REMOVAL OF SPARSE-EXCITATION ARTIFACTS IN CELP

Roar Hagen[⋆]*, Erik Ekudden*[⋆]*, Björn Johansson*[⋆]*, and W. Bastiaan Kleijn*[†]

[⋆] Speech Coding Research
Ericsson Radio Systems AB
S-164 80 Stockholm, Sweden

[†] Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
S-100 44 Stockholm, Sweden

ABSTRACT

In CELP, the use of codebooks with entries with only a few non-zero samples provides high speech quality and facilitates fast computation. With decreasing bit-rate, the intervals between the pulses increase, and the quality of the reconstructed signal begins to suffer from a particular type of artifact, which is strongest for noise-like segments. In this paper we describe experiments which show that the perceived artifacts are mainly concentrated at frequencies above 3 kHz, and this is consistent with our understanding of auditory theory. Our analysis leads to simple strategies to eliminate the artifacts, even at lower bit rates. We describe both a non-adaptive and an adaptive post-processing method to remove the artifacts. The methods are demonstrated to be efficient when used in the ACELP algorithm. A closed-loop method for ACELP is also described.

1 INTRODUCTION

The CELP speech coding algorithm is the basis for many speech coding standards, and is used in many applications. While the potential of the algorithm for efficient coding of speech was recognized immediately, it was initially thought that the computational complexity of the algorithm would prevent its practical application. These fears were alleviated by a large number of fast procedures developed by researchers. In particular the introduction of zero samples in the so-called fixed codebook is used in a large number of fast algorithms [1, 2]. In this paper we will refer to such codebooks as *sparse* codebooks. Among these algorithms, the ACELP algorithm [3, 4] is currently most commonplace and used in the ITU-T G.729 coder [4], the enhanced full-rate GSM coder [5], and the enhanced full-rate D-AMPS coder [6]. The multi-pulse LPC coder is another example of a coder relying on sparse codebooks [7].

While the primary motivation for sparseness is to facilitate fast computations, artifacts are introduced in the reconstructed speech when the number of non-zero samples decreases too far, limiting the usefulness of the method at lower bit rates. This motivated us to search for methods to remove these artifacts. We started with a perceptual analysis of the artifacts. It turns out that the artifacts are time-domain effects residing mainly above 3000 Hz. We have designed both post-processing and closed-loop methods to eliminate the artifacts. The post-processing procedure eliminates most of the artifacts. The closed-loop procedure is slightly more efficient as the method is incorporated into the analysis-by-synthesis. Furthermore, the procedures can be enhanced by adapting the parameter settings.

The outline of the paper is as follows. We start with a section defining sparseness more precisely, both from the coder viewpoint and, more importantly, from the perceptual viewpoint. In Section 3 we describe experiments which confirm the conclusions of Section 2. Section 4 then describes a practical method to remove the artifacts using post-processing only and Section 5 describes a closed-loop procedure. In Section 6 subjective test results for a 6.4 kbps ACELP coder with a sparse codebook are given. Section 7 provides our conclusions.

2 EFFECTS OF SPARSE CODEBOOKS

In the design of a CELP coder, sparseness simply means that the entries of the fixed codebook consist mainly of zero samples. For example, in the G.729 coder and the D-AMPS EFR coder, there are only 4 non-zero samples per 40 samples sub-frame (at 8 kHz sampling rate). These 4 non-zero samples require the relatively high bit-rate of 3.4 kbps. With a decrease in bit-rate, the number of non-zero samples in the codebook decreases. This decrease is accompanied by an increasingly strong perceptual artifact which is most dominant in noise-like signal segments, such as unvoiced speech and background noise, and less noticeable in nearly periodic signal segments such as voiced speech. With only 2 non-zero samples per sub-frame, the artifacts are quite strong and impair the quality of the reconstructed signal significantly. The artifacts are perceived as an annoying "quasi-periodic" component added to the signal.

It is well-known that the first stage processing of the human auditory periphery is accurately described by a filterbank (e.g. [8, 9, 10]). We will refer to this filterbank as the auditory filterbank. Exploiting the auditory filterbank, we define *time-domain* features as variations in the signal power of a particular filter, and a *frequency-domain* feature as a structure in the signal power outputs across a number of adjacent filters averaged over a time interval. (For a quantitative treatment it is perceptually more accurate to use the low-pass filtered, half-wave rectified signal instead of the signal power.) In general, when the filter bandwidths are small, frequency-domain effects are more dominant, while when the filter bandwidths are large, time-domain effects are more dominant.

