IMPROVEMENTS OF BACKGROUND SOUND CODING IN LINEAR PREDICTIVE SPEECH CODERS

Torbjörn Wigren   Anders Bergström   Susanne Harrysson   Fredrik Jansson   Hans Nilsson
Ericsson Radio Systems AB, S–164 80 Stockholm, Sweden

ABSTRACT

A proper coding of background sounds has proven to be important in digital cellular telephone systems. It is therefore of considerable interest to find methods that improve also background sound coding, in particular in medium and low bitrate speech coders. In many coders operating around 8 *kbps*, one significant effect has been a swirling distortion when coding for example smooth car noise. The paper discusses the origin of this problem, and presents an algorithm that reduces swirling, giving a significant overall perceptual improvement, as shown by field tests in commercial D–AMPS systems in the US. A related new algorithm for error concealment is also presented. This method makes it possible to perceptually remove the effect of fading dips persisting for a few seconds, during nonspeech periods.

1. INTRODUCTION

Traditionally, the main emphasis when developing and studying speech codecs for low to medium bit rates have been on their performance when coding speech. The study of the effects of disturbances has mostly been confined to investigations of the resulting quality of the coded voice. Little attention has been given to the coding of pure background sounds like car and street noise, music, and background speech. In a digital cellular telephone system these background sounds occur frequently, and they must be well coded during nonspeech periods in order not to be annoying at the far end.

However, background sound coding has turned out to be a significant problem in the North American D–AMPS system where the VSELP [1],[2] speech coder is the standard. In particular, it has been noted that coding of smooth background noise originating from for example fans or cars results in strong swirling or waterfall distortion. Perceptually, the original disturbance is transformed into a sound with close resemblance to flowing water. Swirling also occurs in other linear predictive codecs than the VSELP, like in [3],[4]. Also the backward adaptive LD–CELP codec [5] was reported to suffer somewhat from the problem with swirling, at least at 8 *kbps* [6]. However, the window tuning techniques proposed in [6] to remove swirling, appears to be of less value when the bitrate is reduced, since the problem with swirling increases with decreasing bitrate.

One major purpose of the paper is therefore to study the origin of swirling (section 2). With this analysis as starting point, an algorithm is then presented that can remove the distortion during nonspeech periods (section 3). The algorithm is applicable to any linear predictive coder, and it requires a voice activity detector (VAD) to operate [7]. The method has been implemented in the D–AMPS system, and results and experiences from field tests are reported in the paper (sections 5 and 6).

A new algorithm for error concealment during nonspeech periods is also presented (section 4). The conventional muting technique [8] that is applied after detection of a number of consecutive bad speech frames, is then replaced by a signal tailored from the previously decoded disturbance. Effects of shadowing or very slow Rayleigh fading, can then be well masked during non speech activity. The method utilizes ideas from the algorithm used in order to reduce swirling.

2. BACKGROUND SOUND MISCODING

A. Swirling Distortion

Simulation experiments were performed with a number of codecs in order to capture effects of varying bitrates, codec structures, frame lengths, real time implementations, and channel impairments. The codecs were fixed and floating point versions of the 7.95 *kbps* VSELP [1],[2], the 4 *kbps* Ericsson D–AMPS half rate candidate [3],[4], and a standard LPC vocoder [9], forced to unvoiced mode. The test material included flat and bandpass filtered (300–3400 Hz) car and babble noise (40 s each), as well as flat and bandpass filtered speech with the above disturbances added (10–20 dB signal to noise ratio). Effects of channel impairments were evaluated for the VSELP and the D–AMPS half rate candidate (1% BER and 3% BER, 7 Hz and 77 Hz doppler frequency). The performance of each codec was evaluated in informal listening tests by skilled listeners.

Experiment 1: The four codecs were run configured as described in [1]–[4]. Conclusions: 1) Swirling exists in all tested coders, and the distortion increases with reduced bitrate. Distortion is severe at 4 *kbps* (40 *ms* frames), cf. [6]. 2) Swirling is significant for ideal channels, and the distortion increases when channel quality is reduced. 3) Swirling increases in tandem configurations. 4) Input filtering does not affect swirling significantly. 5) Swirling is not caused by quantization or fixed point implementation.

