DECORRELATION OF INPUT SIGNALS FOR STEREOPHONIC ACOUSTIC ECHO CANCELLATION USING THE CLASS OF PERCEPTUAL EQUIVALENCE

Anis Ben Aicha and Sofia Ben Jebara

Ecole Supérieure des Communications de Tunis
Research unit TECHTRA
Route de Raoued 3.5 Km, Cité El Ghazala, Ariana, 2083, TUNISIA
anis_ben_aicha@yahoo.fr, sofia.benjebara@supcom.rnu.tn

ABSTRACT

In communication systems using stereophonic signals, the high correlation of input signals decreases dramatically the behavior of acoustic echo cancelers (AEC's) especially the misalignment between the estimated impulse response and the real impulse response of receiving room. In this paper, we focus on the decorrelation of input signals without any modification of auditive quality. Using perceptual properties, we show that it is possible to find a set of signals which are perceptually equivalent to input signals, in spite of their different spectral and temporal shapes. Hence, the 'class of perceptual equivalence' (CPE) is defined. It is an interval build in the frequency domain limited by two bounds: the upper bound of perceptual equivalence (UBPE) and the lower bound of perceptual equivalence (LBPE). These two bounds are used as perceptual transparent non linear transformations of input signals in order to decorrelate them. We show experimentally that the improvements yielded by this method are higher than those of classical non linear transformations.

1. INTRODUCTION

In variety of speech communication systems such as teleconferencing systems, acoustic echo cancelers are necessary to remove undesirable echo that results from coupling between loudspeakers and microphones. Early systems consider the single channel case. Novel applications such as hands-free communications, home entertainment and virtual reality contribute to the development of multi-channel signal processing techniques providing the users with enhanced sound quality.

In this paper, we consider specifically the stereophonic case. Fig.1, represents the conventional acoustic echo cancellation (AEC) scheme. The transmitted signals from remote room $x_1(n)$ and $x_2(n)$ are restored in local room by two loudspeakers and picked up by two microphones. If nothing is done, the picked signal is retransmitted to the remote room producing the undesired acoustic echo.

Conventional AEC's seek to estimate the echo using adaptive finite impulse response filters $(\hat{\mathrm{H}}_1, \hat{\mathrm{H}}_2)$ to model the acoustic impulse response of the two echo paths $(\mathrm{H}_1, \mathrm{H}_2)$ between loudspeakers and one of the two microphones. Similar paths couple to the other microphone, but for the scheme simplicity, they are not shown.

In contrast to the case of mono-channel, the problem of stereophonic echo cancellation is much more difficult to solve. The reasons why the problem is difficult are well explained in [1] [2] and are mainly due to the strong correlation of the transmitted signals x_1 and x_2. Indeed, the two signals x_1 and x_2 are obtained from the common source $s(n)$ by filtering it with impulse response of remote room G_1 and G_2.

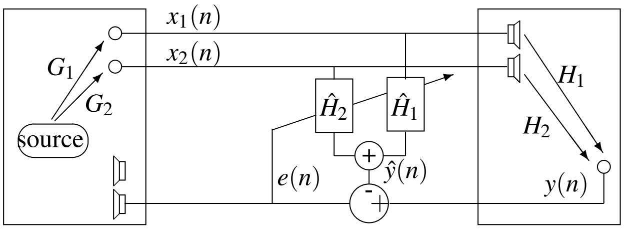

Figure 1: Basic stereophonic acoustic echo canceler.

To improve AEC's behaviors, many methods are proposed in literature to decorrelate the transmitted signals x_1 and x_2. We can classify them into two categories according to the usage or not of the human auditory properties.

The first category groups together decorrelation techniques without perceptual considerations. In [2], an independent random noise with low level is added in each channel in order to reduce the correlation between them. In [4], a channel dependent signal is added, it is obtained by a nonlinear transformation of each channel.

In the second category, human perceptual properties are taken into account when adding to each input signal another one which should be quasi-white and under the input signal masking threshold rendering them inaudible. This technique turns out to be more efficient than non perceptual decorrelating methods [5].

Instead of adding another signal, we propose, in this paper, to perceptual modify the input signals in order to decorrelate them without perceptible degradation. This fact ensures the perceptual transparency of the proposed transformation and improve the behaviors of stereophonic AEC's.

