PERCEPTUALLY-MOTIVATED NONLINEAR CHANNEL DECORRELATION FOR STEREO ACOUSTIC ECHO CANCELLATION

Jean-Marc Valin

CSIRO ICT Centre, Sydney, Australia
jean-marc.valin@csiro.au

ABSTRACT

Acoustic echo cancellation with stereo signals is generally an underdetermined problem because of the high coherence between the left and right channels. In this paper, we present a novel method of significantly reducing inter-channel coherence without affecting the audio quality. Our work takes into account psychoacoustic masking and binaural auditory cues. The proposed non-linear processing combines a shaped comb-allpass (SCAL) filter with the injection of psychoacoustically masked noise. We show that the proposed method performs significantly better than other known methods for reducing inter-channel coherence.

Index Terms— stereo acoustic echo cancellation, non-linear audio processing, all-pass filters, psychoacoustic masking

1. INTRODUCTION

As videoconferencing applications incorporate higher sampling rates and multiple channels, the problem of cancelling acoustic echo becomes harder. One of the main difficulties in stereo echo cancellation is the strong coherence that exists between the left and right channel, making it hard or even impossible to correctly estimate the acoustic impulse response [1].

To improve the performance of a stereo acoustic echo canceller, it is very useful to reduce the coherence between channels [1, 2]. This can be done by altering the signals using some form of nonlinear transformation, as illustrated in Fig. 1. However, most of the methods proposed so far to reduce inter-channel coherence [3, 4, 5] do not take into account human perception and tend to introduce too much audible distortion to the signal, especially at high sampling rates. In this paper, we propose a non-linear processing that closely matches human perception in order to minimise coherence while maximising audio quality. We show that by combining a shaped comb-allpass filter and psychoacoustically masked noise, it is possible to achieve better reduction in coherence while preserving higher audio quality than other nonlinear algorithms.

The paper is divided as follows. Section 2 presents an overview of the stereo acoustic echo cancellation problem. Sections 3 and 4 describe the two parts of the algorithm, which are the all-pass filtering and the noise injection, respectively. Section 5 presents comparative results and Section 6 concludes this paper.

2. OVERVIEW

In a multi-channel audio system, there usually exists a strong coherence between channels (loudspeakers) that causes the filter optimisation problem to be ill-conditioned. It is shown in [6] that the normalised misalignment $\eta(n)$ is inversely proportional to $\left(1-\gamma^2\right)$,

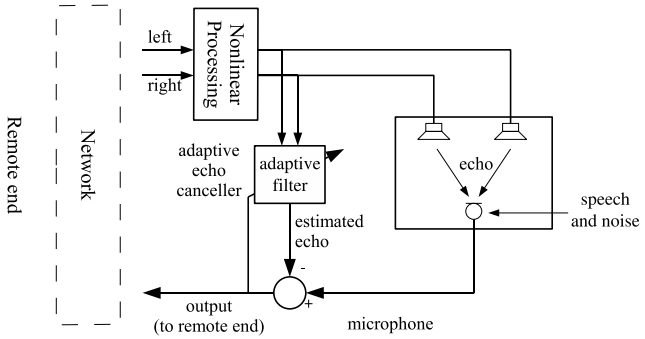

Fig. 1. Stereo echo cancellation system.

where γ is the inter-channel coherence. In the frequency domain, the square inter-channel coherence is defined as [1]

$$\gamma^2(f) = \frac{\left|E\left\{X_1^*(f)\,X_2(f)\right\}\right|^2}{E\left\{\left|X_1(f)\right|^2\right\}E\left\{\left|X_2(f)\right|^2\right\}}, \qquad (1)$$

where $X_j(f)$ denotes the Fourier transform of channel j, $E\{\cdot\}$ denotes the mathematical expectation, and $(\cdot)^*$ denotes the complex conjugate.

It is desirable to maximise the audio quality, while minimising the inter-channel coherence. Several approaches proposed so far to reduce inter-channel coherence have focused on using memoryless non-linearities [3]. Although they have the main advantage of being easy to compute, these non-linearities introduce a great amount of inter-modulation distortion, which quickly degrades the audio quality. They also provide little control regarding how much perturbation is caused as a function of frequency.

