Developments of a Hybrid Pre-Processor Based on Frequency Shifting for Stereophonic Acoustic Echo Cancellation

Bruno C. Bispo, Diamantino da S. Freitas

Department of Electrical and Computer Engineering, University of Porto - Porto, Portugal
bruno.bispo@fe.up.pt, dfreitas@fe.up.pt

Abstract
In a multi-channel hands-free communication, decorrelation algorithms are necessary, along with adaptive filters, to efficiently remove the acoustic echo in real time applications. But the quality of the signals and the spatial position of the sound source must not be perceptually affected by those algorithms. This paper proposes a new hybrid solution that uses frequency shifts to improve the performance of a state-of-art solution based on addition of half-wave rectified signals. The results show that the new method achieves a significant improvement in the identification process of the real echo paths as well as in the global perceptual quality of the processed signals.

Index Terms: stereophonic acoustic echo cancellation, frequency shifting, speech decorrelation

1. Introduction
In hands-free communication, acoustic coupling between loudspeakers and microphones is inevitable and causes that, after talking, the user receives back his own voice with delay. The occurrence of this acoustic echo is annoying and must be eliminated or, at least, attenuated.

The use of adaptive filters in acoustic echo cancellation has been established during the last decades. But in a stereophonic acoustic echo cancellation (SAEC) situation, as illustrated in Figure 1, it can be shown that a theoretical non-uniqueness of the solution provided by the adaptive filters exists [1].

Fortunately in real scenarios with finite length adaptive filters, this non-uniqueness problem is avoided. However, a bias is introduced in the filter coefficients due to the strong correlation between the channels’ signals if they are originated from the same sound source [2]. As consequence, the adaptive filters $f_{1,2}$ generally converge to a solution that does not correctly match the real echo paths $f_{1,2}$ of the reception room.

To overcome this bias problem, a pre-processing technique is built into the SAEC system to decorrelate the channels’ signals before the application of the adaptive filters. Nevertheless, the pre-processing techniques must not insert perceptible degradations, including modifications in the spatial image of the sound source, since the signals will be played through the loudspeakers in the reception room, while keeping a low complexity to be applied in real time systems.

2. Proposed Hybrid Solution
2.1. Half-wave rectifier technique
The half-wave rectifier (hwr) technique adds a positive half-wave rectified version of the signal in one channel and a negative version in the other according to [2]

$$x'(n) = x(n) + \alpha \left( \frac{x(n) + |x(n)|}{2} \right), \quad (1)$$

where $\alpha$ is a parameter that controls the level of the added nonlinearity. This method is able to obtain a good performance while the stereo perception is not affect even with
α = 0.5 [2]. It is worth mentioning that it is also necessary to remove the DC level.

Preliminary tests showed that, despite achieving a reasonable decorrelation in the lower frequencies, the method may not decorrelate properly the higher frequency components depending on the impulse responses of the transmission room and the source signal. This may occur due to the fact that the power of speech signals is concentrated in the lower frequencies [5].

2.2. Frequency shifting technique

The frequency shifting (fs) technique was initially proposed to increase the stability margin of public address systems by Schroeder in 1962, and evaluations of this method in this context have already been made [6].

In a SAEC scenario, an interchannel frequency shifting was already tried such that a channels’ signal was shifted in frequency relative to the other, but this caused a quite perceivable destruction of the stereo perception of the signals [1]. Preliminary listening tests confirmed this effect since the position of the sound source seemed to oscillate in function of the applied frequency offset. But the ability of this technique to decorrelate the channels’ signals was quite high, thereby stimulating more attention and analysis.

It was understood that a frequency shift is critically perceived in the low frequencies of stereophonic images since, in this range, the human perception of the azimuthal position of sound sources is highly dependent on the interaural time difference [7]. On the other hand, this dependence gradually reduces with increasing frequency until it vanishes [4, 7]. Informal listening tests showed that a small frequency shift in higher frequencies is difficult to be perceptually detected and still produces a substantial decorrelation between the channels’ signals in the frequency range where it was applied.