In previous works [6] [7], we demonstrate that it is possible, for each signal, to construct a class of perceptual equivalence (CPE). It is defined as an interval where signals have the same auditory properties as the original one. This class is built in the spectral domain with perceptual rules, it is limited by two boundaries: the low boundary of perceptual equivalence (LBPE) and the upper boundary of perceptual equivalence (UBPE). This concept introduces a great degree of freedom to choose signals exciting the adaptive filters which have the same perceptual quality as the original ones and characterized by low cross-correlation. Hence, this paper aims finding and justifying such signals.

The paper is organized as follows. Section 2 recalls the fundamental problem of stereophonic acoustic echo cancellation and gives an overview of nonlinear methods to decorrelate input signals. In section 3, we construct the CPE using auditory properties of human ear. In section 4, we detail the proposed technique. In section 5, we compare the proposed technique over two techniques proposed in [3] and [5]. The results show the improvement of proposed AEC's versus these techniques. Then we conclude the paper.

2. UNIFYING SCHEME FOR INPUT SIGNALS DECORRELATION

It has been shown for stereo AEC that the strong correlation between input signals yields to the convergence of adaptive filters to a solution that does not correctly model the transfer functions between the loudspeakers and the microphones. In fact the cross correlation matrix of input signals plays a significant role in the convergence of adaptive filters to real impulse responses of the receiving room [2]. The well conditioned the cross correlation matrix is, the better behaviors of AEC's are obtained [4]. If the cross correlation matrix is ill conditioned, which is the case with high correlated signals, adaptive filters may converge to non optimal solution.

To make cross correlation matrix well conditioned permitting to adaptive filters to converge to real impulse responses of local rooms with small bias, it is obvious that we need to decorrelate input signals.

Fig. 2 represents a unifying scheme to decorrelate input signals. $x_1(n)$ and $x_2(n)$ are non linearly transformed to other signals $x'_1(n)$ and $x'_2(n)$. The purpose of this transformation is to make the new signals $x'_1(n)$ and $x'_2(n)$ non correlated as possible with constraint of no modification of perceptual quality. In the following, we recall some known methods in literature according to their usage of perceptual properties.

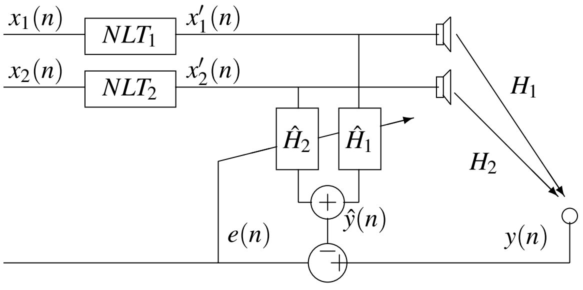

Figure 2: Unifying scheme to decorrelate stereophonic input signals by nonlinear transformations.

2.1 Non perceptual techniques (NPT)

The first idea to partially decorrelate the input signals is based on the addition of low level of independent random noise to each channel in order to reduce the coherence between $x'_1(n)$ and $x'_2(n)$ when compared to the coherence between $x_1(n)$ and $x_2(n)$ [2]. Another technique consists on the modulation of each input signal with an independent random noise [9]. However, it is shown that even if the level of added noise is very low, the quality of speech is significantly degraded [4].

Without perceptual considerations and in order to minimize audible degradation, it is preferable to add something like the original signal. The main idea proposed in [4] is to add to each input signal $x_1(n)$ and $x_2(n)$ a nonlinear transformation of the signals themselves.

$$x'_i(n) = x_i(n) + \alpha \mathscr{F}[x_i(n)], \qquad (1)$$

where $\mathscr{F}$ is the chosen nonlinear function and α is a parameter which controls the amount of the added signal.

2.2 Perceptual techniques (PT)

The basic idea is to take advantage of human auditory properties, namely the simultaneous masking which happens in frequency domain. In [5], the proposed method is based on the addition of a random noise spectrally shaped to be masked by the presence of the input signal. To achieve the complete masking of the added noise, the well known noise masking threshold is computed from each input signal. It expresses the maximum level of added noise to be not audible. The masking threshold serves to shape the added noise.

