# 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
Emails: {bruno.bispo, dfreitas}@fe.up.pt

## ABSTRACT

In a multi-channel hands-free communication, reasonable complexity decorrelation algorithms are necessary, along with adaptive filters, to efficiently remove the acoustic echo in real time applications. But, at the same time, 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, in a stereophonic echo cancellation environment, 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 [1] and their application is performed following the diagram illustrated, for a stereo case, in Figure 1. Assuming the existence of only one of the channels, the adaptive filter $\hat{\mathbf{f}}_\mathbf{1}$ models the real echo path $\mathbf{f_1}$ of the reception room and obtains an estimate of the echo signal $y(n)$. Then, this estimate is subtracted from the microphone signal $d(n)$ generating the error signal $e(n)$, which is effectively the signal to be transmitted and is used in the updating process of the adaptive filter.

This scheme usually works quite well in a mono-channel system. But in a stereophonic acoustic echo cancellation (SAEC) situation, it can be shown that a theoretical non-uniqueness of the solution provided by the adaptive filters exists [2].

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 [3]. As consequence, the adaptive filters $\hat{\mathbf{f}}_{\mathbf{1,2}}$ generally converge to a solution that does not correctly match the real echo paths $\mathbf{f}_{\mathbf{1,2}}$ of the reception room.

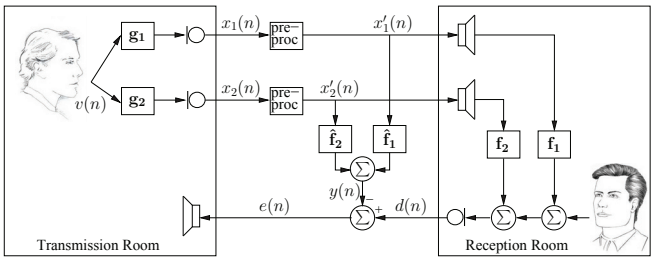


**Fig. 1**. Stereophonic acoustic echo cancellation.

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.

The first pre-processing technique, proposed by [2], adds a low level white noise signal to each channel which highly degrades the signals' quality, mainly in periods of silence. Instead of using white noises, the reference [4] applies perceptual audio coding/decoding (MPEG-1 Layer III) to add noise shaping according to the human psychoacoustic model, but the degradation of the sound quality is still perceptible.

A third method uses a half-wave rectifier function to insert non-linearities in both channels [3, 5] and is one of the most common state-of-art solutions. This method has a very low complexity and provides an improvement in the convergence of the adaptive filters, however it can generate audible distortions depending on the level of the inserted non-linearity and on the nature of the sound contents.

Recently a sub-band approach was proposed [6]. The method applies a sine wave phase modulation with constant frequency but amplitude dependent on the sub-band, and it achieves as good performances as the half-wave rectification method but with much better global perceptual sound quality.

Among other solutions, shifts in the entire spectrum of the channels's signals were already tried but the effect in the

stereo perception was disastrous [2]. However, in order to make a preliminary evaluation of a solution based on sub-band frequency shifting, the present work proposes a new hybrid approach to improve the performance of the half-wave rectification technique by combining it with frequency shifts in sub-bands of the spectrum of the channels' signals.

The paper is organized as follows: Section 2 discusses some preliminary results of the original methods and the theoretical concepts that led to the development of the proposed technique; Section 3 describes the configuration of the simulated experiments; in Section 4 the obtained results are presented and discussed. Finally, Section 5 concludes the paper emphasizing its main contributions.

## 2. PROPOSED HYBRID SOLUTION

### 2.1. Half-wave rectification technique

The half-wave rectification (HWR) technique adds a positive half-wave rectified version of the signal in one channel and a negative version in the other according to [3]

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

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 affected even with $\alpha = 0.5$ [3, 5]. It is worth mentioning that after the application of the technique it is 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 [7].