A digital frequency shifter can be implemented using cosine and sine as modulation functions along with a Hilbert filter following the scheme presented in Fig. 2 [6], where ω0 is the desired frequency shift value. The impulse response of the Hilbert filter can be calculated according to

\[
h_{hil,k} = \begin{cases} 0, & \text{if } k \text{ is even}, \\ \frac{2}{k\pi}, & \text{else}. \end{cases} \tag{2}
\]

Due to its infinite length, this response must be truncated to a range \( k = -N_{hil} \) to \( k = N_{hil} \) using a window function, resulting in a Hilbert filter \( \hat{h}_{hil} \) with length \( L_{hil} = 2N_{hil} + 1 \). Moreover, to avoid non-causality it is necessary to shift the truncated solution by \( N_{hil} \) coefficients and, consequently, to delay the cosine modulated signal by \( N_{hil} \) samples.

2.3. Hybrid techniques

Due to the above considerations, a new hybrid technique emerged in order to combine the strengths of both solutions. The hybrid configuration uses a filter bank to apply a specific decorrelation method in each sub-band of the signals’ spectrum.

Among many possible combinations, two hybrid configurations were chosen to initially evaluate the effect of frequency shifts to the bias problem of the SAEC. Considering 8 kHz band-limited speech signals, the techniques to be evaluated are summarized in Table 1.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Spectrum band</th>
</tr>
</thead>
<tbody>
<tr>
<td>hwr</td>
<td>0-2 kHz</td>
</tr>
<tr>
<td>hwr</td>
<td>2-4 kHz</td>
</tr>
<tr>
<td>hwr</td>
<td>4-8 kHz</td>
</tr>
</tbody>
</table>

The fs technique applied a positive frequency shift in one channel and a negative in the other, and it used \( N_{hil} \) equivalent to 20 ms. Because of this intrinsic delay from the fs algorithm, in the sub-bands of the hybrid methods where the hwr were applied, the signals had to be properly delayed.

3. Experiments configurations

To assess the relative performances of the above techniques in a SAEC system, a first experiment measured their abilities to decorrelate the channels’ signals and, consequently, to improve the echo path estimates provided by the adaptive filters using two quantitative metrics. A second experiment evaluated the audible distortions inserted by the methods in speech signals using a standardized subjective test. For this purpose, the following configuration was used.

3.1. Environment setup

In order to simulate a teleconference environment, measured impulse responses \( g_{1,2} \) and \( f_{1,2} \) with lengths \( L_G = L_F = 4000 \) samples and downsampling to \( f_s = 16 \) kHz were used.

The adaptive filters \( \hat{f}_{1,2} \) were GSFAP algorithms [8] with 20 projections and \( L_F = 2000 \) samples. Their step-size and regularization parameters were optimized by minimizing the average misalignment. An echo-to-noise ratio of 30 dB and a voice activity detector were also applied to simulate real world conditions and to avoid adaptation in presence of only noise.
3.2. Quantitative metrics

3.2.1. Coherence function

In reference [2], a link is established between the conditioning of the covariance matrix of the signals $x_1'$ and $x_2'$ and the coherence function

$$\gamma(f) = \frac{S_{x_1'x_2'}(f)}{S_{x_1'x_1'}(f)S_{x_2'x_2'}(f)},$$

(3)

where $S_{x_1'x_2'}(f)$ is the cross-power spectral density of signals $x_1'$ and $x_2'$. In practice, the coherence function is used to evaluate the cross-correlation between two signals and, hence, works as a metric of the decorrelation algorithms efficiency.

For each signal section of 2000 samples taken with 50\% overlap, the power spectral densities were estimated using a FFT with 320000 points and zero-padding in order to achieve a resolution of 0.05 Hz per bin, so that small values of $\omega$ could be evaluated.

3.2.2. Misalignment

The performance of the adaptive filters, and also of the decorrelation methods, were measured by the normalized misalignment defined as [2]

$$\text{mis}(n) = \sum_{k=1}^{2} \frac{||f_k(n) - \hat{f}_k(n)||}{||f_k(n)||}.$$  

(4)

3.3. Qualitative metric

The perceived quality of the processed signals was evaluated by the standardized subjective listening test denominated Multi Stimulus test with Hidden Reference and Anchor (MUSHRA)[9].

