INFLUENCE OF AUDIO CODING ON STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Tomas Gänsler and Peter Eneroth

Dept. of Applied Electronics, Signal Processing Group
Lund University, Sweden
tg@tde.lth.se, pe@tde.lth.se

ABSTRACT

Stereophonic acoustic echo cancellation has been found more difficult than echo cancellation in mono due to a high correlation between the two audio channels. Different methods to decorrelate the channels have been proposed so that the stereophonic echo canceller identifies the true echo paths and its convergence rate increases. In this paper it is shown that the use of a perceptual audio coder effectively reduces the correlation between the channels and thus convergence to the true echo paths is insured. Furthermore, in those frequency regions where the encoder introduced quantization noise is below the global perceptual masking threshold, an extra amount of inaudible noise can be added to the channels. Thereby the channel correlation is further decreased and the solution is stabilized. In subband audio coders with high frequency resolution only minor modifications are needed in the decoder.

1. INTRODUCTION

Two emerging applications for stereophonic acoustic echo cancellation are high quality videoconferencing and tele-gaming. In the future, desktop based conference systems will also need stereophonic acoustic echo cancellers (SAEC). These systems have different quality demands influencing the bandwidth and bitrates etc.

Stereophonic acoustic echo cancellation, however, has been found far more complicated than its monophonic counterpart. This is due to the fact that the two channels carry linearly related signals, [1], which leads to convergence problems of the echo canceller. Due to the linear relation between the channels there is, in theory, no unique solution for the echo canceller to identify. Moreover, the non-unique solutions that exist are all dependent on the echo paths in the *far-end* (remote) room, [1, 2, 3]. In real situations, however, the solution to the problem is not truly singular only extremely ill-conditioned due to uncorrelated microphone noise and infinite impulse responses of the remote room's echo paths, [4, 5]. The convergence rate of the NLMS algorithm is highly dependent on the condition number of the correlation matrix thus more sophisticated algorithms must be used in stereophonic echo cancelling, e.g. [2, 6, 7].

Despite using more sophisticated algorithms there are still problems with unstable estimates of the echo paths, [3]. In order to stabilize the solution the correlation between the stereo channels has to be reduced without introducing annoying distortion. A number of solutions to this problem has been suggested, see e.g. [1], but rejected for different reasons. The most promising solution so far is to distort the stereo channels non-linearly as proposed in [5] where a half-wave rectified portion (α) of the signal is added to the signal itself. This distortion does not destroy the stereophonic perception but introduces a noise that most often is inaudible but may be perceived depending on the level of introduced non-linearity, [8].

This work was supported by Telia Research AB, Stockholm, Sweden.

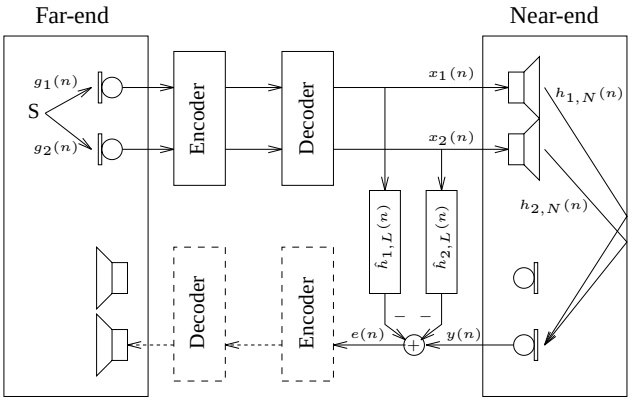

Figure 1: Audio coder and Stereophonic AEC. Only one return part is shown.

The objectives for this paper is to study perceptual *audio coding* as another option to reduce the correlation between the channels. A perceptual audio coder, depending on bitrate etc., introduces a quantization noise that most often is below the hearing threshold. The question is if this distortion is strong enough in order to make the solution to the stereophonic acoustic echo canceller problem "well-conditioned."

2. PROBLEM FORMULATION

The circumstances under which convergence to the true echo paths of an SAEC is achieved has been thoroughly analyzed in [5]. This section summarize some of their results that are used in this paper to formulate the problem and analyze the performance of the

SAEC.

