WIDEBAND SPEECH CODING USING MULTIPLE CODEBOOKS AND GLOTTTAL PULSES

C. McElroy, B.P. Murray, A.D. Fagan

DSP Research Group,
Dept. of Electronic and Electrical Engineering,
University College, Dublin,
Ireland.
Tel. +353-1-7061964, Fax. +353-1-2830921
E-Mail : ciaranmc@maxwell.ucd.ie

ABSTRACT

We propose a coder that achieves near transparent wideband speech coding by parameterising the prediction residual through the use of multiple codebooks and synthetic glottal pulses coupled with adaptive bit allocation. The use of synthetic glottal pulses improves the performance of the coder compared to a previous coder using a single impulse [1] without increasing the bit rate. This multiple codebook approach results in a coder operating at 16 kb/s and 24 kb/s that provides comparable speech quality to the CCITT G.722 coder operating at 64 kb/s.

1. INTRODUCTION

With the recent introduction of ISDN and the reality of video teleconferencing and videophones comes a demand for more natural sounding speech than that traditionally supplied by the narrowband telephone network. Increasing the speech bandwidth from 3.4 to 7 kHz dramatically improves the naturalness, presence and intelligibility of the speech. This is particularly important in video conferencing applications were it has been shown that increasing the audio quality produces a greater perceived improvement in the overall system. The current standard for wideband speech coding is the G.722 coder. This uses subband ADPCM and operates at 48, 56 and 64 kb/s. When operating at 64 kb/s G.722 gives transparent speech coding for most speakers. However the bit rate is still too high for widespread practical use. A coder operating at 16 kb/s while maintaining the high quality of G.722 is needed.

The CELP algorithm [2] has achieved great success in coding narrow band speech (0-4 kHz) and is more recently being used in wideband applications also [3]. This coder models the speech production process in two stages, namely a vocal tract model and its excitation. The excitation consists of the weighted sum of a Gaussian sequence and previous excitations which, while it works well, is not sufficient to provide transparent coding.

In this paper we propose a coder that uses multiple codebooks and synthetic glottal pulses, coupled with dynamic bit allocation to achieve near transparent coding at 24 and 16 kb/s.

2. IMPROVING THE EXCITATION MODEL

Conventional CELP coders use a combination of a weighted white noise sequence and previous excitations as the excitation for the synthesis filter. This model is not suited to accurately modelling unvoiced to voiced transitions [4] and so a distinctive warble is introduced into the speech.

A considerable amount of work has been carried out in low bit rate narrowband speech coding towards solving the onset problem [4]. This work indicates that the problem is due to the adaptive codebook of past excitations being incapable of introducing the required pitch pulses at the onset. The adaptive codebook works well in steady state voiced speech, however it is not suited to coding transitions. If the adaptive codebook misses the pitch pulse then it must be modelled by the fixed codebook. Since this codebook is stochastic it cannot do this without distorting the excitation around the pulse. If a single impulse is used in combination with the stochastic codebook then this problem is reduced [1]. However, from looking at prediction residuals it is clear that glottal pulses are not single impulses but contain several impulses. This suggests that the performance of the coder could be improved further if more than one impulse could be used.

3. DESIGN OF GLOTTAL PULSE CODEBOOK

The exact number of non-zero samples in a glottal impulse is determined from observing how a multi-pulse coder deals with onsets. It appears that the onsets tend to be coded using just two pulses. Therefore the

0-7803-2431-5/95 $4.00

synthetic glottal pulses used in this work consist of just two impulses, see Figure 1. The number of samples between the impulses is allowed to vary from 0 to 4. The relative amplitudes of the second pulse with respect to the first are determined by allowing them to vary in the range -0.5 to -2.95 in steps of -0.05, a single impulse is also included. This large number of pulses is then used in a coder and the pulses actually used are recorded. The optimum pulses can then be determined using the LBG algorithm for designing vector quantisers [5], the vectors being the number of zeros between the non-zero samples and the relative amplitude of the second pulse with respect to the first.

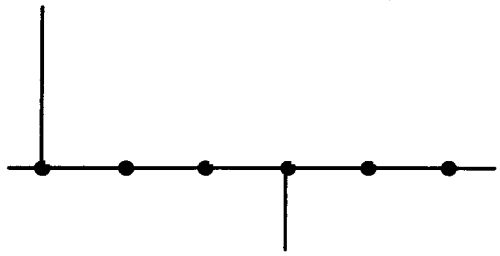

Figure 1 : The synthetic glottal pulse.

