EMBEDDING SIDE INFORMATION INTO A SPEECH CODEC RESIDUAL

Nicolas Chétry and Mike Davies

Centre for Digital Music
Queen Mary University, University of London
Mile End Road
E1 4NS London

{nicolas.chetry,mike.davies}@elec.qmul.ac.uk

ABSTRACT

We introduce a technique for embedding side information into a speech codec residual. While conserving the backward compatibility with existing decoders, it is described how it is possible to hide information into the speech long-term residual signal when it is encoded using a uniform or quasi-uniform quantiser. The method consists of embedding multiple parity bits at the quantiser level in a configuration that minimises the distortion over the whole subframe. The system has been evaluated quantitatively in terms of MOS and subjectively using "double blind" listening tests using the GSM-FR speech codec. It has been found that data can be embedded without severe perceptual degradation of the signal quality. Beyond this particular application to speech coding, it is shown how simple parity-check techniques can be developed to transparently transmit any binary data alongside the main encoded signal by using such joint quantisation and embedding scheme.

1. INTRODUCTION

Information embedding and watermarking are concerned with the process of *hiding* information into a digital representation of a signal. Common applications can be found in Digital Rights Management (DRM) for authenticating the provenance of an image, a piece of music or to control and restrict the use of a digital media content. However, some applications of information hiding do not necessarily need to be secure to attacks. As an example, embedding side information can be used to avoid the transmission of multiple synchronised data streams, to authenticate the speaker using a public PGP key [1] or to attach a business card alongside the speech data. In speech coding, applications aimed at hiding bandwidth extension information into the narrow-band speech signal have been proposed in [2]. After encoding and transmission, this information is used at the decoder-end to reconstruct the wide-band signal. These techniques have the advantage of transmitting a wide-band signal at the same bitrate as the narrow-band signal while conserving the backward compatibility with existing decoders.

Most watermarking techniques encountered in the literature act at the signal level or in a transform domain (*source embedding*). The task consists of designing algorithms that are robust to multiple encoding/decoding processes, filtering operations or collusion attacks. Techniques of spread spectrum [3], transform encryption coding [4] or using the masking property of the human auditory system [5] have been, among others, described in the literature.

Whereas conventional watermarking techniques can be used to transparently transmit side information through the channel, it can be argued that the constraints imposed by robustness considerations reduce the performance of the watermarking. Indeed capacity has to be sacrificed in order to make the embedding robust. For this reason, the development of *joint quantisation and embedding* techniques is highly desirable. The purpose of this paper is to study and elaborate techniques that can be applied at the quantiser level. More specifically, we propose an algorithm based on a binary block coset code to hide data into the long-term residual of a traditional speech codec. In essence, a convolutive code and a Viterbi search algorithm are used for the embedding operation while at the receiver-end, the decoding process simply involves a trivial parity bit calculation.

The paper is organised as follows. In Section 2, details about the GSM-FR codec are provided and the emphasis is on the Regular Pulse Excitation (RPE) quantiser. Next, in Section 3, a novel technique used to hide information into a bitstream using multiple parity bit embedding and convolutive codes is presented. In Section 4, details about the evaluation protocol are given and the quantitative and subjective results are reported. Finally, a discussion about the application of the method to the bandwidth extension of speech signals is presented in Section 5 while conclusions about the advantages and the limitations of the method close the paper.

2. THE GSM-FR SPEECH CODEC

The GSM-6.10 [6] full rate speech codec is the first standardised codec to be used over GSM networks. It encodes signals sampled at 8 kHz, operates at 13 kbits/s, has a MOS of 3.5 and is based on a RPE-LTP (Regular Pulse Excitation - Long-Term Prediction) encoding scheme. An overview of the encoding process is given in this section.