Thus, swirling appears to be inherent in linear predictive coders as is also concluded in [6]. □

Experiment 2: To verify that swirling is caused by the LPC analysis block, and to secure that swirling has no

0-7803-2431-5/95 $4.00 © 1995 IEEE

other source, a number of tests were performed on modified versions of the codecs. Conclusions: 1) Swirling is significantly reduced when the temporal variability of the LPC parameters is reduced by 500 Hz of bandwidth expansion. 2) Nonstationary background sounds are deteriorated if the temporal variability is reduced too much. 3) Swirling is not affected by disconnection of the LTP. 4) Swirling increases when a postfilter is used.

Thus swirling is caused by an excessive temporal variability of the estimated short term spectrum, when encoding stationary noise, cf. Fig. 1.

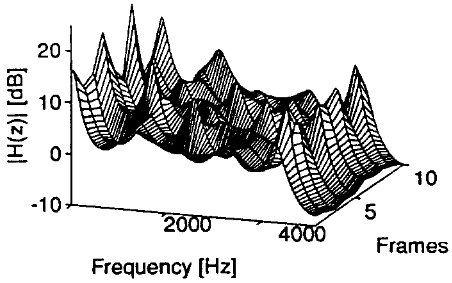

Fig.1. Consecutive short term spectra from the D-AMPS encoder. Stationary car noise input.

□

One way to understand the origin of the problems is to formally consider the least–squares solution commonly employed in LPC analysis. Denote the measured signal samples by $s(n)$ where n is discrete time, and collect the parameters, a_i, $i = 1, \ldots, m$, of the estimated all–pole model in the vector θ. Using a covariance formulation, the least squares estimate of θ is

$$\hat{\theta}_N = \left(\frac{1}{N} \sum_{n=1}^{N} \varphi(n)\varphi^T(n) \right)^{-1} \frac{1}{N} \sum_{n=1}^{N} \varphi(n)s(n) \qquad (1)$$

$$\varphi(n) = (s(n-1) \; \ldots \; s(n-m))^T \qquad (2)$$

where N is the length of the analysis window.

For (voiced) speech signals with high SNR, (1) and (2) computes the predictor that best matches the measured data. This usually works well, since a low order all–pole model is known to accurately model the vocal tract. Relatively few bits are therefore required to encode the residual signal with VQ–techniques.

A prediction of a spectrally shaped stationary and ergodic noise signal, with an accuracy resembling that of voiced speech is in general not possible. The reason is that the prediction gain achievable with LPC analysis in general is much lower for a noise signal than for voiced speech. However, if the underlying short term spectrum were known with a sufficiently high accuracy, a spectrally equivalent signal could be generated with white noise as source in the receiver. "Sufficiently high accuracy" is here tied to the ergodicity relation (3) obtained from (1)

$$\hat{\theta}_\infty = \lim_{N \to \infty} \hat{\theta}_N = \left(E\varphi(n)\varphi^T(n) \right)^{-1} E\varphi(n)s(n) \qquad (3)$$

where E denotes expectation. Unless N is large enough $\hat{\theta}_N$ may be far from its limiting value, making it possible for the short term spectrum to vary from frame to frame. Proper noise encoding, utilizing (3) can therfore be expected to require longer frames than (voiced) speech coding. Stated differently, when encoding noise the LPC analysis block should be operated more as a spectral modeler than as a linear predictor. The above discussion also indicates why swirling distortion increases with reduced bitrate.

B. Error Concealment

One typical error concealment scheme is proposed in [8]. There the LPC parameters and the frame energy $R(0)$, are replaced by the values obtained from the most recent frame that was classified as correct by the channel decoding scheme. [8] also specifies a state machine that totally mutes the output signal after six consecutively detected bad frames (120 ms).