Figure 2 (a) shows the actual prediction residual of an onset. The first two pitch pulses are shown. Figure 2 (b) shows the residual when coded by a conventional CELP coder and Figure 2 (c) shows it when coded using a white codebook and glottal pulses. The codebook using the pulses is obviously better at capturing the perceptually important pulses.

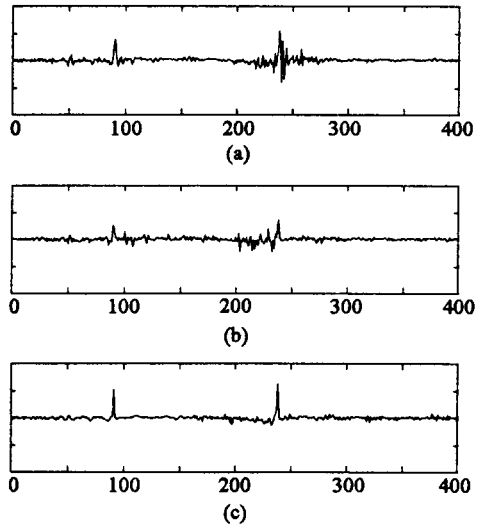

Figure 2 : Excitation signals, (a) the original prediction residual, (b) the excitation using a white codebook and pitch prediction, (c) the excitation when using a white codebook, pulses and pitch prediction

4. THE MULTIPLE CODEBOOK CODER

The white nature of the excitation fails to take account of the relative perceptual importance of certain frequency ranges. Only 10-20 % of the information contained in a speech signal is contained in the frequencies above 4 kHz. When a white codebook is used all frequencies are given equal importance. This is inefficient, especially with a wideband coder. The situation can be improved by using bandlimited codebooks to provide greater emphasis on the perceptually important frequencies [1].

The coders tested in this work use six codebooks, a glottal pulse codebook, a stochastic codebook and four bandlimited codebooks with the following bandwidths 0-1 kHz, 1-2.5 kHz, 2.5-5 kHz and 5-8 kHz. The bandwidths were chosen to increase roughly logarithmically with frequency.

The excitation is the sum of all the selected codewords so the presence or absence of certain codewords is unimportant. Likewise the order in which they were searched is irrelevant to the synthesiser. However, the order in which the codebooks are searched does affect the resulting speech quality. For example, when an onset is present the glottal pulse is important and the white codebook is less so but during a fricative the white codebook is more important than the pulse. This means that the order in which the codebooks are searched must change with the characteristics of the speech. In the coder suggested here all the codebooks are searched to find the codeword which will produce the highest SNR. This codebook is then excluded and the rest of the codebooks are searched, allowing for the effect of the previously selected codeword.

Using many codebooks to parameterise the excitation results in high quality speech but at the expense of a high bit rate. The bit rate can be reduced by only using the codebooks that yield the greatest improvement in the SNR. This is easily done since the dynamic ordering has already assessed the effect of each codebook on the SNR. Since only a subset of the codebooks are used in any given subframe, side information indicating which codebooks are used has to be include. This adaptive bit allocation does not affect the speech quality significantly but it dramatically reduces the bit rate.

The pitch codebook is not included in the above scheme as this was found to reduce speech quality, instead the pitch search is carried out first and is always included in the excitation coder. This is most likely due to the pitch codebook improving the periodicity of the speech even if it does not always produce the highest SNR.

Periodicity in voiced speech is critical to the perceived speech quality [6].

Once the excitation vectors have been determined their gains can be reoptimised. This produces a slight increase in the SNR without changing the bit rate.

Of the six available codebooks only three are used at any one time, requiring 5 bits to describe the allocation. The glottal pulse is chosen first in approximately 12 % of subframes as opposed to 2.5 % for the single impulse used in a previous coder [1]. Also the glottal pulse is used at some stage in 48 % of subframes compared to only 12 % of subframes for the single impulse.