Assume that the far-end microphone signals are given by, Fig. 1,

$$x_i(n) \quad = \quad g_i(n) * s(n),\ i = 1, 2, \tag{1}$$

where $s(n)$ is the source signal and $g_i(n),\ i = 1, 2$ are the far-end echo paths of length M. "$*$" denotes convolution. The residual echo $e(n)$ after the EC is

$$\begin{aligned} e(n) &= y(n) - \hat{\mathbf{h}}_{1,L}^T \mathbf{x}_{1,L} - \hat{\mathbf{h}}_{2,L}^T \mathbf{x}_{2,L} && (2)\\ y(n) &= \mathbf{h}_{1,N}^T \mathbf{x}_{1,N}(n) + \mathbf{h}_{2,N}^T \mathbf{x}_{2,N}(n) && (3)\\ \mathbf{h}_{i,N} &= [h_{i,0} \cdots h_{i,N-1}]^T, && (4)\\ \mathbf{x}_{i,N}(n) &= [x_i(n) \cdots x_i(n-N+1)]^T. && (5) \end{aligned}$$

$\mathbf{h}_{i,N},\ i = 1, 2$ are the true responses of length N of the near-end room and $\hat{\mathbf{h}}_{i,L},\ i = 1, 2$ are the estimated responses of length L.

Minimization of the weighted least squares criterion

$$J(n) \quad = \quad \sum_{l=1}^{n} \lambda^{n-l} |e(n)|^2,\ 0 < \lambda \leq 1, \tag{6}$$

results in solving the system of linear equations

$$\mathbf{R}_{xx}(n) \begin{bmatrix} \hat{\mathbf{h}}_{1,L} \\ \hat{\mathbf{h}}_{2,L} \end{bmatrix} = \mathbf{r}_{yx}(n), \tag{7}$$

where $\mathbf{r}_{yx}(n)$ is the estimated cross-correlation vector and $\mathbf{R}_{xx}(n)$ is the correlation matrix,

$$\mathbf{R}_{xx}(n) = \tag{8}$$
$$\sum_{l=1}^{n} \lambda^{n-l} \begin{bmatrix} \mathbf{x}_{1,L}(l)\mathbf{x}_{1,L}^T(l) & \mathbf{x}_{2,L}(l)\mathbf{x}_{1,L}^T(l) \\ \mathbf{x}_{1,L}(l)\mathbf{x}_{2,L}^T(l) & \mathbf{x}_{2,L}(l)\mathbf{x}_{2,L}^T(l) \end{bmatrix}.$$

The challenging problem with stereophonic echo cancelling all lies in the condition number of this matrix. It is also shown in [5] that,

$$\begin{aligned} L \geq M &\Rightarrow \mathbf{R}_{xx}(n) \text{ is singular } \forall n, \\ L < M &\Rightarrow \mathbf{R}_{xx}(n) \text{ is ill-conditioned }, \\ L \geq N &\Rightarrow \text{misalignment, } \varepsilon(n) \to 0,\ n \to \infty, \\ L < N &\Rightarrow \text{misalignment, } \varepsilon(n) \neq 0,\ \forall n, \end{aligned} \tag{9}$$

where the misalignment is $\varepsilon(n) = ||\mathbf{h} - \hat{\mathbf{h}}||^2 / ||\mathbf{h}||^2$ and $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_{1,L}^T\ \hat{\mathbf{h}}_{2,L}^T]^T,\ \mathbf{h} = [\mathbf{h}_{1,L}^T\ \mathbf{h}_{2,L}^T]^T$. An ill-conditioned $\mathbf{R}_{xx}(n)$ increases the misalignment in (9). Thus there is a contradiction, if $L << M$ the solution of (7) is better conditioned, on the other hand $L = N$ reduces misalignment, but practically $L < M \approx N$. The solution to this misalignment problem is therefore to *decrease the correlation between the stereo channels, thus reducing the condition number of* $\mathbf{R}_{xx}(n)$.

The eigenvalues of the correlation matrix can be lower bounded by $[1 - |\gamma(f)|^2]$, where $\gamma(f)$ is the coherence between the stereo channels [5]. Ill-conditioning can therefore be monitored by the coherence function which serves as a measure of achieved decorrelation. The next section explains how decorrelation can be achieved by having a perceptual audio coder in the transmission path, Fig. 1.