However, for slowly Rayleigh fading channels the bit error rate is often close to zero for most of the time, except when a fading dip occurs. Then the bit error rate rises to about 50 % for a number of consecutive frames. This means that the output from the speech decoder is muted periodically at a low rate, typically a few Hz. This chopped output can be very annoying. Muting also distorts the output signal during shadowing periods.

For speech signals which are only short term stationary, fading dips persisting for more than a few frames inevitably give rise to a chopped signal as described above. However, the stationarity of certain types of background sounds like car noise should make it possible to mask fading dips persisting for much longer periods of time. This requires that different error concealment actions are applied during speech and nonspeech periods.

3. SWIRLING REDUCING ALGORITHM

A swirling reducing algorithm, intended to operate during nonspeech parts of the coded signal, is described in this section. In most systems, the main problem is the uplink direction. Therefore, the described add on algorithm is located in the *decoder* of an IS-54 VSELP [1],[2]. Then swirling can be reduced, no matter what type of terminal that is used.

As reported above, the origin of swirling distortion is too much variability in consecutive estimated short term spectra. A frame length of 40 ms was shown to be insufficient to solve the problem for low and medium rate codecs. Hence, the main idea is to remove the distortion by using a longer analysis window during *nonspeech* periods. To obtain this, a VAD is used [7]. A block diagram of the algorithm is given in Fig.2.

A longer spectral analysis window means that the autocorrelation function is computed over a longer frame of data. This can be implemented as

$$acf_K(l) = \frac{1}{K} \sum_{j=0}^{K-1} acf(j,l) \qquad l = 0, \ldots, m \qquad (4)$$

where K is the number of frames used for averaging, and where $acf(j,l)$ denotes the autocorrelation for lag l in the j:th frame of the sum. This formulation is beneficial when using the VAD of [7]. Since the algorithm resides in the decoder, the autocorrelation function must be retrieved from the quantized reflection coefficients and the frame energy. A slightly modified step up algorithm, described in the appendix is used for this purpose.

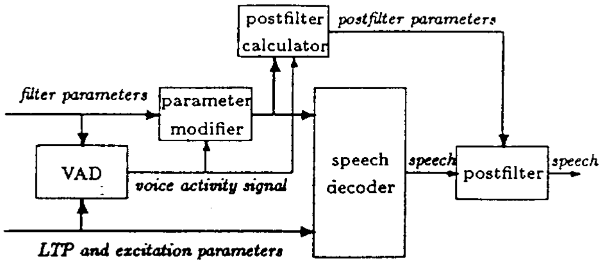

Fig.2. Swirling reducing algorithm.

The swirling reducing algorithm has an initial convergence period before it starts to operate, because of properties of [7]. For a typical SNR of 20 dB this time is about one second. At lower SNRs, the time increases.

In case the algorithm is located in the decoder with a postfilter, this filter should also be modified. A slight additional bandwidth expansion, in combination with enhanced low frequency characteristics is beneficial perceptually. The reason is that spectral averaging and bandwidth expansion tend to emphasize the high frequency region. When the algorithm is placed in the encoder, the decisions from the VAD can be expected to be improved, since decisions are not based on quantized data. Note that the modification of the filter parameters should then be performed after the excitation search, just prior to transmittal of the parameters. Informal tests show an increase of swirling distortion when the filter is modified before the excitation search. One reason may be that the signal is then whitened with the locally estimated filter and hence the excitation search performs better as compared to a modified filter.

To characterize the performance a number of tests were performed. It was found that $K = 8$ is a good compromise that gives a significant reduction of swirling distortion, without modifying performance in other situations. The postfilter "brightness" (see [1],[2]) is typically reduced from 0.40 to 0.15.

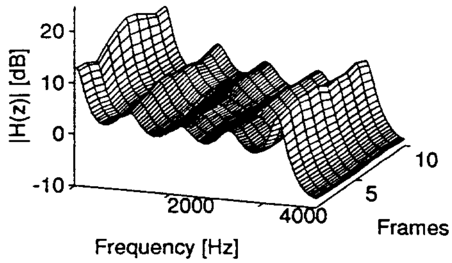

Fig.3. Consecutive averaged short term spectra obtained from the D-AMPS encoder.