Another popular approach is to alter the phase of the signal in a time varying way [4, 5]. The time-varying aspect of the transformation is important because the transformation would otherwise be linear and thus have no effect on inter-channel coherence. The phase of an audio can be altered either through the use of an all-pass filter, or in the short-term Fourier transform (STFT) domain.

The algorithm we propose in this work was designed to minimise inter-channel coherence, while maintaining good quality audio, including the stereo image. Additionally, it is important not to add any significant delay to the audio because latency is a very important aspect in the perception of acoustic echo. Although this rules out analysis/synthesis algorithms based on the DFT, it still allows the use overlapping windows, as long as the processing within each window is causal.

The human auditory system localises sounds using two sets of binaural cues. Interaural phase difference (IPD) is used at low frequencies ($\leq 1.5\,\mathrm{kHz}$) and while interaural intensity difference (IID)

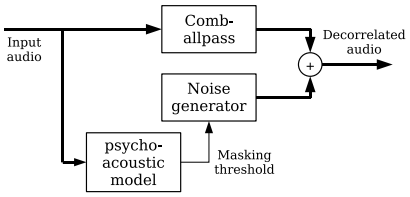

Fig. 2. Overview of the proposed algorithm (for each channel).

is used at higher frequencies ($\geq$ 2 kHz). We propose a strategy that takes into account these binaural cues when altering the audio signals. At higher frequencies, we propose to reduce coherence by altering the phase (IPD) while preserving the IID. Because it is not possible to alter IID without altering IPD at lower frequencies, we use wideband psychoacoustically-masked noise with emphasis on lower frequencies. Note that, unlike [1], the noise is not only shaped, but predominantly added to the lower frequencies. The approach is illustrated in Fig. 2.

3. SHAPED COMB-ALLPASS (SCAL) FILTERING

Allpass filters have a flat frequency response with non-linear phase and can be represented by the general causal form

$$A(z) = \frac{\sum_{k=1}^{N} a_k z^{k-N} + z^{-N}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \ . \tag{2}$$

It is hard to design an all-pass filter starting from the general form (2) for high orders. However, it is possible to construct a filter that alters the phase similarly across all frequencies by using a simple comb-allpass filter of the form

$$A(z) = \frac{\alpha + z^{-N}}{1 - \alpha z^{-N}} \ . \tag{3}$$

The filter in (3) combines an all-pole comb filter to a maximum-phase all-zero comb filter, so the poles and zeros are equally spread around the unit circle with radii of respectively $\alpha^{1/N}$ and $\alpha^{-1/N}$.

For the processing to be non-linear, it is required to vary the coefficient α controlling the filter. This is achieved through using overlapping windows with α held constant over each window. We use both an analysis window and a synthesis window to prevent any blocking artifacts. The signal is reconstructed using weighted overlap-add (WOLA). Because all-pass filtering is a time-domain process, no extra delay is added because at any given time, we do not need to apply the allpass filter on the whole window. The analysis-synthesis window is required to meet the Princen-Bradley criterion [7] and we use the Vorbis window [8].

Interaural phase difference (IPD) is the main localisation cue at lower frequencies, so the human ear is more sensitive to phase distortion in the low frequencies. For that reason, it is important to "shape" the phase modulation as a function of frequency so as to limit distortion of the stereo image. It is desirable to introduce less distortion to the phase at lower frequencies than at higher frequencies. To do so, we propose a shaped comb-allpass (SCAL) filter of the form

$$A(z) = \frac{\alpha\left(1 - \beta z^{-1}\right) + z^{-N}}{1 - \alpha\left(-\beta z^{-N+1} + z^{-N}\right)} \ , \tag{4}$$

where α controls the *depth* of the filter and β controls the *tilt*. Stability is guaranteed (sufficient condition) as long as

$$|\alpha|\left(1 + |\beta|\right) < 1 \ , \tag{5}$$

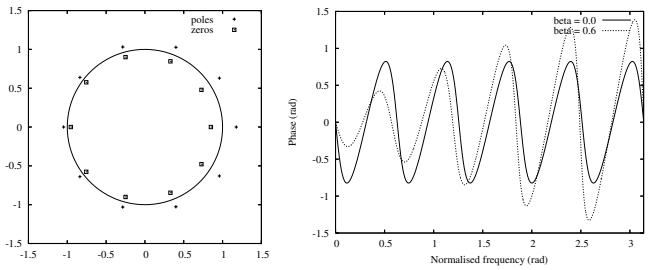

Fig. 3. Effect of the *tilt* parameter β (with $\alpha = 0.4$, $N = 10$). Location of the poles and zeros (left) and phase response for different values of β (right).

so (5) can be used to determine the upper bound on α as a function of β. The effect of the *tilt* parameter β is demonstrated in Fig. 3 and can be explained by the fact that as β increases, the poles and zeros of the all-pass filter move closer to the unit circle at high frequencies and away from the unit circle at lower frequencies.

