the preferred coder is provided in Table 1, with average bit rates between 980-1088 bps achieved.

Informal listening tests found the coder to produce natural sounding speech of good intelligibility, which maintained speaker characteristics. Naturalness is noticeably absent in other coders at these rates.

5.1 Event and Target Quantisation

For quantisation of the event function vectors, both 5 and 6-bit codebooks were trained using event functions derived from around 30 minutes of male and female speech. For vector quantisation of the target vectors, this speech was also used in training a 20-bit multistage codebook, with 2 10-bit stages. A 24 bit multistage codebook with two 12-bit stages was also trained. Event widths required 4 bits for accurate representation.

Informal listening tests found no significant difference in the reconstructed speech when using 5-bits or 6-bits for event vector quantisation. Informal listening tests preferred 24-bit vector quantisation of the targets to 20-bit vector quantisation. Hence, the preferred coder used 5 bits for event shape quantisation and 24 bits for target quantisation.

5.2 Prototype Quantisation

The final implementation of the coder reconstructed speech by interpolating between ten prototypes (reconstructed from SEW and REW surfaces) for each event.

The lower 800 Hz of the SEW magnitude was quantised using both 7 bits and 6 bits, once per event. Informal listening tests found that 6-bit quantisation did not produce a significant degradation compared with 7-bit quantisation. However, it was found that using a pulse model to represent the SEW magnitude, requiring no bits for transmission, produced output speech of higher quality to that reconstructed using 7-bit quantised SEW magnitudes. When using the pulse scheme, a separate SEW model was found for each REW before reconstructing the ten prototypes for the event.

As previously mentioned, the REW magnitude was quantised using 3 bits. Informal listening tests found that using a random value for the REW magnitude samples and scaling to the quantised REW gain was preferred over using a codebook of magnitude shapes and so was adopted in the final implementation. To minimise the bit rate, REW gain information was sent once per event. To recreate the remaining 9 REWs, their magnitude was linearly interpolated from the transmitted REW, while their phase was set to a random value.

5.3 Pitch and Gain Quantisation

When using fixed transmission of pitch information, informal listening tests found that non-uniform 6-bit pitch quantisation resulted in no significant degradations in the reconstructed speech compared with 7-bit pitch quantisation. For comparison purposes, pitch and gain information was sent twice per event when using variable transmission to achieve a similar bit rate to fixed transmission. However, variable transmission of the pitch was found to cause inaccuracies in the interpolated pitch track and produce distortions in the reconstructed speech during informal listening tests.

Parameter	Bit allocation	Rate (Hz)
pitch	6	40
gain	5	40
event function	9	15-18
target vector	24	15-18
SEW	0	15-18
REW	3	15-18

Table 1. Bit allocations for the proposed coder.

Analysis of the true pitch track showed that it often has significant frame-to-frame variations especially during events located in unvoiced regions. Since variable transmission results in interpolation between points separated by more than one frame, it leads to pitch track inaccuracies in these regions. Fixed transmission does not suffer from this problem (providing overall transmission rate is sufficient), and hence is preferred in the final implementation. Variable gain transmission also produces distortions due to similar reasons as those for variable rate pitch transmission. Hence, fixed transmission of the gain information is also preferred.

6. CONCLUSION

A very low bit rate coder with good intelligibility operating at around 1kbps has been described. The coder exploits the commonalities between WI and Temporal Decomposition and results in a variable rate WI implementation. The current implementation transmits pitch and gain information at a fixed rate of 40 Hz, while the SEW/REW surfaces are transmitted at variable rates reflecting their evolutionary behaviour. The pitch and gain transmission occupies a significant proportion of the total number of bits in the coder. Hence, future implementations will attempt to reduce the bits rate of these parameters.

7. ACKNOWLEDGEMENTS

C.H. Ritz is in receipt of an Australian Postgraduate Award and a Motorola (Australia) Partnerships in Research Grant.

8. REFERENCES

[1] Atal, B. S., "Efficient coding of LPC parameters by Temporal Decomposition", *Proc. ICASSP'83*, pp. 81-84, Boston, 1983.

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

[3] Kim, S. J. and Oh, Y.H., "Efficient quantisation method for LSF parameters based on restricted temporal decomposition", *Electronics Letters*, vol. 35, issue 12, pp. 962-964, 1999.

[4] Van Dijk-Kappers, A. M. L. and Marcus, S. S., "Temporal decomposition of speech", *Speech Communications*, vol. 8, no. 2, pp. 125-135, 1989.