2.1 General principle

The codec is based on the linear predictive principle and the encoder contains a closed analysis/synthesis loop to limit the effect of the quantisation on the filter coefficients on one hand and on the long-term predictor parameters on the other hand. During the LPC analysis stage, 8 reflection coefficients are calculated using the Schur algorithm. These parameters are then transformed to Log Area Ratios (LAR) before quantisation and transmission (LAR are derived from the reflection coefficients but provide better quantisation properties [7]). At the same time, the set of LPC parameters is interpolated and transformed back into a set of reflection coefficients. The latter is used in conjunction with an LPC lattice analysis filter to produce 160 samples of short-term residual signal.

During the LTP analysis stage, the short-term residual signal is divided into 4 sub-frames of 40 samples. For each sub-frame, the parameters of the long-term one tap analysis filter, the LTP lag and the LTP gain, are estimated. The lag is found with values ranging from 40 to 120 by determining the maximum cross-correlation between the current short-term sub-frame and the previous 3 reconstructed ones. A block of 40 long-term residual signal samples is then obtained by subtracting 40 estimates of the short-term residual signal from the short-term residual signal itself. The resulting block of 40 long-term residual samples is fed to the RPE anti-aliasing filter. Next, the block of 40 input long-term residual samples is represented by one of 4 candidate sub-series of 13 pulses each. The selected sub-series (i.e. the one having the maximum energy) is identified by its RPE grid position. The maximum amplitude of the 13 RPE samples is estimated, and then quantised to 3 bits. Next, each of the 13 RPE samples of the current sub-frame is normalised

N. Chétry is supported by the Department of Electronic Engineering at Queen Mary, University of London.

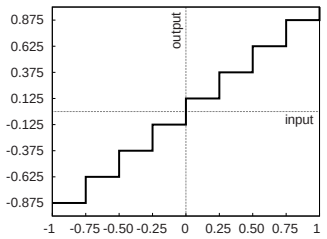
using the de-quantised normalisation factor and fed to the quantiser described in the following section.

2.2 RPE quantiser

For each sub-frame, each of the 13 RPE samples is independently quantised to 3 bits using the quantiser described in Figure 1.

Min	Max	Level
-32768	-24577	**-28672**
-24576	-16385	**-20480**
-16384	-8193	**-12288**
-8192	-1	**-4096**
0	8191	**4096**
8192	16383	**12288**
16384	24575	**20480**
24576	32767	**28672**

(a)



(b)

Figure 1: RPE quantiser levels. (a) fixed point representation on 15 bits and (b) floating point graphical representation.

The main thrust of the GSM-FR codec – and more generally of LPC-based speech codecs – is to remove as much spectral redundancy from the signal as possible so that the normalised RPE samples can be subsequently treated as uniformly distributed memoryless random variables.

While more modern speech codecs utilise more sophisticated perceptual and Vector Quantisation (VQ) based schemes for the long-term residual signal encoding, the inefficiencies in the scalar quantisation can in theory be "soaked up" by a well designed information embedding strategy, thereby rendering the joint coding/embedding scheme more efficient [8]. In the next section, we present one approach to generate such an embedding scheme using multiple parity bit embedding.

3. SIDE INFORMATION EMBEDDING

The noiseless information embedding problem can be solved with the use of 'Wyners method' based on coset error correcting codes as nicely illustrated by Chou *et al.* in [9]. The simplest such codes are probably the Quantisation Index Modulation (QIM) or Dither Modulation (DM) schemes introduced in [10]. Binary DM can be applied to the normalised RPE signal by splitting the quantiser codebook into 2 subsets (those with odd least significant bit and those with even least significant bit) and encoding a '1' or a '0' from the message sequence by choosing which subset to use to encode a given sample. This idea is illustrated in Figure 2.

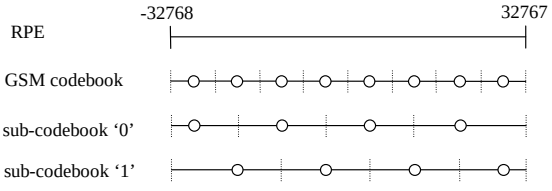


Figure 2: A schematic representation of the Dither Modulation principle for embedding at a rate of 1 bit per sample.