Generally, informal tests showed good performance of the algorithm. In Fig.3, consecutive spectra obtained from the VSELP [1],[2] are plotted. The same input data as in Fig.1 were used.

4. ERROR CONCEALMENT ALGORITHM

When a VAD is available in the decoder, it becomes possible to use different error concealment techniques depending on whether speech or nonspeech is detected. As described above, slow fading results in a chopped output from the speech decoder. During nonspeech periods "stationary" sounds, like car noise, could be replaced by artificial sounds when long fading dips occur. Such artificial sounds have to match the spectral character of the original sound very well. This means that the particular sound needs to be stationary over the whole fading dip.

The spectrum of a background sound needs to be temporally stationary in order to be classified as nonspeech by the VAD [7]. The swirling reducing actions then adds to this stationarity, as described in section 3, which makes the situation even more favorable.

The error concealment algorithm operates as follows. In case nonspeech was detected by the VAD just before a burst of detected bad frames occured, the following actions are initiated: 1) The last correct LPC parameters obtained from the swirling reduction algorithm are used instead of the received parameters, during the whole burst. 2) The gain parameters of the LTP and the innovation are locked to the values from the last correct frame, during the whole burst.

The effect is that error bursts persisting for more than 100 frames can be well masked during nonspeech periods, as shown by informal listening tests.

5. IMPLEMENTATION

The swirling reduction and the error concealment algorithms have been implemented in the Texas Instruments TMS320C50 DSP, in the base station of the Ericsson D-AMPS system. The implementation was performed by systematic conversion from high level floating point code, to high level fixed point code, and then to fixed point assembly code. Objective and subjective comparisons between floating point and fixed point simulations have shown that there is no performance loss when implementing the algorithms using fixed point code. The assembler implementation has been verified to be bit exact with the fixed point implementation.

0.3 $MIPS$ are required for the VAD and the swirling reducing actions. The corresponding memory requirements are 900 words of ROM, 200 words of static RAM and 100 words of dynamic RAM (temporary variables). The computational complexity of the error concealment algorithm is low, $< 0.1 MIPS$. The extra memory required is 20 words of ROM and < 20 words of RAM.

6. FIELD TESTS

The swirling reducing algorithm has been further tested in a commercial D-AMPS system in West Palm Beach, FL, USA. The modified speech decoder software was installed at four sites in order to perform the tests. IS-54 without swirling reduction was used as a reference.

Configurations tested included swirling reduction in the encoder, the decoder and both. Mobiles, as well as handportable phones from Ericsson GE, Motorola and Hughes Network Systems were tested. The test cases included clean speech, speech and car noise, speech and car noise in handsfree mode, speech and restaurant noise, varying channel conditions, speech and street noise, as well as speech and car noise with music added. Quality was judged by a number of skilled listeners. Since there is no standardized methodology for characterization of performance for background sound coding, and since field tests are difficult to control in detail, only informal results are available. These results are summarized in Table 1. Note that the presented quality mostly relates to the particular background sound and not to the speech.

RESULTS OF FIELD TESTS				
Case	**IS-54**	**Enc.**	**Dec.**	**Mixed**
Clean speech	good	good	good	good
Car noise	poor	fair	fair	fair
Handsfree	poor	fair	fair	fair
Restaurant noise	fair	fair	fair	fair
Street noise	poor	fair	fair	fair
Music	poor/fair	poor	poor	poor
Car noise/BER	poor	fair	fair	fair

Table 1: Results of US field tests.

The performance for different mobiles were similar in all tests. The slight quality reduction in background music is present only for low level music which is interpreted by the VAD as background noise. The above results indicate that the proposed method for reduction of swirling works well in a majority of test cases. No reduction of speech quality because of possible wrong decisions of the VAD was noticed. This is most likely a result of the conservative setting of the swirling reducing actions.