When using a filter of order N, there are $N + 1$ points on the frequency axis where the phase response is zero, regardless of α. In other words, there are frequencies where no coherence reduction occurs. For this reason, we also vary the order N of the filter so that the "nulls" in the phase response change as a function of time.

The order N is changed randomly for each new window, subject to $N_{min} \leq N \leq N_{max}$. We then vary α using

$$\alpha(N) = \min\left(\left(\alpha(N-1) + r_0\right), \frac{1-\epsilon}{1+|\beta|}\right) \ , \tag{6}$$

where r_0 is a uniformly-distributed random variable chosen in the $[-r_{max}, r_{max}]$ range (typically $r_{max} = 0.6$) and $\epsilon \ll 1$ controls the distance to the unit circle of the high frequency poles. The SCAL filter has a delay of $N_{max} = 10$ and an overall complexity of only 23 operations per sample, which is negligible when compared to the complexity of the adaptive filtering used to cancel the echo.

4. PSYCHOACOUSTICALLY-MASKED NOISE

The SCAL processing in Section 3 is mainly effective for frequencies above 2 kHz. For lower frequencies, the ear is more sensitive to phase distortion (altering stereo image), so it is preferable to inject noise that is uncorrelated to the audio signal. In this work, we use the psychoacoustic model from the Vorbis audio codec, as described in [9]. The output of the psychoacoustic model determines the amount and spectral shape of the noise that can be added without significantly altering perceptual audio quality. The psychoacoustic model is also tuned to introduce less noise in higher frequencies because those are already decorrelated by the SCAL filter.

The noise to be added is generated in the frequency domain. Again, we make use of weighted overlap-and-add to reconstruct the time-domain signal. To avoid adding a delay to the signal of interest, only the noise is delayed and is added to the non-delayed input signal. This can be done without significant audio quality degradation because of the temporal masking effect. Because we are adding a random signal during the WOLA process, it is the power that is additive and not the amplitudes. For that reason, we again need to use a window that satisfies the Princen-Bradley criterion, even though it is only applied once. Although a lossy codec could be used [10] to add the noise, it would cause a significant increase in the total delay (>100 ms with MP3).

5. EVALUATION AND RESULTS

We compare four different coherence reduction algorithms:

- Proposed algorithm, with shaping and variable order (**SCAL**), $\beta = 0.43$, $5 \leq N \leq 10$
- Proposed algorithm, without shaping or variable order (**Comb-allpass**), $\beta = 0$, $N = 7$
- Smoothed absolute value (**smoothed absolute**), $\alpha_{abs} = 0.3$
- First-order, time varying all-pass filter (**allpass**), $\alpha_{min} = -0.985$

The smoothed absolute value non-linearity is included because it was shown in [3] to be among the best memoryless non-linearity. The time-varying first order all-pass filter is implemented as described by [4] but using $\alpha_{min} = -0.985$ to account for the different sampling rate used in this work.

The block-based phase alteration method proposed in [5] is excluded from the comparison because the boundary artifacts caused by the implicit rectangular window causes major quality degradation at high sampling rate, even for very small amounts of decorrelation. While a WOLA approach could be used, it would involve additional delay, something which is not acceptable in this context.

In both the proposed algorithm and the first-order all-pass filter, there is a random component, so it is possible to independently process each channel with the algorithm. On the other hand, applying the same memoryless nonlinearity (smoothed absolute value in this case) to each channel would not reduce the coherence. For that reason, we invert the sign of the gain α_{abs} used for each channel.

5.1. Methodology

We evaluate the algorithms on eight audio excerpts sampled at 44.1 kHz. Four of the samples are voice samples (male and female speech, Suzanne Vega, quartet) taken from the EBU Tech 3253 - Sound Quality Assessment Material (SQAM), while the other four are various music samples (classical, folk, rock, castanets).