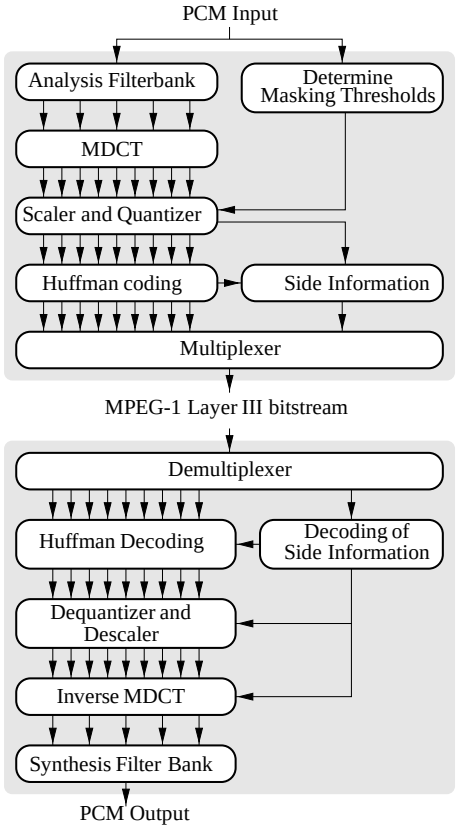

Figure 2: MPEG-1 Layer III encoder and decoder.

3. AUDIO CODING

The Moving Picture Experts Group (MPEG) has developed two of the first international high-quality audio-visual coding standards, known as MPEG-1 and MPEG-2. Both these standards include high-quality stereophonic audio coders and MPEG-2 even includes multi-channel audio coding. These features are needed in todays and tomorrows interactive multi-media applications.

The MPEG-1 Layer III audio coder, the most advanced audio coder in MPEG-1, typically compresses stereophonic audio up to 12 times with insignificant audio quality loss. It is included in communications standards as H.310 Broadband Audio-visual Communications systems and H.323 Visual Telephone Systems and Equipment for Local Area Networks. The Layer III coder is also commonly used as a high-quality audio coder on the World Wide Web.

The high compression ratio is possible by removing components of the source signal that are perceptually irrelevant to the ear. In *Simultaneous masking*, a large frequency component will mask smaller ones in a nearby frequency band, whereas in *temporal masking*, components just before or right after (in the time domain) a large audio component are masked. Using this knowledge, the audio-encoder dynamically estimates the *global masking thresh-*

old, that describe the just noticeable distortion as a function of frequency and time segment [9], Fig. 3.

The actual audio encoder operates in parallel with the global mask estimation algorithm. The audio source signal is decomposed into 32 critically downsampled bandpass signals by a filter bank. The frequency resolution is increased by processing each bandpass signal with a Modified Discrete Cosine Transform (MDCT) in the Layer III coder. Depending on the input signal, each bandpass signal is decomposed in either 6 or 18 MDCT components, where the shorter window (generating 6 MDCT components) may be used during transients in the audio source. After this decomposition the MDCT components are scaled and quantized [10]. The main key in perceptual coders is to select enough quantization levels in the every subband, so that the introduced quantization noise level is below the global masking threshold. Data redundancy is reduced by Huffman coding the signal before transmitting it over the channel, Fig. 2. The decoder operates almost like an encoder in reverse, as illustrated in Fig. 2.

When the channels are not identical, the quantization noise introduced to the two channels are almost independent. As a result, the correlation between the two channels is decreased.

Correlation between the two channels can be decreased even more if independent noise is added to the channels. Due to large overhead, every single DCT-band cannot be optimally quantized. Instead they are divided into five regions, with specific numbers of quantization levels. Define quantization-noise to mask ratio (QMR) as the difference between the level of quantization noise and the level where distortion may become just audible in a given MDCT band. Then it is possible to add non-perceivable noise in those MDCT bands where QMR is positive. That is, for all MDCT bands in the frequency region where the channel correlation need to be reduced, perform

$$\begin{aligned} \mathrm{QMR}(j) > 0 \quad &\Rightarrow \quad \tilde{X}^j_{\mathrm{MDCT}} = X^j_{\mathrm{MDCT}} + f(\mathrm{QMR}(j)) \cdot v \\ \mathrm{QMR}(j) < 0 \quad &\Rightarrow \quad \tilde{X}^j_{\mathrm{MDCT}} = X^j_{\mathrm{MDCT}} \end{aligned}$$

where X^j_{MDCT} is the MDCT component in band j and $f(\cdot)$, given by the global masking threshold, amplifies the noise component v to be added, Fig. 3. A block implementing this channel decorrelation is added to the decoder right before the Inverse MDCT, Fig. 2. The global masking information is not available in the decoder, but because of the high frequency resolution of the MDCT, a simplified global masking estimate can be calculated with low complexity.