2.3 Motivation

Previous perceptual techniques add a data-dependent or a data-independent signal to partially decorrelate stereophonic input signals. A simple and an intuitive idea arises: is it possible to eliminate some parts of input signals instead of adding another signal? Of course, this elimination must operate as a nonlinear operator and must not introduce any audible degradation on input signals. Once again, we must consider human auditory properties but in different manner.

The basic idea is inspired from our previous works dealing with speech denoising techniques [6] [7], where we proposed a class of perceptual equivalence (CPE). Each signal having a spectrum belonging to this class is heard as the original one. We think that if we consider the class bounds, using the upper bound as a transformed spectrum for the first input signal $x_1(n)$ and the lower bound as a transformed spectrum for the second input signal $x_2(n)$, we can reduce the correlation between them.

First of all and before detailing the idea, let us describe the perceptual class of equivalence, its bounds and its usefulness in the next paragraph.

3. CLASS OF PERCEPTUAL EQUIVALENCE

We aim to find an interval where possible signals belonging to it have the same auditive properties as the considered signal. For such purpose, auditory properties of human ear are considered. More precisely, the masking concept is used: a masked signal is made inaudible by a masker if the masked signal magnitude is below the perceptual masking threshold MT [8].

Using both signal spectrum and its masking threshold, we look for decision rules to decide on the audibility of a modified spectrum obtained by adding, subtracting or modifying some frequency components of the considered signal. These rules will permit to construct the perceptual class of equivalence.

3.1 Upper Bound of Perceptual Equivalence UBPE

The masking threshold is a curve computed in short time frequency domain from power spectrum of the considered signal. It represents for each frequency component, the maximum level of added noise to be inaudible. Hence, an additive noise, with power spectrum under MT, will be inaudible.

However, it modifies the power spectrum shape of the considered signal. It is obvious, that we can add many 'inaudible' noises so that we get several shapes of power spectrum with the same auditive quality as the original signal.

We seek to determine the intervals formed by the power spectrum of the original speech, as a lower limit, and another curve, as an upper limit, so that, any modified signal (i.e., original signal + inaudible noise) with power spectrum between the two limits will not impair the perceptual quality of the original signal.

Since the masking threshold MT represents the maximum power of an additive inaudible noise, the upper limit that we look for is the curve that results from adding the power spectrum of the original signal with its own masking threshold.

The resulting spectrum is called upper bound of perceptual equivalence "UBPE" and is defined as follows [6] [7]

$$UBPE(m,f) = \Gamma_s(m,f) + MT(m,f), \tag{2}$$

where m (resp. f) denotes frames index (resp. frequency index). $\Gamma_s(m,f)$ is the clean speech power spectrum.

3.2 Lower Bound of Perceptual Equivalence LBPE

By duality, some attenuations of frequency components can be heard as speech distortion of input signal. Thus, by analogy to UBPE, we propose to calculate a second curve which expresses the lower bound under which any attenuation of frequency components is heard as a distortion. We call it lower bound of perceptual equivalence "LBPE". To compute LBPE, we used the audible spectrum introduced by Tsoukalas *and al* for audio signal enhancement [10]. In such case, audible spectrum is calculated by considering the maximum between the clean speech spectrum and the masking threshold.

When speech components are under MT, they are not heard and we can replace them by a chosen threshold $\sigma(m,f)$.
The proposed LBPE is defined as follows [6], [7]

$$LBPE(m,f) = \begin{cases} \Gamma_s(m,f) & \text{if } \Gamma_s(m,f) \geq MT(m,f) \\ \sigma(m,f) & \text{otherwise .} \end{cases} \tag{3}$$

The choice of $\sigma(m,f)$ obeys only one condition. It must be under the masking threshold $\sigma(m,f) < MT(m,f)$. We choose it equal to the absolute threshold of hearing. It is defined as the minimum power of a signal to be audible in absolute silence [8].

3.3 Usefulness of UBPE and LBPE

Using UBPE and LBPE, we can define three regions characterizing the perceptual quality of a modified input signal: modified frequency components between UBPE and LBPE are perceptually equivalent to the original input signal components. Frequency components above UBPE contain a audible noise and frequency components under LBPE are characterized by speech distortion. Hence, UBPE and LBPE constitute the limits of admissible transparent modification of input signals: every signal having spectral shape included between UBPE and LBPE is perceptually equivalent to the original one. The set of these signals forms the class of perceptual equivalence "CPE".