The edges of the quantisation cells that give minimum distortion are marked by the vertical lines. Under the high resolution assumption DM can embed 1 bit/sample with an added distortion of 6 dB (i.e. equivalent to 1 bit reduction in quantiser resolution). Note that this is 2.43 dB better than simply replacing the Least Significant Bit (LSB) with the message bit to embed [10].

In principle, one can embed information even more efficiently by building block embedding codes. Interestingly, for low embedding rates ($\leq 1/2$ bit/sample) there is also theoretically no significant loss in performance by restricting ourselves to binary block codes based around the DM sub-codes illustrated in Figure 2 [8].

3.1 Embedding one parity bit

The simplest way to extend DM to a binary block code is to treat DM as selecting the codeword with minimum distortion that also satisfies the constraint:

$$b_k + m_k = 0 \mod 2 \tag{1}$$

where b_k is the LSB of the quantised sample and m_k is the message bit to be embedded.

It is now simple to replace Equation (1) by a parity constraint over multiple samples:

$$\sum_{i \in \mathscr{I}_k} b_i + m_k = 0 \mod 2 \tag{2}$$

where $\mathscr{I}_k$ is the parity set. That is: a set of K consecutive samples.[1]

If the individual parity sets $\mathscr{I}_k$ are contiguous, then the embedding rate is simply $R = 1/K$. Furthermore, encoding only requires at most one of the original quantised samples to be modified per parity set (none if Equation (2) is satisfied with the original quantisation) and the optimal sample to modify is the one that introduces the most distortion in the original quantisation. This codeword should thus be replaced by the minimum distortion codeword coming from the other sub-code.

Decoding can, of course, be simply implemented by direct calculation of Equation (2).

Such a scheme can also be shown to be a very efficient form of information embedding at very low embedding rates (long blocks). Indeed, under the high resolution assumption, it can be shown [8] that the distortion D introduced by the embedding is:

$$D/D_0 = 1 + \frac{6R^2}{1+R} \tag{3}$$

where D_0 is the original quantisation distortion.

3.2 Overlapping parity embedding codes

The main problem with embedding parity bits into contiguous sets of samples is that we can either have a powerful code (large sets) with a very low embedding rate or a weak code (small sets) with a faster embedding rate. That is: the embedding rate and parity set size are rigidly linked. A simple way to overcome this is to remove the constraint that the parity blocks be contiguous and instead, allow them to overlap. This case is illustrated in Figure 3.

The price to pay for such a scheme is that minimum distortion embedding can no longer be calculated by simply changing the quantisation of one sample per parity set. Instead we will have interaction between the individual parity set equations. Fortunately as can be see from Figure 3, this is a particular instance of a convolutive coset code that has a simple decoding rule. The minimum distortion embedding can therefore be calculated using a form of the Viterbi algorithm [13], in a similar manner to encoding with trellis based VQ [14].

It has also to be noted that while the embedding complexity has increased, the complexity of decoding remains unchanged, again simply requiring the calculation of the parity bits using the Equation (2).

3.3 Algorithm implementation

In the GSM-FR codec, independent sub-frames of 13 RPE samples are considered. In Figure 3 is depicted an example of an overlapping parity embedding code for one sub-frame. In this particular case, the parity set size has been set to 7, the sample shift to 1 and the message to embed m_k is 7 bits long.

[1] This idea of embedding a parity bit in multiple samples is not new in information hiding and steganography (see for example "The power of parity" [11] or more recently [12], where a technique for speech authentication and integrity verification using the GSM-FR RPE residual has been proposed).