Because we need the ground truth, all recordings are simulated using real impulse responses measured in a room with around 220 ms reverberation time (-60 dB). On the remote end, the samples are played one meter in front of an XY-stereo microphone (16,384-sample impulse response). These first recordings are processed using each of the algorithms in the comparison. The quality of the processed files is evaluated using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [11] methodology[1]. Although in the final application the listeners hear the audio through loudspeakers, we have decided to perform the evaluation using headphones to make artifacts – especially stereo image artifacts – more noticeable.

The processed signals are played on the near end at the back of another XY-stereo microphone (16,384-sample impulse response) and some background noise is added to obtain a signal-to-noise ratio (SNR) of 40 dB. Stereo echo cancellation is then performed using a variant of the MDF algorithm [12] for a 8,192-sample filter length. The filter misalignment is measured between the filter found by the echo canceller and the first 8192 samples of the real impulse response.

[1]We used the RateIt graphical interface available at http://rateit.sf.net/

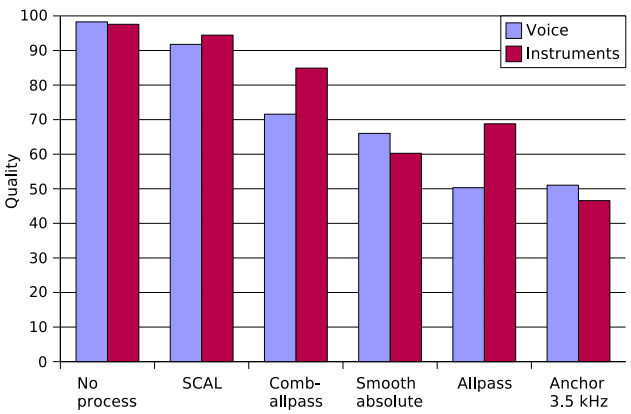

Fig. 4. Quality of the signal from different algorithms obtained from a MUSHRA test (higher is better).

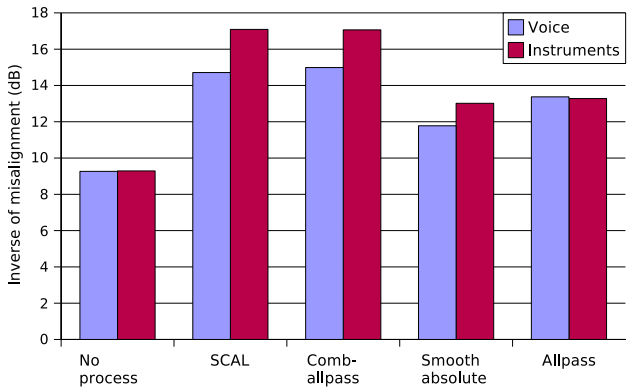

Fig. 5. Inverse of the filter misalignment (higher is better).

5.2. Quality and Misalignment

Fig. 4 shows the results we obtained with 10 listeners[2] on a MUSHRA quality comparison of the algorithms. We observe that for both voice and instruments, the quality with the proposed (SCAL) algorithm is very close to that of the reference ("no process") and significantly higher than all other algorithms, with greater than 99% confidence on a permutation test. This also shows that the shaped aspect (parameter β) of the comb-allpass filter is important, since the simple comb-allpass filter has significantly lower quality.

Fig. 5 shows the misalignment obtained using each algorithm after 10 seconds. We can observe that convergence with the proposed method is a significant improvement over the smoothed absolute value, the first-order allpass filter and the unprocessed signal. The total misalignment is approximately the same as with the simple comb-allpass filter.

Lastly, the average inter-channel coherence is shown in Fig. 6 as a function of frequency for a male speech sample. We see that the proposed algorithm is more constant than other algorithms. It can also be observed that for the comb-allpass filter with a constant order some frequencies are still highly correlated. This is due to "nulls" in the phase response that do not depend on the value of α.

[2]Results from one additional listener were removed during post-screening because they were inconsistent (e.g. all algorithms rated 0 including the reference).