### 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 [8].

In a SAEC scenario, an interchannel FS 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 [2]. 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 [9]. On the other hand, this dependence gradually reduces with increasing frequency until it vanishes [6, 9]. 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.

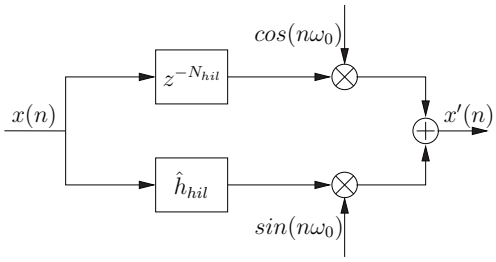


**Fig. 2**. Block diagram of the frequency shifter.

A digital frequency shifter can be implemented using cosine and sine as modulation functions along with a Hilbert filter following the scheme presented in Figure 2 [8], where $\omega_0$ is the desired frequency shift value. The impulse response of the Hilbert filter can be calculated according to

$$h_{hil,k} = \begin{cases} 0, & \text{if } k \text{ is even,} \\ \dfrac{2}{k\pi}, & \text{else.} \end{cases} \qquad (2)$$

Due to its infinite length, this response must be truncated to a range $k = -N_{hil}, \ldots, N_{hil}$ 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.

It is evident that the efficiency of this implementation of the frequency shifter depends on the length of the Hilbert filter: higher values of $N_{hil}$ provide more accurate solutions but, at the same time, insert longer delays in the output signal. Fortunately, as the more $|k|$ increases the more the filter coefficients tend to zero, $N_{hil}$ values do not need to be very large to have an accurate solution.

### 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.

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

**Table 1**. Configurations of the hybrid techniques.

| | Spectrum band | | |
|---|---|---|---|
| Technique | 0-2 kHz | 2-4 kHz | 4-8 kHz |
| HWR | HWR: $\alpha = 0.5$ | HWR: $\alpha = 0.5$ | HWR: $\alpha = 0.5$ |
| hybrid1 | HWR: $\alpha = 0.5$ | HWR: $\alpha = 0.5$ | FS: $\omega_0 = 5$ Hz |
| hybrid2 | HWR: $\alpha = 0.5$ | FS: $\omega_0 = 1$ Hz | FS: $\omega_0 = 5$ Hz |

algorithm, in the sub-bands of the hybrid methods where the HWR were applied, the signals had to be properly delayed.

With no major concerns about the computational complexity, an orthogonal two-channel filter bank with 198 coefficients was used to split the frequency bands. As consequence, the hybrid1 and hybrid2 techniques presented, respectively, a delay of 32.3 ms and 56.9 ms.

## 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 introduced 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 room impulse responses $\mathbf{g}_{1,2}$ [10] and $\mathbf{f}_{1,2}$ [11] were used. They were truncated to lengths $L_G = L_F = 4000$ samples, after being downsampled to $f_s = 16$ kHz, and are illustrated in Figure 3.

The filters $\hat{\mathbf{f}}_{1,2}$ were adapted using the Gauss-Seidel Fast Affine Projection (GSFAP) algorithm [12]with 20 projections and $L_{\hat{F}} = 2000$ samples. Their stepsize 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 [3], a link is established between the conditioning of the covariance matrix of the signals $x_1'$ e $x_2'$ and the coherence function

$$\gamma(f) = \frac{S_{x_1'x_2'}(f)}{\sqrt{S_{x_1'x_1'}(f) S_{x_2'x_2'}(f)}}, \tag{3}$$

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.