4. MEASUREMENT STUDIES

The influence audio coding has on the convergence of the SAEC is exemplified by simulations. The far-end speech is recorded in a room of size $330 \times 500 \times 272$ cm, having a reverberation time of 0.3 seconds. The two far-end microphones are positioned 60 cm apart. Recordings were made using 48 kHz sampling rate.

To be able to access the misalignment of the estimated echo paths we use synthetic near-end room responses. The lengths of these responses are $N = 4096$ samples when the sample rate is 16 kHz and they have been estimated using data from a room of size $460 \times 670 \times 272$ cm. Echo cancellation is performed using a sample rate of 16 kHz where the length of the filters are $L = 2048$ each. No ambient near-end noise is added in the simulations.

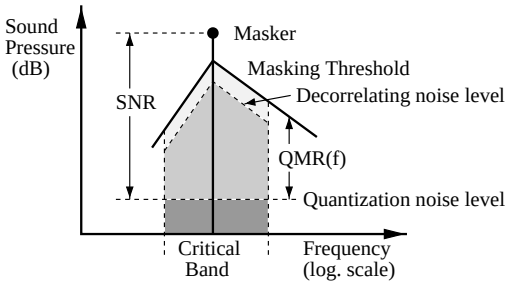

Figure 3: Masking Threshold: The grey area is masked by the tone.

As adaptive algorithm the two-channel FRLS, [11], is used, which is in principle the same algorithm as in [2] without numerical stabilization. The algorithm, [11], has been modified in order to remain stable when speech is used as input. The MPEG-1 Layer III coder used can be found in [12].

Results from four far-end cases, original recording, MPEG Layer III encoded/decoded speech, MPEG Layer III encoded/decoded with modified decoder, non-linearly modified far-end, $\alpha = 0.5$, are shown. The last case is shown as a reference since this technique has been proved effective, [5]. The MPEG coder is set to produce a bitrate of 192 kbits/s at a compression ratio of 8:1.

Possible ill-conditioning of the correlation matrix is indicated by the coherence function. Figure 4 shows the coherence between the far-end channels of the four cases. It is clearly seen that introduction of a non-linearity as well as audio coding results in smaller coherence, especially at higher frequencies. However, the coherence is still fairly close to one in lower frequencies. By modifying the decoding procedure according to Section 3 the coherence can be made smaller also in the lower frequencies without noticeable distortion, Fig. 4c.

Convergence of the FRLS algorithm is shown in Fig. 5. The convergence is presented in Fig. 5c as the Echo Return Loss Enhancement (ERLE) and in Fig. 5d as misalignment. In all cases the algorithm achieve the same high ERLE. The misalignment is on the other hand highly dependent on preprocessing of the far-end speech.

5. CONCLUSIONS

A perceptual audio coder indeed improves the ability of an SAEC algorithm to converge to the true solution. Further improvement of the misalignment can be made by decoding the data utilizing the QMR-margin that often exists after coding. Very good perceptual quality is achieved while the complexity is maintained low since only a few simple operations need to be added in the decoder.

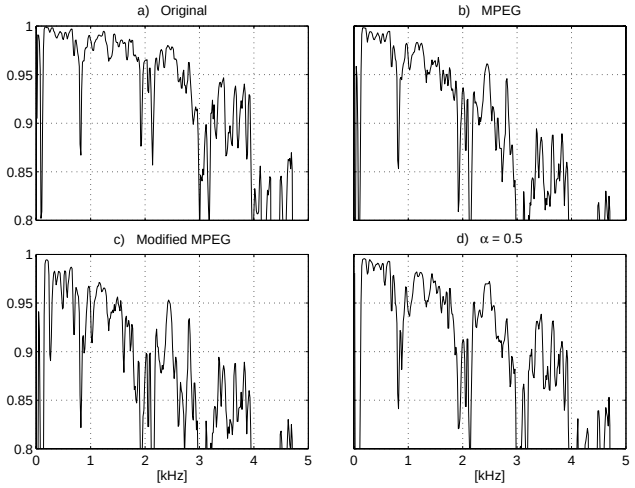

Figure 4: Coherence of four studied cases.