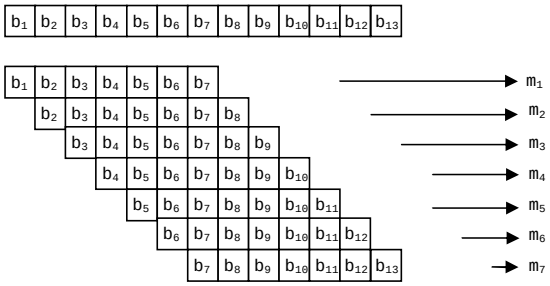


Figure 3: An overlapping parity embedding code depicted for a block of 13 samples. The least significant bit for each sample (b_i) is used in the parity calculation for a message of 7 bits (m_k). In this case, parity set size (K) is equal to 7, and the sample shift is equal to 1.

The code has two parameters that can be adjusted. The parity-set size, K, and the sample shift between consecutive parity-sets (or equivalently the number of bits per block). The fixed overall block length of 13 samples in our case also provides a restriction on the set of possible codes that can be used. That is:

$$K + \text{sample shift} \times (n \text{ bits/block} - 1) = 13 \qquad (4)$$

We have experimented with different parameter settings and in our experience, the parity set sizes given in Table 1 for the different embedding rates have produced the best results in terms of MOS equivalent. Further, it has to be noticed that the computational requirement of the embedding process grows exponentially with the parity set size K, thus setting some limits in the case where practical real-time applications are envisaged.

4. ALGORITHM EVALUATION

The performances of the data hiding algorithm are evaluated. The database consists of a subset of the TIMIT database and contains 10 files from 10 speakers (5 males and 5 females) that have been resampled at 8 kHz. The message to be embedded in each sub-frame of 13 RPE samples consists of a random binary sequence of 3, 4, 5 and 6 bits respectively.

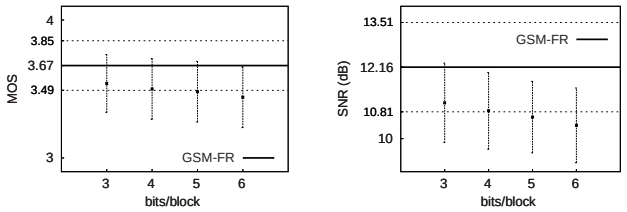
4.1 Quantitative evaluation

The Perceptual Evaluation of Speech Quality (PESQ) code [15] has been used to estimate the MOS equivalent. In order to calibrate our implementation of the GSM-FR speech codec, its corresponding average MOS has been calculated using all the files in the database. For 100 files, the average MOS equals 3.67 ± 0.18 while the average signal to noise ratio is 12.16 ± 1.35 dB.

	3 bits/block	4 bits/block	5 bits/block	6 bits/block
	600 bits/s	800 bits/s	1000 bits/s	1200 bits/s
K	7	7	5	8
MOS	3.54 ± 0.21	3.50 ± 0.22	3.48 ± 0.22	3.44 ± 0.22
SNR	11.08 ± 1.20	10.84 ± 1.16	10.65 ± 1.08	10.40 ± 1.13

Table 1: Average MOS equivalent (using the PESQ algorithm provided in [15]) and SNR for different embedding rates. Corresponding standard deviations are also reported. K are the parity set sizes that have been empirically determined to minimise the corresponding MOS equivalent.

In Table 1 are given the MOS and SNR for different embedding rates, ranging from 3 bits per block of 5 ms (equivalent to 600 bits/s) to 6 bits per block of 5 ms (1.2 kbits/s). Although it is acknowledged



(a) MOS (b) SNR

Figure 4: (a) MOS equivalent and (b) SNR for the embedding rates of 3, 4, 5, 6 bits per block respectively. The average MOS and its standard deviation values for the GSM-FR are shown in horizontal bold and dashed lines respectively.

that the signal to noise ratio is not a perceptually relevant measure of the distortion of a signal, the comparison of its evolution with the MOS can give information about the overall level of degradation introduced by the embedded data. It can be noticed in Figure 4 that hiding 5 bits per block degrades the signal quality to just below one standard deviation of the average performance of the GSM-FR, both in terms of MOS and SNR. On the other hand, the mean performances of embedding 3 and 4 bits per block fall within the GSM-FR natural distortion range. Overall, the system behaves as expected as the MOS values decrease as the embedding rate increases. The next section is concerned with the subjective evaluation of the signals quality as a function of the embedding rate.