An increase in sparseness of the fixed codebook of a CELP coder will decrease the reconstruction accuracy of both time-domain and frequency-domain features. The relative decrease is affected by the statistics of the original signal. If the ideal excitation for the LP-based synthesis filter is similar to white noise, then the time-domain features of the signal will be affected much more significantly than the frequency domain features. On the other hand, if the ideal excitation is itself sparse in nature, then both time and frequency-domain features will be affected similarly. Since the artifacts are much stronger in regions of the first class, we conclude that the increasing distortion of the time-domain features of the signal is most significant in the artifacts associated with sparseness.

The role of the long-term predictor (LTP) in CELP offers another explanation for the larger problems in regions

where the ideal excitation is noise-like. In such regions, the relative contribution of the LTP to the excitation is small compared to the sparse fixed codebook which alone cannot create a smooth signal power contour. In more periodic signal segments, the LTP is the major contributor to the excitation and the sparseness of the fixed codebook leads to much less prominent artifacts.

The bandwidth of the filters of the filterbank describing the first stage of processing by the auditory periphery increases from around 100 Hz at 100 Hz to about 500 Hz at 3000 Hz. Independent of any processing of the signal beyond this filterbank, this pre-preprocessing stage makes time-domain effects more likely at high frequencies.

The discussion of the auditory system in this section suggests that the artifacts associated with sparseness are due to distortion of the time-domain features, and that this distortion should be most audible in the higher frequency ranges. In the next section we describe experiments which confirms these findings.

3 BASIC EXPERIMENTS

The goal of the basic experiments was to characterize the perceptual artifacts in a signal with a relatively high level of sparseness. For this purpose we used signals with both speech and background noise. The signal was first divided into overlapping windows with a length of 10 ms. The windows were shaped so that the addition of the windowed signals resulted in exact reconstruction. On each of the windowed signals we performed a discrete Fourier transform, and we manipulated the transform in various frequency bands. The coder used for the experiments was the D-AMPS EFR coder [6], where the fixed algebraic codebook was replaced with an algebraic codebook with only 2 pulses.

In an initial informal experiment, the magnitude or phase spectra of the coded signal was replaced with the corresponding spectra from the original uncoded signal. This experiment confirmed that the sparseness artifacts are due to time-domain effects as replacing the magnitude did not remove the artifacts whereas the replacement of the phase removed the artifacts completely. Therefore, the following experiments concentrated on altering the phase spectrum in various frequency bands.

The artifacts induced by sparseness are strongest for noise-like signals. For such signals, the power of the output from the filters of the auditory filterbank should be smooth in time. This is changed with the sparse excitation. To make the power output from the filters smoother, a random component was added to the phase spectra in various frequency bands of the decoded signal. Note that randomizing the phase of the discrete Fourier transform coefficient results in an elimination of periodicity present in the frequency band. However, the perception of pitch is dominated by low frequencies, where we expect the artifacts induced by sparseness not to be of any significance. It was found that the artifacts were eliminated by adding a random component to the phase of the discrete Fourier transform coefficients. As expected, the artifacts dominate for the higher frequency ranges. Modifying the phase spectra in the 3000-4000 Hz band removed the artifacts, especially for background noise conditions. However, modifying the phase of the decoded speech signal itself introduced some amount of unnaturalness to clean speech.

To reduce any possible effects of block-wise modifying the Fourier spectrum, a further experiment was performed where the excitation signal to the LP synthesis filter was modified instead of the decoded speech signal. In this way the modified excitation signal is smoothed by both the synthesis filter and the postfilter. Informal listening tests confirmed that the unnaturalness problems due to block processing in the speech domain were eliminated. The method still maintained its efficiency in removing the sparseness artifacts. This means that the reduction in periodicity by our procedure to eliminate the sparseness artifacts did not lead to any significant adverse effects in the perceived quality of the reconstructed signal.

4 POST-PROCESSING ARTIFACT REMOVAL

The decoder structure used to include the phase modification as a post-processing only technique into a CELP coder is given in Figure 1. The phase modified excitation may alternatively be used to update the LTP history. In this case, however, the encoder also use the modified signal and the technique is no longer a pure post-processing operation.