7. CONCLUSIONS

Two add-on algorithms that improve background sound (primarily noise) coding in low to medium rate linear predictive speech coders were presented. The first algorithm reduces thre swirling distortion that is caused by temporal variability of the encoded short term spectrum. A related technique is the window tuning suggested in [6]. However, for low bitrates a 40 *ms* window as proposed in [6] appears to be too short, which means that a VAD is needed. The reduction of variability must be as moderate as possible so that no undesired distortion is introduced.

The other algorithm improves error concealment during nonspeech periods in cases where long fading dips occur (slow Rayleigh fading or shadowing). The method uses a VAD and it operates by replacement of conventional muting by a tailored disturbance signal derived from previously decoded frames. This allows masking of the effect of about 100 consecutive lost frames, or equivalently, of about 2 *s* of background noise.

The complexities of both algorithms are low. The swirling reducing algorithm is in use in Ericssons North American D-AMPS system.

APPENDIX

In this appendix, a slightly modified version of the stepup algorithm of [9] is described. The input is the frame energy, $R(0)$, of the input signal, [1],[2], and the reflection coefficients, $k_l, l = 1, \ldots, m$. The output is the autocorrelation function, and the direct form filter coefficients, $a_l, l = 1, \ldots, m$. The algorithm is

$$\begin{array}{rcl}
acf(0) & = & R(0) \\
\alpha^{(0)} & = & acf(0) \\
acf(1) & = & -k_1\alpha^{(0)} \\
a_1^{(1)} & = & k_1 \\
\alpha^{(1)} & = & \alpha^{(0)}(1-k_1^2) \\
\text{for} \quad l = 2 & \text{to} & m \quad \text{repeat} \\
acf(l) & = & -k_l\alpha^{(l-1)} - \sum_{j=1}^{l-1} a_j^{(l-1)} acf(l-j) \\
a_l^{(l)} & = & k_l \\
a_i^{(l)} & = & a_i^{(l-1)} + a_{l-i}^{(l-1)} k_l \qquad i = 1, ..., l-1 \\
\alpha^{(l)} & = & \alpha^{(l-1)}(1-k_l^2)
\end{array} \tag{5}$$

REFERENCES

[1] I. A. Gerson and M. A. Jasiuk, "Vector sum excited linear prediction (VSELP)", *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 66–68, 1989.

[2] I. A. Gerson and M. A. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps", *Proc. ICASSP*, pp. 461–464, 1990.

[3] H. Hermansson, T. B. Minde, J. Ahlberg and P. Lundqvist, "A speech codec for cellular radio at a gross bit rate of 11.4 kb/s", *Proc. ICASSP*, pp. 625–628, 1991.

[4] T. B. Minde, T. Wigren, J. Ahlberg and H. Hermansson, "Techniques for low bit rate speech coding using long analysis frames", *Proc. ICASSP*, Pt. II, pp. 604–607, 1993.

[5] J. Chen, R. V. Cox, Y. Lin and N. Jayant, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard", *IEEE J. Sel. Areas Com.*, vol. SAC–10, no. 5, 1992.

[6] J. Chen and R. V. Cox, "Convergence and numerical sensitivity of backward-adaptive LPC-predictors", *Proc. IEEE Workshop on Speech Coding for Telecommunications*, pp. 83–84, 1993.

[7] D. K. Freeman, G. Cosier, C. B. Southcott and I. Boyd, "The voice activity detector for the pan-European digital cellular mobile telephone service", *Proc. ICASSP*, pp. 369–372, 1989.

[8] I. A. Gerson, M. A. Jasiuk, M. J. McLaughlin and E. H. Winter, "Combined speech and channel coding at 11.2 kbps", *in Signal Processing V: Theories and Applications*, pp. 1339–1342, 1990.

[9] J. D. Markel and A. H. Gray, *Linear Prediction of Speech, 3:rd printing*. Berlin, Germany: Springer–Verlag, 1982.