In this subjective test, the evaluators should assess the processed signals, a hidden reference and a 3.5 kHz band-limited reference (anchor) comparing them with the reference signal according to the scale presented in Fig. 3. The listening test was performed by a balanced group of 10 evaluators where half of them were experienced listeners.

![Figure 3: Grading scale of the MUSHRA test.](image)

3.4. Speech database

The speech database used in the simulations is formed by a total of 100 signals recorded by 10 different talkers (10 signals per talker). Each signal consists of one short sentence with duration of 4 s and original sampling rate of 48 kHz but downsampled to 16 kHz. The average power spectral density of the signals is showed in Fig. 4.

4. Experiments Results

4.1. Experiment 1

With regards to performance assessment, Fig. 5 exhibits the average coherence function for each decorrelation technique. It illustrates the poor performance that the hwr method can obtain in the higher frequencies and the improvements achieved by the proposed hybrid schemes using frequency shifts.

As can be observed in Fig. 5(b), the coherence values for the hwr method are quite near unity in the higher frequencies, showing some useful decorrelation only in the lower frequencies. In Fig. 5(c), the good effect of the fs technique can already be noticed in the highest sub-band (above 4 kHz), where coherence values approximately equal to half of the ones obtained with the hwr method are achieved. In the hybrid2 method, the superiority of the fs technique in terms of decorrelation is extended to the middle sub-band (2-4 kHz), as illustrated in Fig. 5(d).

Fig. 6 confirms the last results showing the misalignment obtained by the 3 evaluated methods in the SAEC system. Due to its greater capacity to decorrelate the channels’ signals, the hybrid2 technique allowed a gain of 4 dB compared to the original hwr method what means that the proposed technique achieved solutions closer to the impulse responses of the real echo paths.

It should be mentioned that the absolute value of the misalignment is also highly dependent on the convergence speed of the adaptive algorithm. However, a reduced number of tests using a fast RLS adaptive algorithm [2], and shorter signals, corroborated that the hybrid techniques always surpass the hwr method in terms of the relative misalignment.

4.2. Experiment 2

In experiment 2, the global quality of the processed stereo signals was subjectively evaluated using the standardized
listening test described previously. The average grades with a 95% confidence interval for each speech signal and for each decorrelation technique are depicted in Fig. 7. The hidden reference and the anchor are also included as well as the overall average grade.

The results showed that both proposed hybrid methods outperformed the hwr method with a slight average advantage for the hybrid2 technique. Since the difference between the processed signals reside only in the higher frequencies, it can be concluded that the distortions inserted by the hwr method in this frequency range are more sensitive to the human audition than those generated by the frequency shifts.

In some of the depicted cases, the size of the 95% confidence interval is greater than the desired. This was due to the subjective nature of the test and to the restricted number of evaluators. Moreover, the use of non-experts listeners usually tends to increase the variance of the results. But even so, the results are quite significant since in all runs the new proposed methods presented an average performance superior to the hwr method.

Furthermore, the results indicate that the application of frequency shifts with an appropriate value of $\omega$ depending on the sub-band is a promising technique to decorrelate multichannel speech signals with small degradations in the global quality. Further investigation is necessary to build a decorrelation method using only frequency shifts.

![Figure 5: Coherence function using: a) no decorrelation technique, b) hwr, c) hybrid1, d) hybrid2.](image)

![Figure 6: Misalignment: convergence comparison in SAEC.](image)

![Figure 7: MUSHRA results: quality comparison.](image)

5. Conclusions

Shifts in the entire spectrum had already been tried to decorrelate the channels’ signals in a stereophonic acoustic echo cancellation system, but the effect in the stereo perception was disastrous. In this work, a sub-band frequency shift scheme had its theoretical fundamentals discussed and it was used to improve the performance of the well known half-wave rectifier method.

The results showed that the application of frequency shifts with an appropriate value in the higher frequencies, instead of the half-wave rectifier, provides a better estimation of the real echo paths while obtaining processed signals with less perceptible degradations relatively to the original ones.

6. References