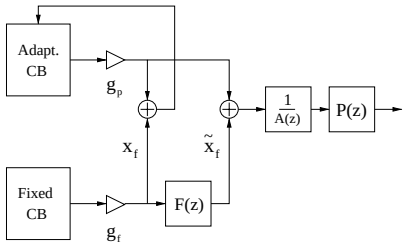

Figure 1. Post-processing phase modification in CELP.

4.1 Non-adaptive processing

When trying to find a method, based on the findings in the previous section, that would be efficiently implementable in a CELP coder, several questions were raised:

1. Can the phase randomization be applied to the fixed codebook excitation only in each sub-frame without using overlapped transforms?
2. Can a fixed modification of the phase in each sub-frame be employed?
3. Can the method be implemented in the time-domain?

Further informal experiments based on the methodology of the previous section gave a positive answer to the first two questions. This means that the method does not require additional delay. The complexity will also be manageable since the method can be applied to the sparse fixed codebook signal. Furthermore, the usage of a random generator is avoided.

The method does, at this point, consist of multiplying the Fourier transform of the fixed codebook entry in each subframe by a transfer function modifying the high-frequency phase only. This is expressed by

$$\tilde{X}_f(\omega) = F(\omega) \cdot X_f(\omega) \tag{1}$$

where

$$F(\omega) = e^{j\theta(\omega)} \tag{2}$$

and $\theta(\omega)$ is the random component added to the phase spectrum. $X_f(\omega)$ is the Fourier transform of the fixed codebook signal whereas $\tilde{X}_f(\omega)$ is the modified Fourier transform. The post-processed fixed codebook entry is obtained by an inverse Fourier transform of $\tilde{X}_f(\omega)$. Multiplying by a transfer function $F(\omega)$ in the Fourier-transform domain is equivalent to circular convolution in the time-domain with the impulse response $f(n) = IFFT\{F(\omega)\}$ corresponding to the transfer function. Therefore, the answer to question

3 is also positive. We have now obtained a low-complexity method since we only need to convolve with the few non-zeros samples of the fixed codebook signal. The modified fixed codebook signal, $\tilde{x}_f(n)$, is thus given by

$$\tilde{x}_f(n) = \sum_{k=0}^{P-1} \alpha_k f(n, m_k) \quad (3)$$

where α_k is the amplitude (including sign) and m_k is the position of the k'th pulse in the sparse codebook, P is the number of pulses, and $f(n, m_k)$ is the circularly shifted version of $f(n)$.

In Figure 2 the impulse-response $f(n)$ is shown for two examples. In both cases, the phase is modified only in the 3-4 kHz band whereas the amount of modification is in the range $-\pi/2$ to $\pi/2$ and $-\pi$ to π, respectively.

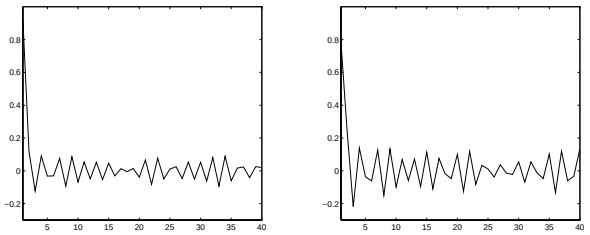

Figure 2. Impulse responses $f(n)$ for the phase modification. Left: Modification from $-\pi/2$ to $\pi/2$. Right: Modification from $-\pi$ to π.

In order to assess the efficiency of the phase modification method, we performed A-B listening tests involving 3 coders all based on the D-AMPS EFR coder [6]. Only the fixed codebook was different among the coders. The fixed codebook had 2, 3, and 4 pulses in the algebraic codebook, respectively. The bit rates of the coders are 6.2 kbps, 6.8 kbps, and 7.4 kbps (the original D-AMPS EFR coder), respectively. A number of pair-wise comparisons were made between these coders where one of the candidates had post-processing phase modification. The test material consisted of 14 speech samples including clean speech (4 samples), speech with car background noise (4 samples), speech with babble background noise (4 samples), pure car noise (1 sample), and pure babble noise (1 sample). Seven listeners performed the test, using telephone handsets. The results are shown in Table 1. The horizontal items employ phase modification. The impulse response of the phase modification filter used is the one shown in the right part of Fig. 2 (i.e. phase modification in the 3-4 kHz band and from $-\pi$ to π).

Table 1. Preference of horizontal item over vertical item.