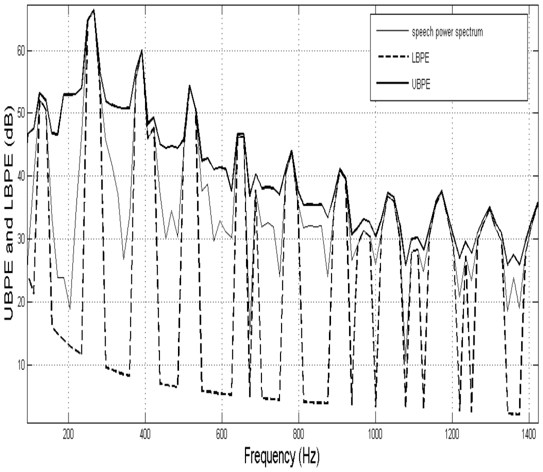

Figure 3: An illustration of UBPE and LBPE (in dB) of a speech frame.

As illustration, we present in Fig.3 an example of speech frame power spectrum and its related curves UBPE (upper curve in bold line) and LBPE (bottom curve in dash line). The original speech power spectrum is, for all frequencies index, between the two curves UBPE and LBPE.

4. PROPOSED TECHNIQUE TO DECORRELATE STEREOPHONIC SIGNALS

4.1 Proposed technique principle

The CPE permits multiple choices of signals with equivalent perceptual quality. For our application, we seek to reduce the correlation between input signals. In other words, our purpose is to change temporal forms of input signals as possible such as they become less correlated without any audible degradation. Bringing up the problem to frequency domain, we can, intuitively, choose the two limits UBPE for the first input signal $x_1(n)$ and the LBPE for the second input signal $x_2(n)$. In fact, the two limits lead to the most different signals.

The proposed nonlinear transformations (NLT) are written as follows.

$$\begin{cases} NLT_1(x_1) = UBPE_{x_1} & \text{computed from } x_1(n) \\ NLT_2(x_2) = LBPE_{x_2} & \text{computed from } x_2(n) \end{cases} \tag{4}$$

4.2 Reduction of coherence function

To mathematically justify our choice, let us recall the expression of coherence function which is computed in spectral domain as follows.

$$C(f) = \frac{\gamma_{x_1x_2}(f)}{\sqrt{\gamma_{x_1x_1}(f)\gamma_{x_2x_2}(f)}}, \tag{5}$$

where $\gamma_{x_ix_i}(f)$ denotes the power spectrum density of x_i (for $i=1$ and $i=2$), $\gamma_{x_1x_2}(f)$ denotes the inter-spectral density of signals x_1 and x_2.

The coherence function measures the similarity in frequency domain between two signals. If $C(f)$ is near 1, it means that the two signals are highly correlated. In the opposite case, when $C(f)$ is close to zero, the two signals are completely decorrelated.

In our case of acoustic echo cancellation, the significance of the coherence function is related to the cross correlation matrix of input signals. For decorrelated input signals the cross correlation matrix is well conditioned which leads to the convergence of adaptive filters to the real impulse responses of the receiving room [4].

In Fig.4, we represent the coherence magnitude (CM) of original signals and the CM obtained with three considered methods: non perceptual technique NPT developed in [3], perceptual technique PT developed in [5] and our proposed technique. The input signals $x_1(n)$ and $x_2(n)$ were obtained by convolving a clean speech with two impulse responses G_1 and G_2 of length 4096.

It can be seen that the coherence function of original signals is high for all frequencies, which corresponds to an ill conditioned cross-correlation matrix. NPT method reduces the CM, but PT method does better. This is an expected result since PT method takes into account perceptual concepts to decorrelate input signals. However, the best reduction is achieved by our proposed technique, since the CM is much reduced even in the low and medium frequencies. Thus, better behavior of adaptive algorithms can be expected with this latter method.