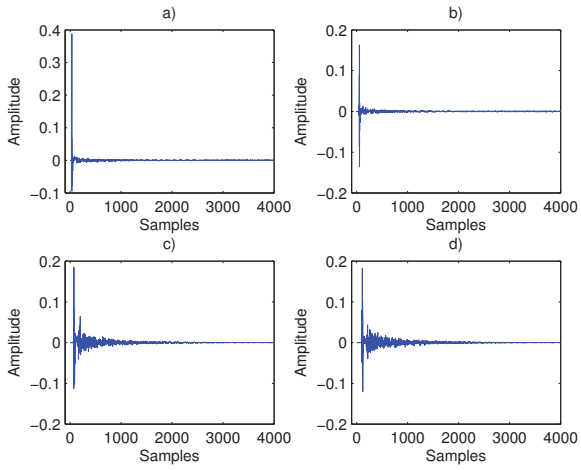


**Fig. 3**. Simulation environment setup: a) $\mathbf{g}_1$, b) $\mathbf{g}_2$, c) $\mathbf{f}_1$, d) $\mathbf{f}_2$.

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 $\omega_0$ 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 [3]

$$\text{mis}(n) = \sum_{k=1}^{2} \frac{\|\mathbf{f}_k(n) - \hat{\mathbf{f}}_k(n)\|}{\|\mathbf{f}_k(n)\|}. \tag{4}$$

### 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)[13].

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 Figure 4.

The listening test was performed by a balanced group of 10 evaluators where half of them were experienced listeners, and both quality and stereo perception of the signals were considered in the grading procedure.

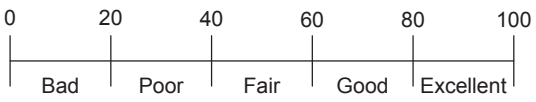


**Fig. 4**. Grading scale of the MUSHRA test.

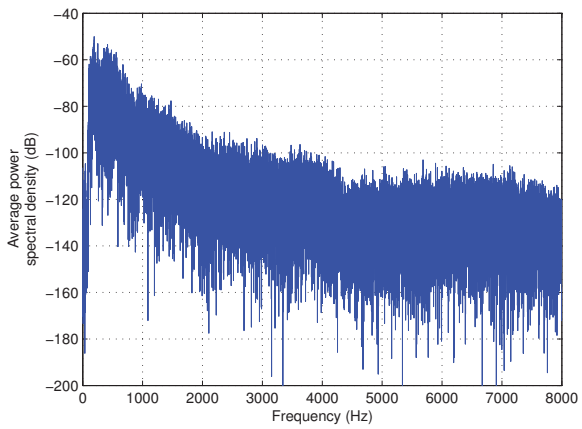


**Fig. 5**. Average power spectral density of the input signals.

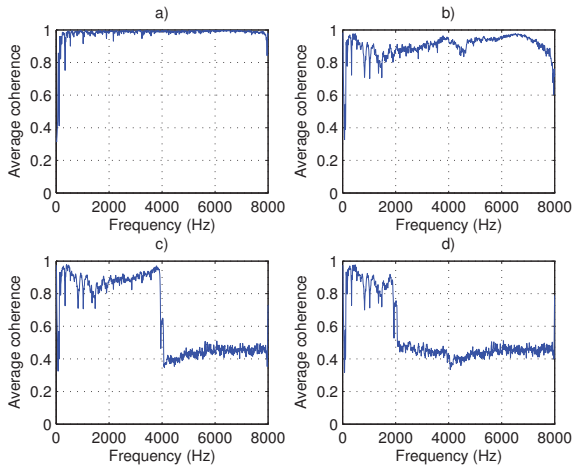


**Fig. 6**. Coherence function using: a) no decorrelation technique, b) HWR, c) hybrid1, d) hybrid2.

### 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 Figure 5.

All signals were recorded in the talkers' native language and their nationalities and genders are resumed below: 4 Americans (2 males and 2 females), 2 British (1 male and 1 female), 2 French (1 male and 1 female), 2 Germans (1 male and 1 female).

However, this original database was only used for the estimation of the coherence function. Since the assessment of the performance of the adaptive filters needs longer signals, all signals of each talker were concatenated getting a total of 10 signals with duration of 40 s (1 signal by talker) in this experiment. Moreover, due to the time consumption of the subjective tests, the subjective quality evaluation was performed using only 5 of the signals recorded in English.