	2-pulse	3-pulse	4-pulse
2-pulse mod	72%	42%	-
3-pulse mod	-	61%	50%
4-pulse mod	-	-	52%

As the number of algebraic codebook pulses is increased, the distortion due to sparse excitation is reduced. Therefore, with four pulses, the improvements by using phase modification is marginal. For three and two pulses, however, the increased quality is noticeable. Significantly, the 3-pulse codebook with phase modification is judged to have a quality identical to that of the original 4-pulse codebook.

4.2 Adaptive processing

It was noted in Section 3 that the sparseness of the fixed codebook is perceptually most noticeable for unvoiced speech and for speech in background noise. In voiced segments of speech, too strong phase modification may result in a slight noisyness. A further enhancement of the performance may be achieved if the phase modification filter is made signal adaptive. Two basic control parameters for the filter were selected:

- The cut-off frequency for phase modification.
- The maximum phase modification angle.

The selection of a particular filter to use for a segment (e.g. sub-frame) of speech can be made according to an implicit signal classification. A simple but efficient method is the use of the quantized LTP gain as illustrated in Table 2. The filter cut-off frequency, f_c, and the the maximum phase modification, θ_{max}, are given as a function of the quantized LTP gain, $\hat{g}_{LTP}$.

Table 2. Adaptive phase modification parameter settings.

$\hat{g}_{LTP}$	f_c	θ_{max}
> 0.9	-	0
$[0.5, 0.9]$	3000	$\pi/2$
< 0.5	2000	π

The improved performance using an adaptive filter selection as described above has been verified through informal listening tests. Especially the performance for background noise is improved due to the smoother excitation resulting from the stronger phase modification when $\hat{g}_{LTP}$ is low. The effect is close to that of adding a random noise component to the excitation. The quality is also improved for strongly voiced segments. With a short memory in the adaptive filter selection, the adaptive technique does not degrade noticeably under channel-error conditions.

The simple LTP gain adaptation may be further enhanced by incorporating more information describing the signal, such as the energy and the spectral shape, in order to detect e.g. background noise.

5 CLOSED-LOOP ARTIFACT REMOVAL

The phase modification of the excitation signal, as described in the earlier sections, may also be applied closed-loop in the analysis-by-synthesis stage. The structure for the closed-loop phase modification is given in Fig. 3.

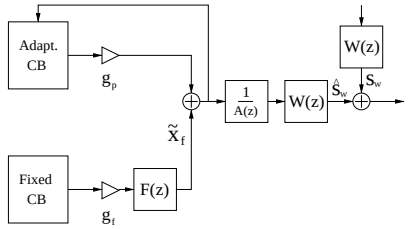

Figure 3. Analysis-by-synthesis phase modification in CELP.

Using an algebraic codebook, the fixed codebook search is performed by maximizing the term

$$T_k = \frac{(C_k)^2}{E_k} = \frac{(d^t c_k)^2}{c_k^t \Phi c_k} \quad (4)$$

where $d = H^t x$ and $\Phi = H^t H$. x is the target signal and H is the lower triangular matrix performing linear convolution with the search filter. c_k is the k'th entry in the sparse algebraic codebook.

With the phase modification filter included in the search, a phase modified algebraic codebook entry should be used in eq. 4. Therefore, c_k is replaced by $c_{f_k} = Fc_k$, where F is a matrix performing the circular convolution with the

phase modification filter impulse response $f(n)$. It is easily shown that this is equivalent to replacing d and Φ in eq. 4 by $d_f = (HF)^t x$ and $\Phi_f = (HF)^t(HF)$. The efficient algebraic codebook search algorithm in the ACELP coder can therefore be used unmodified, only the inputs d and Φ are changed.

Informal listening indicates that the closed-loop version provides slightly higher quality than the post-processing only version. It was also noted that a milder phase modification is suitable for the closed-loop case. We believe this is due to the fact that the phase modified fixed codebook signal is now used to update the adaptive codebook. An alternative structure to Fig. 3 updates the adaptive codebook with an excitation signal consisting of the (unmodified) fixed codebook contribution and the adaptive codebook contribution. Adaptive phase modification can also be incorporated in a straightforward manner.

The increase in complexity by including the phase modification is due to using the matrix HF instead of the matrix H. For circular convolution, F is not lower-triangular as the linear convolution matrix H. Thus, HF is not lower-triangular. An increase in complexity arises from both performing the matrix multiplication HF and using HF to compute d_f and Φ_f. The increase is managable since it occurs prior to the actual search in each sub-frame. It is, however, significant and strongly dependent on the sub-frame length.