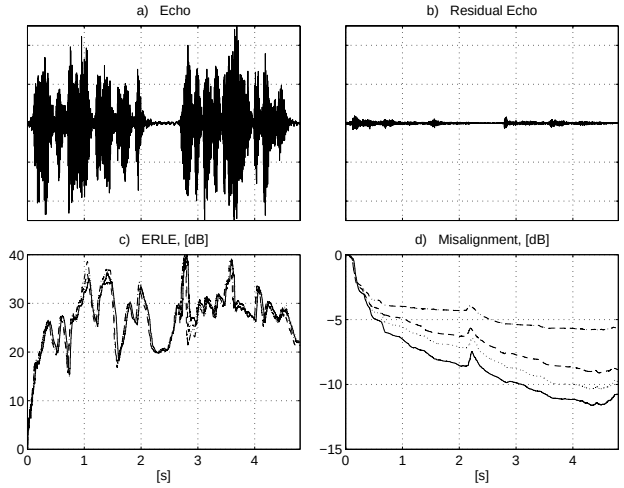

Figure 5: (a) Echo. (b) Residual echo. (c) ERLE. (d) Misalignment of the SAEC. Dashed-dotted line: Result from original far-end speech. Dashed line: Non-linearly modified. Dotted line: MPEG encoded/decoded. Solid line: MPEG encoded/decoded with modification. Each adaptive filter has 2048 taps.

Acknowledgments

The authors thank N. Johansson, B. Rodger and O. Till, Telia Research AB, for supplying measured data and performing listening tests. J. Benesty should also be acknowledged for sharing his paper [5].

6. REFERENCES

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall. Stereophonic acoustic echo cancellation - an overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8):148–151, 1995.

[2] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier. Adaptive filtering algorithms for stereophonic acoustic echo cancellation. In *Proc. of ICASSP*, pages 3099–3102. Telecom Paris, 1995.

[3] S. Gay. Algorithms for acoustic echo cancellation. Keynote Talk, The International Workshop on Acoustic Echo and Noise Control, September 1997.

[4] F. Amand, A. Gilloire, and J. Benesty. Identifying the true echo path impulse response in stereophonic acoustic echo cancellation. In *Proc. of EUSIPCO*, pages 1119–1122. Politecnico di Milano, 1996.

[5] J. Benesty, D. R. Morgan, and M. M. Sondhi. A better understanding and an improved solution to the problems of stereophonic acoustic echo cancellation. *IEEE Trans. on Speech and Audio Processing*. To appear. A short version can be found in Proc. of ICASSP 1997 pages 303-306.

[6] J. Benesty, P. Duhamel, and Y. Grenier. A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation. *IEEE Signal Processing Letters*, 3(2):35–37, February 1996.

[7] S. Makino et. al. Subband stereo echo canceller using the projection algorithm with fast convergence to the true echo path. In *Proc. of ICASSP*, pages 299–302. NTT, 1997.

[8] O. Till, B. Rodger, and N. Johansson. Informal listening test of stereophonic perception after non-linear tranformation. Unpublished work, Telia Research AB, May 1997.

[9] P. Noll. MPEG digital audio coding. *IEEE Signal Processing Magazine*, 14(5):59–81, September 1997.

[10] B. G. Haskell, A. Puri, and A. N. Netravali. *Digital Video: An Introduction to* MPEG-*2*, chapter 4, pages 55–79. Digital Multimedia Standards Series. Chapman & Hall, 1st edition, 1997. http://www.thomson.com/.

[11] M. G. Bellanger. *Adaptive Digital Filters and Signal Analysis*. Marcel Dekker, 1987. ISBN: 0-8247-7784-0.

[12] MPEG-1 LAYER III shareware audio coder, 1995. Am Weichselgarten 3 D-91058 Erlangen Germany, Encoder and decoder code: http://www.iis.fhg.de/departs/amm/layer3/, Public domain decoder source code (ANSI c): ftp://ftp.fhg.de/pub/iis/layer3/public_c/.