4.2 Subjective evaluation

In order to subjectively assess the effect of embedding side information on the signals quality, "ABX–double blind" listening tests have been conducted. In each test, three files were presented to the listeners, 'A' and 'B' being the references – in our case, one was the original GSM-FR encoded/decoded file and the other one had data embedded – and 'X' the unknown sample to identify. Note that at each run, the listeners had no information about which file was the original GSM-FR and which file was the one with embedded data. Listeners have then to identify 'X' as being either 'A' or 'B'. Four expert listeners were asked to evaluate 20 pairs of files for each of the four considered embedding rates. The results are presented in Table 2:

	3 bits/block	4 bits/block	5 bits/block	6 bits/block
(1)	50%	60%	50%	45%
(2)	70%	65%	65%	60%
(3)	60%	55%	60%	55%
(4)	50%	65%	55%	55%
Total	**57.5**%	**61.3**%	**57.5**%	**53.8**%
p	**10.9**%	**2.8**%	**10.9**%	**28.8**%

Table 2: Results of the ABX listening tests for the four considered embedding rates. In rows numbered (1), (2), (3) and (4) are reported the percentages of correct answers for each listener as a function of the embedding rates. In the row labelled "Total" are reported the percentages of correct answers for each embedding rate averaged across listeners. Finally, p represents the probability of achieving at least this many correct answers by chance alone.

Note that listener (2) is one of the authors, thus explaining the greater ability to discriminate between the two files. Based on these results, the following conclusions can be drawn. It can firstly be observed that there is no great difference between the percentages of correct answers for the four considered embedding rates. For

instance, at 5 bits/block, listeners performed just as badly at identifying the correct files as for an embedding rate of 3 bits/block. This may suggest that the quality of the signals with embedded data greatly depends on the utterance. Although this type of listening tests does not allow to directly assess the amount of distortion introduced by the embedding operation, it can nevertheless be concluded that for 3, 5 and 6 bits/block, none of these results are statistically significant ($p < 5\%$). That is: we cannot rule out achieving these results by chance alone. However, for 4 bits/block, this is statistically significant ($p = 2.8\%$) so that the signals are not always indistinguishable.

5. DISCUSSION

A particular field of applications targeted with this technique is concerned with the bandwidth extension of speech signals. The aim of bandwidth extension is to generate the high frequency content of signal using small side information.[2] This side data conveys pertinent and essential information for the accurate reconstruction of the high frequency content at the user-end. In contrast to approaches using watermarking algorithms (e.g. [2]), the technique presented here *works* with the codec and is applied at the quantiser level. Thus the embedding is allowed to be fragile and therefore more efficient. On the other hand, one can note that watermarking techniques have the advantage of being independent of the codec used for the transmission.

In practice, and in the case of LPC-based encoding schemes, bandwidth extension techniques encountered in the literature split the process into two steps: the regeneration of the short-term residual on one hand and the regeneration of the wide-band spectral envelope on the other hand. The techniques of *spectral folding* or *spectral translation* [16], for example, can be used to recover a wide-band short-term residual signal. The latter is then fed into a reconstructed wide-band LPC synthesis filter to generate the wide-band signal. Therefore, the high-frequency spectral envelope information is the only data needed to be transmitted alongside the quantised base-band signal.