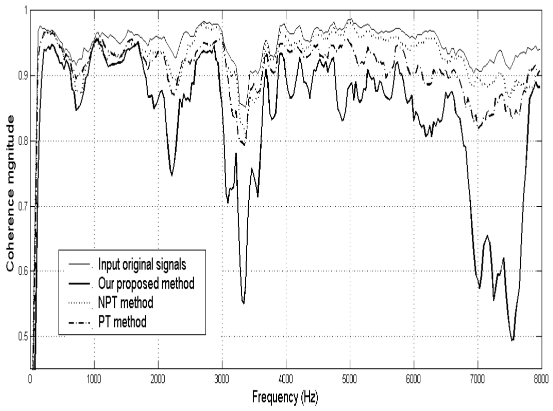

Figure 4: Coherence magnitude comparison using three methods (NPT technique, PT technique and our proposed technique).

5. EXPERIMENTAL RESULTS

5.1 Evaluation criteria

We wish to evaluate and compare our proposed method over NPT method and PT method using adequate criteria for acoustic echo cancellation.

- Minimum square error (MSE)

$$MSE(n) = 10\log_{10}\left\{e(n)^2\right\}, \tag{6}$$

where $e(n) = y(n) - \hat{y}(n)$ is the estimation error.

- Misalignment (M)

$$M_i(n) = \frac{\left[\hat{H}_i(n) - H_i\right]^T \left[\hat{H}_i(n) - H_i\right]}{H_i^T H_i}, \tag{7}$$

where the upper script T denotes the transposition operator.

$M_i(n)$expresses the estimation error of the impulse response of the i^{th} path in local room at iteration n.

- Echo return loss enhancement (ERLE)

ERLE represents the attenuation of acoustic echo. It is computed by block of N samples. The $ERLE(k)$ of the k^{th} block is given by:

$$ERLE(k) = 10\log_{10}\frac{\sum\limits_{n=(k-1)N+1}^{kN} y(n)^2}{\sum\limits_{n=(k-1)N+1}^{kN} [y(n) - \hat{y}(n)]^2}. \tag{8}$$

5.2 Simulation results

The microphone output signal $y(n)$ in the receiving room is obtained by summing the two convolutions $(H_1 * x_1)$ and $(H_2 * x_2)$, where H_1 and H_2 are impulse responses of receiving room, each of length 4096 points and truncated to 512 points. For all of our simulations, we have used the normalized least mean square algorithm (NLMS). To smooth the curves, misalignment and MSE are averaged over 128 points.

Fig.5 represents the evolution of MSE over the number of iterations for the considered methods (NPT developed in [3], PT developed in [5] and our proposed technique). We remark that the MSE decreases when n increases. The less residual echo is obtained with the proposed technique. It means that the echo is more reduced when regarding NPT and PT techniques.

Fig.6 represents the misalignment of considered techniques. Once again, the proposed method is seen to have the best performances since it greatly reduces the misalignment.

Fig.7 represents the ERLE of considered methods. The best ERLE is obtained with the proposed method.

6. CONCLUSION

In this paper, we have developed and tested a new method based on the use of human auditory properties to improve behaviors of stereophonic acoustic echo cancelers . Indeed, two boundaries UBPE and LBPE are constructed in spectral domain leading to multiple choice of perceptual equivalent signals. The choice of UBPE for the first input signal and LBPE for the second input signal yields to an efficient reduction of the coherence of stereophonic signals, improving the behaviors of AEC. The new technique was compared with previous proposal and was found to be more efficient.

REFERENCES

[1] H. Buchner, J. Benesty, and W. Kellermann, “Multichannel Frequency-Domain Adaptive Filtering with Application to Multichannel Acoustic Echo Cancellation,” in *Adaptive Signal Processing: Applications to Real-World Problems*, Berlin: Springer-Verlag, J. Benesty and Y. Huang, Eds., 2003, ch. 4.