## 4. EXPERIMENTS RESULTS

### 4.1. Experiment 1

With regards to performance assessment, Figure 6 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 Figure 6(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 Figure 6(c), the good effect of the FS technique can already be noticed in the highest sub-band (above 4 kHz), where coherence values approximately

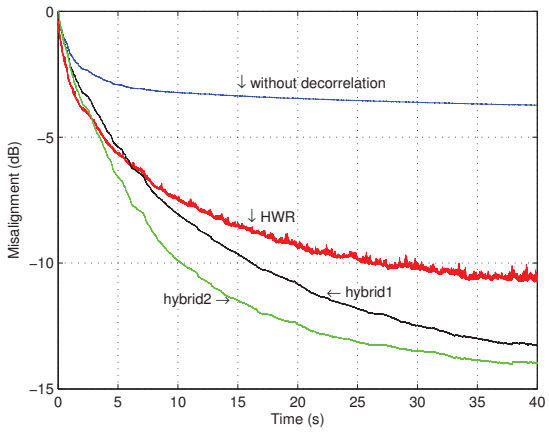


**Fig. 7**. Misalignment: convergence comparison in SAEC.

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 Figure 6(d).

Figure 7 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 algorithm [3], and shorter signals, corroborated that the hybrid techniques always surpass the HWR method in terms of the relative misalignment.

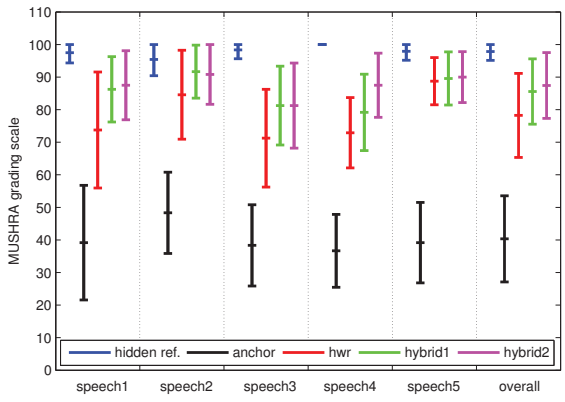


**Fig. 8**. MUSHRA results: quality comparison.

### 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 Figure 8. 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 resides 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 $\omega_0$ 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.

## 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 relative to the original ones.

## 6. ACKNOWLEDGEMENTS

This work was partially funded by FCT - Portuguese Science and Technology Foundation (SFRH/BD/49038/2008).

## 7. REFERENCES

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, New Jersey: Prentice Hall, 1985.

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

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

[4] T. Gansler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, Seatle, USA, May 1998, pp. 3649–3652.

[5] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 6, pp. 686–696, September 2001.

[6] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *Proc. IEEE ICASSP*, Honolulu, Hawaii, USA, April 2007, pp. 17–20.

[7] G. M. Ballou, *Handbook for Sound Engineers*, 3rd ed. Waltham, Massachusetts: Focal Press, 2002.

[8] E. Hansler and G. Schmidt, *Acoustic Echo and Noise Control*. Hoboken, New Jersey: John Wiley & Sons, 2004.

[9] J. Blauert, *Spatial Hearing*, 2nd ed. Cambridge: MIT Press, 1983.

[10] ITU-T, "Recommendation G.191: Software tools for speech and audio coding standardization," International Telecommunications Union, Geneva, Switzerland 2010.

[11] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. International Conference on Digital Signal Processing*, Santorini, Greece, July 2009.

[12] F. Albu, J. Kadlec, N. Coleman, and A. Fagan, "The gauss-seidel fast affine projection algorithm," in *IEEE Workshop on Signal Processing Systems*, San Diego, USA, October 2002, pp. 109–114.

[13] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Geneva, Switzerland 2003.