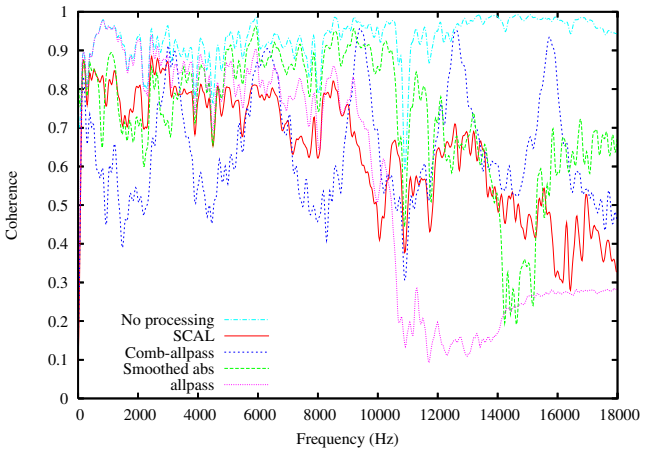

Fig. 6. Inter-channel coherence after processing (lower is better)

5.3. Qualifying Algorithm Artifacts

Listening to the samples makes it possible to qualify the artifacts caused by the algorithms under test and make the following remarks.

Proposed algorithm

Some listeners reported a mild "flanging" effect when listening to the proposed algorithm, along with slight movement of the stereo image from left to right. Overall, the artifacts are less severe than those of the other algorithms and the stereo image distortion would likely become less noticeable when listening with loudspeakers.

Proposed algorithm, without shaping and variable order

When we remove the shaping of the comb-allpass filter, the flanging effect becomes worse and the distortion in stereo image becomes much more noticeable. This is due to increased changes in phase below 1 kHz, which affects the interaural phase difference (IPD). That degradation is most noticeable on the well known Suzanne Vega *a cappella* excerpt.

Smoothed absolute value

Being a non-linear function, the smoothed absolute value produces inter-modulation distortion. On harmonic signals such as speech, the distortion is mostly masked and is perceived as additional "harshness". Another artifact can be observed in the castanets sample. Because castanets produce strongly asymmetric time-domain impulses, the smoothed absolute value causes one of the channels to be amplified more than the other, resulting in a very disturbing "bouncing" stereo image. On tonal non-harmonic signals such as the glockenspiel (not included in the formal evaluation), the inter-modulation distortion effect can cause new tones to appear in some regions of the spectrum. Some tones even appear at low frequencies, which has the effect of changing the perceived fundamental frequency.

First-order all-pass filter

The main artifact introduced by the first-order all-pass filter is a nearly white crackling noise that is the result of varying the filter coefficient α from one sample to another. For most samples, the noise is masked at lower frequency, so it is usually perceived as a high-frequency crackling noise. It is mainly perceivable on very tonal samples, that do not leave much room for masking noise components.

6. CONCLUSION

In this paper, we have demonstrated that it is possible to reduce the coherence between the left and right channels in a video-conference application without significantly reducing the audio quality. The proposed method includes a shaped comb-allpass (SCAL) filter to reduce coherence at higher frequencies and psychoacoustically masked noise injection at lower frequencies. Novel aspects of this work include the shaping of the phase alteration to better match human stereo perception, as well as the use of windowing the allpass filter output to prevent blocking artifacts.

The proposed method was shown to outperform other existing methods both in terms of quality and amount of decorrelation provided, leading to better echo cancellation results. Moreover, the total complexity of the proposed algorithm is kept small so that it does not significantly increase the complexity of a complete echo cancellation system.

7. REFERENCES

[1] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 156–165, 1998.

[2] M. M. Sondhi, D. R. Morgan, and J. L Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, 1995.

[3] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. SAP*, vol. 9, no. 6, 2001.

[4] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," in *Proc. ICASSP*, 1998, pp. 3689–3692.

[5] M. Wu, Z. Lin, and X. Qiu, "A frequency domain nonlinearity for stereo echo cancellation," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences – Letters*, pp. 1757–1759, 2005.

[6] T. Gansler and J. Benesty, "New insights to the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, 2002.

[7] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. ASSP*, vol. 34, no. 5, pp. 1153 – 1161, 1986.

[8] C. Montgomery, "Vorbis I specification," 2004, `http://www.xiph.org/vorbis/doc/Vorbis_I_spec.html`.

[9] J.-M. Valin and C. Montgomery, "Improved noise weighting in CELP coding of speech – applying the Vorbis psychoacoustic model to Speex," in *Proc. 120th AES Convention*, 2006.

[10] T. Gansler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proc. ICASSP*, 1998.

[11] ITU-R, *BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems*, 2001.

[12] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. ASSP*, vol. 38, no. 2, 1990.