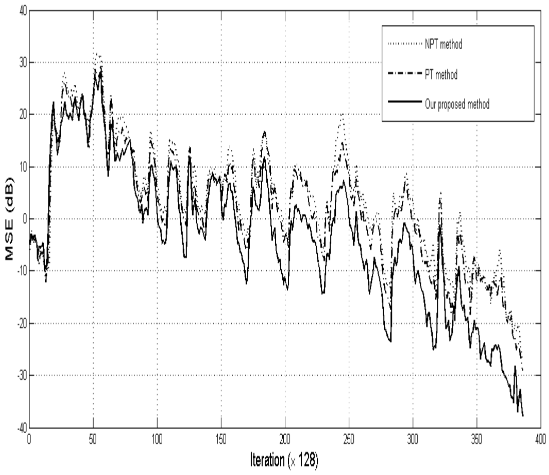

Figure 5: Evolution of MSE for the three considered methods using with 2-NLMS.

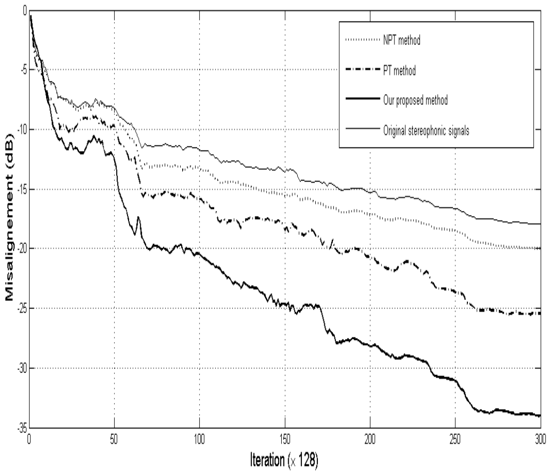

Figure 6: Misalignment of three considered methods obtained with 2-NLMS.

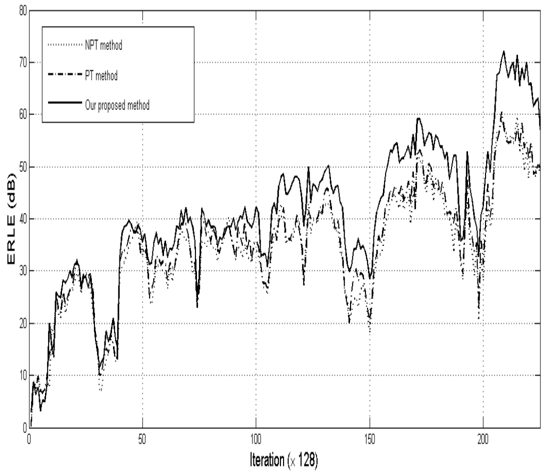

Figure 7: Evolution of ERLE for the three considered methods obtained with 2-NLMS.

[2] M. M. Sondhi and M. R. Morgan, “Stereophonic acoustic echo concellation - an overview of fundamental problem,” *IEEE Signal Processing Letters*, vol. 2, no. 8 pp. 148-151, 1995.

[3] J. Benesty, D. R. Morgan and M. M. Sondhi, “A better undrestanding and an improved solution to the specific problems of stereophonic aoustic echo cancellation,” *IEEE Transaction on Speech and Audio Processing*, vol. 6, no. 8, pp. 156-165, 1998.

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

[5] A. Gilloire and V. Turbin, “Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellation,” in *Porc. IEEE ICASSP*, pp. 3681-3684, 1998.

[6] A. Ben aicha et S. Ben Jebara, “Caractérisation perceptuelle de la dégradation apportée par les techniques de débruitage de la parole,” *Traitement et Analyse de l’Information Méthodes et Applications TAIMA*, Tunisia, 2007.

[7] A. Ben aicha et S. Ben Jebara, “Quantitative perceptual separation of two kinds of degradation in speech denoising applications,” *Workshop on Non-Llinear Speech Processing NOLISP*, France, 2007.

[8] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *in Proc. IEEE*, vol. 88, pp. 451-513, 2000.

[9] S. Shimauchi and S. Makino, “Stereo projection echo canceler with true echo path estimation,” in *Porc. IEEE ICASSP*, pp. 3059-3062, 1995.

[10] D. E. Tsoukalas, J. N. Mourjopoulos and G. K. Kokkinakis, “Speech enhancement based on audible noise suppression,” *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6, pp. 497-517, 1997.

[11] J. D. Johnston, “Transform coding of audio signal using perceptual noise criteria,” *IEEE J. Select. Areas Commun.*, vol 6, pp. 314-323, 1988.