In an alternative implementation, we employed linear convolution for the phase modification. This introduces a very small complexity increase since one can simply convolve the impulse responses for the search filter and the phase modification and use the resulting impulse response to construct the matrix H. The ringing of the phase modification must now be taken into account. With this lower-complexity method, we obtained quality indistinguishable from using the original circular convolution.

6 FORMAL SUBJECTIVE EVALUATION

The non-adaptive post-processing phase modification has been included in a 6.4 kbps downscaled version of ITU-T G.729. The coder uses an algebraic codebook with two pulses for every sub-frame of 40 samples. The two pulses are positioned in two overlapping tracks with 16 and 32 positions resulting in an 11-bit codebook including the two sign bits.

Results from a formal subjective ACR (Absolute Category Rating) test in Swedish including 24 naive listeners are given in Table 3. Modified IRS filtered speech material was used and the listening was performed in handsets. Results for the reference coders G.726 (ADPCM) at 24 kbps and G.729 at 8 kbps are also given.

The results show that the 6.4 kbps coder gives high quality for clean speech. The 6.4 kbps coder performs significantly better than G.726 at 24 kbps and close to the G.729 coder at 8 kbps. Therefore, the 6.4 kbps coder using the very sparse algebraic codebook with phase modification gives high quality suitable for many lower bit-rate applications.

Background noise performance has been evaluated both using DCR (Degradation Category Rating) tests and using A-B comparison tests. The results indicate that the 6.4 kbps coder performs worse than G.726 at 24 kbps, but that it has a performance close to G.729 at 8 kbps.

Table 3. Subjective ACR test results for the 6.4 kbps coder.

Condition	MOS	95% CI
G.726, 24k (-16 dBov)	3.03	0.182
G.726, 24k (-26 dBov)	2.84	0.164
G.726, 24k (-36 dBov)	2.57	0.158
G.726, 24k (tandem)	2.35	0.164
6.4k coder (-16 dBov)	3.27	0.186
6.4k coder (-26 dBov)	3.40	0.186
6.4k coder (-36 dBov)	3.33	0.176
6.4k coder (tandem)	2.59	0.188
G.729, 8k (-26 dBov)	3.81	0.188
MNRU 6 dB	1.20	0.126
MNRU 12 dB	1.68	0.162
MNRU 18 dB	2.45	0.167
MNRU 24 dB	3.29	0.178
MNRU 30 dB	3.97	0.176
MNRU 36 dB	4.26	0.168

7 CONCLUSION

Many speech coding approaches use sparse codebooks. Such codebooks are associated with artifacts when the bit rate is decreased, and the sparseness increased. In this paper, we have shown that it is possible to eliminate such artifacts effectively using both post-processing and closed-loop procedures. The post-processing procedure is very effective as demonstrated by listening test results. The closed-loop procedure performs somewhat better than the post-processing procedure. The effectiveness of both procedures can be further enhanced by making the operation adaptive. In general, the new procedures significantly increase the coding performance of sparse codebooks at low bit rates.

REFERENCES

[1] J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, "Fast CELP coding based on algebraic codes," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Dallas, TX), pp. 1957–1960, 1987.

[2] W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Fast methods for the CELP speech coding algorithm," *IEEE Trans. Acoust. Speech Sign. Proc.*, vol. 38, no. 8, pp. 1330–1342, 1990.

[3] R. Salami, C. Laflamme, J.-P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," *IEEE Trans. Vehic. Techn.*, vol. 43, no. 3, pp. 808–816, 1994.

[4] R. Salami et al., "Description of the proposed ITU-T 8 kb/s speech coding standard," in *Proc. IEEE Speech Coding Workshop*, (Annapolis, MD), pp. 3–4, 1995.

[5] K. Järvinen et al., "GSM enhanced full rate speech codec," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Munich, Germany), pp. 771–774, IEEE, 1997.

[6] T. Honkanen et al., "Enhanced full rate speech codec for IS-136 digital cellular system," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Munich, Germany), pp. 731–734, IEEE, 1997.

[7] B. S. Atal, "High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, (Tokyo, Japan), pp. 1681–1684, 1986.

[8] R. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, pp. 1560–1586, 1987.

[9] S. Seneff, "A joint synchrony/mean rate model of auditory speech processing," *J. Phonet.*, vol. 16, pp. 55–76, 1988.

[10] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, pp. 453–485, Marcel Dekker, 1991.