The fixed codebooks (including the codebook of glottal pulses) contain 2048 codewords and the gains are quantised to 5 bits, implying 16 bits per codeword. The pitch codebook contains 256 codewords and so uses a total of 13 bits. Therefore each subframe requires 66 bits. The minimum allowed pitch delay is equal to the subframe dimension. When short pitch delays were allowed there was virtually no improvement in the resulting speech quality. This is probably due the improved modelling by the fixed codebooks reducing the importance of the adaptive codebook. A twentieth order predictor is used and its coefficients are updated once per frame, scalar quantised to 70 bits. A frame length of 25 ms was used and the number of subframes per frame dictated the final bit rate. The 24 kb/s coder used 8 subframes per frame and the 16 kb/s coder used 5 subframes per frame.

5. RESULTS

Informal subjective tests were carried out on the above multi-codebook coder (MCELP) operating at 24 kb/s and 16 kb/s. These coders were compared with speech produced by the G.722 coder operating at 64, 56 and 48 kb/s. Four speakers were used, two male and two female, each speaking different sentences for ten seconds. The tests were carried out by ten listeners. Tables 1 and 2 show the preferences expressed in these tests.

Comparison	G.722	MCELP	Neither
64 vs. 24	80 %	10 %	10 %
56 vs. 24	75 %	10 %	15 %
48 vs. 24	15 %	80 %	5 %
64 vs. 16	95 %	0 %	5 %
56 vs. 16	85 %	10 %	5 %
48 vs. 16	55 %	30 %	15 %

Table 1: Subjective test results for female speech.

Comparison	G.722	MCELP	Neither
64 vs. 24	20 %	50 %	30 %
56 vs. 24	0 %	75 %	25 %
48 vs. 24	5 %	70 %	25 %
64 vs. 16	45 %	40 %	15 %
56 vs. 16	35 %	40 %	25 %
48 vs. 16	20 %	70 %	10 %

Table 2: Subjective test results for male speech.

The above results indicate that, as with most CELP coders, the MCELP coder performs better when coding low pitch voices. The 24 kb/s coder is judged to be better tan the 64 kb/s G.722 coder for male speech while the quality is somewhere between that of the G.722 coder at 56 and 48 kb/s for female speech. At 16 kb/s we get a coder comparable to the 64 kb/s G.722 coder for male speech and comparable to the 48 kb/s coder for female speech.

The coders were also compared with another multi-codebook coder with three stochastic codebooks (i.e. no dynamic ordering and no adaptive bit allocation) and a conventional CELP coder using just one stochastic codebook. As expected the MCELP coder is significantly better than the coder using just one codebook. The performance of the coder using three codebooks was speaker specific. For some speakers there was a significant noise component present while for others it was virtually indistinguishable from our coder.

6. CONCLUSIONS

The subjective tests show that the use of multiple codebooks and synthetic glottal pulses coupled with adaptive bit allocation in modelling the excitation can produce very high quality wideband speech at bit rates as low as 24 kb/s and 16 kb/s. This gain in quality is achieved at the expense of a significant increase in the computational complexity of the coder over conventional coders. The complexity could be reduced by taking advantage of the sparsity of the glottal pulse codebook and using a sparse ternary codebook instead of the stochastic codebook. The subband codebooks could be more efficiently searched in the frequency domain

ACKNOWLEDGEMENTS

The authors would like to acknowledge the assistance provided by Teltec Ireland who sponsored the work.

REFERENCES

[1] C. McElroy, B.P. Murray and A.D. Fagan, "On Improving Wideband CELP Speech Coders", *Proc. EUSIPCO-94, Signal Processing VII: Theories and Applications*, Elsevier Science Publishers, 1994, pp. 912-915

[2] Bishnu S. Atal and Manfred R. Schroeder, "Stochastic coding of speech signals at very low bit rates", *Proc. Int. Conf. Acoust, Speech and Signal Processing*, 1984, pp 1610-1613

[3] A Fuldseth, E. Harbourg, F.T. Johansen, J.E. Knudsen, "Wideband Speech Coding at 16kbit/s for a Videophone Application", *Speech Communication*, Vol. 11, pp 139-148, 1992.

[4] Richard L. Zinser, "CELP coding at 4.0 kb/sec and below: improvements to FS-1016", *Proc. Int. Conf. Acoust., Speech and Signal Processing*, 1992, pp I313-I316

[5] Y. Linde, A. Buzo and R.M. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Trans. Comm.*, COM-26, April 1978, pp 702-710.

[6] W. B. Kleijn, "Encoding Speech Using Prototype Waveforms", *IEEE Transactions on Speech and Audio Processing*, Vol 1, No. 4, October 1993, pp 386-399.