In [17], the wide-band spectral envelope parameters are encoded using a 8-bit vector quantiser. The corresponding amount of side information needed to reconstruct a wide-band signal (50–8000 Hz) from its narrow-band version (300-4000 Hz) is evaluated at 500 bits/s. It is further noticed that the reconstructed signals offer a noticeable quality improvement over the base-band signals. This equivalent cost of 2.5 bits/block is therefore compatible with the techniques presented here. In particular, it has been shown that 3 bits/block could be embedded with a loss of 0.13 points on the equivalent MOS scale.

The technique presented here embeds information in the RPE uniform quantiser signal by minimising the total RMS error over the whole block. Although not fully exploited here, a perceptual model could be used in order to spectrally shape the noise generated by the embedding operation below a psycho-acoustically relevant level. Other possible improvements include the study and design of embedding strategies dependent on the physical nature of the signal. For example, one could think of treating voiced and unvoiced frames differently by taking into account the masking properties of the vowel formant structure during the embedding process.

6. CONCLUSION

We have presented a scheme to embed data into a quasi-random host signal when it is encoded using an uniform quantiser. This technique of joint quantisation and embedding consists of embedding parity bits over overlapping blocks of quantised samples. A Viterbi search algorithm is then used to determine the optimum code that minimises the total distortion measure over the whole block. At the decoder end, the decoding consists of a trivial parity bit calculation. The practical use of such technique has been illustrated with an example of application in speech coding. Using an implementation of the GSM-FR speech codec, it has been quantitatively and subjectively shown that 1.2 kbits of data could be embedded every second with very limited perceptual degradation of the encoded/decoded speech signals. Beyond the particular application for data embedding, we have described how this technique could be used to transmit high-frequency band information using a bandwidth extension algorithm. This constitutes the natural extension of this research.

[2] Techniques for *blind* bandwidth extension using no side information exist but generally result in recovered wide-band signals of lower quality.

REFERENCES

[1] P. R. Zimmermann, *PGP: source code and internals*, The MIT Press, 1995.

[2] B. Geiser, P. Jax, and P. Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *Proc. Eurospeech*, 2005.

[3] Q. Cheng and J. Sorensen, "Spread spectrum signaling for speech watermarking," in *Proc. ICASSP*, 2001.

[4] F. J. Ruiz and J. R. Deller, "Digital watermarking of speech signals for the National Gallery of the Spoken Word," in *Proc. ICASSP*, 2000.

[5] A. Sagi and D. Malah, "Data embedding in speech signals using perceptual masking," in *Proc. Eusipco*, 2004.

[6] ETSI-TS 100-961 v 7.1.0 (2000-07), "Digital cellular telecommunications sytem (Phase 2+), Full Rate Speech, Transcoding," *GSM 06.10 version 7.1.0*, 1998.

[7] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 3, pp. 309–321, June 1975.

[8] M. E. Davies, "On the efficiency of embedding information within scalar quantizers," *Submitted to IEEE Trans. Information Theory*, 2006, Preprint available at `http://www.elec.qmul.ac.uk/people/miked/documents/Infor_embedding_submission.pdf`.

[9] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between data hiding and distributed source coding," in *Proc. 33rd Annual Asilomar conference on Signals, Systems, and Computers*, 1999.

[10] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[11] R. J. Anderson and F. A. P. Petitcolas, "On the limitations of steganography," *IEEE Journal of Selected Areas in Comms., Special issue on Copyright & Privacy Protection*, vol. 16(4), pp. 474–478, 1998.

[12] S. Yuan and S. A. Huss, "Audio watermarking algorithm for real-time speech integrity and authentication," in *Proc. of the 2004 workshop on Multimedia and Security MM&Sec*, 2004.

[13] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, pp. 260–269, 1967.

[14] A. Gersho and R. M. Gray, "Vector quantization and signal compression," *Kluwer Academic Publishers*, 1992.

[15] ITU-T P. 862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," February 2001.

[16] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. ICASSP*, 1979.

[17] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc. IEEE Speech Coding Workshop (SCW)*, 2000, pp. 130–132.