Abstract—Decorrelation is a well known issue in the context of Stereophonic Acoustic Echo Cancellation: it is related to the problem of uniquely identifying each pair of room acoustic paths, due to high inter-channel coherence. This paper proposes an improvement to a previous work to decorrelate a stereo signal based on the missing fundamental phenomenon. An adaptive algorithm is employed to track the behavior of one of the two channels, ensuring a continuous decorrelation without affecting the stereo perception. The improvement is obtained through a constraint function, based on a sigmoid function, applied to the adaptive coefficient which controls the fundamental frequency estimation. Several results are presented, comparing our approach with a technique based on non-linearity and the masked noise injection method. The comparison is done in terms of convergence speed improvement and subjective quality perception, in order to confirm the validity of the improved proposed approach.

I. INTRODUCTION

Acoustic Echo Cancellers are used in teleconferencing systems to reduce undesired echoes originated from coupling between the loudspeaker and the microphone. The spreading of systems which allow the presence of more than one participant needs more realistic performances in terms of sound localization, in order to identify the speaker position [1]. This is possible only by increasing the number of system channels: as a consequence, multichannel acoustic echo cancellers become essential. By the use of a two channels system, it is already possible to obtain more realistic performances because listeners have spatial information which help to identify the speaker position [1].

The problem of stereophonic acoustic echo cancellation cannot be viewed as a simple generalization of the monochannel case: besides the fact that more adaptive filters are needed due to the increased system complexity, stereo systems have more problems than mono ones [1], since the system performances depend both on the receiving room and the transmission room. This is due to the fact that the two channels of the stereophonic signal are generated by the same source; therefore, a linear relation between them exists. As a consequence, the normal equation that the adaptive algorithm has to solve is singular and has no unique solution: all these solutions depend on the speaker position in the transmission room. Because of the non stationary nature of the speaker, echo cancellation may be poor and suffers from convergence problems due to room changes: a method to decorrelate the input channels must be introduced in order to obtain a suitable signal for echo cancellation.

In literature, several decorrelation techniques have been proposed in order to weaken the relation between the two stereo channels. These approaches can be divided into two main categories.

The first group of methods are based on the direct alteration of the two channels, e.g. adding to the signal a non linearly processed version of itself [1], [2] or applying a shift of one sample on one of the two channels with a two-tap filter with time-varying coefficients [3] or modifying the phase response with first-order and second-order time-varying all-pass filters [4], [5]. A recent approach based on a phase alteration obtained by phase modulation, has been proposed in [6]: the phase modulation amplitude is a function of the frequency subband on which the modulation is performed. This approach, as previous ones [7], [8], is based on the idea of preserving signal quality, applying a technique that alters the signal as less as possible, especially for low frequencies.

The second group of techniques are based on the introduction of an external signal on both channels: this operation can be controlled by a desired signal-to-noise ratio [1], [9] or by taking advantage of the human auditory properties [10]. In the first case, the simple introduction of noise is not sufficient to achieve satisfactory results ensuring a good audio quality; in [9] the added noise level is adjusted according to the characteristics of the speech inputs in order to obtain a good quality of the decorrelated stereo speech signals. In the second case, random noise controlled by auditory properties is used in order to spectrally shape the noise components according to masking rules.

In this paper, a signal decorrelation technique which combines the characteristics of the above two categories is adopted, starting from the approach derived by the same authors in a previous work [11]. It is based on the psychoacoustic effect of the missing fundamental, which allows the perception of a pitch (i.e. the fundamental frequency) without the corresponding frequency actually being contained in the signal. Therefore, the algorithm consists in the removal of the fundamental frequency through an adaptive notch filter, which depends on a coefficient able to track it. It is evident that a correct pitch tracking is essential to avoid stereo perception alteration. Moreover, the coefficient used to control the pitch estimation has to be limited in a specified range, in order to prevent the filter from diverging [12].

Thus a novelty with respect to [11] has been adopted to improve the tracking performance: the control on the adaptive coefficient based on clipping [12] has been replaced by a control based on a constraint function, in order to vary the coefficient in the specified range, providing absence of discontinuity in the adaptation. As a consequence of this, listening audio quality is improved.

In Section II the method proposed in [11] is reviewed, considering the novel constraint control of the adaptive coefficient with respect to [11], showing the tracking performance improvement. Then, its validation by means of several simulations is reported in Section III: the two main aspects related to stereophonic decorrelation have been taken into consideration, i.e. convergence speed improvement and subjective quality evaluation. Finally, conclusions and future works are described in Section IV.

II. IMPROVED PROPOSED APPROACH

The approach presented in [11] is based on the missing fundamental effect, related to the perception of the fundamental frequency without the corresponding frequency actually being contained in the signal. This phenomenon has been explained as a human brain capability to process the information present in the overtones to
calculate the “missing” fundamental [13]: when the fundamental frequency is removed from the set of harmonics, the sound perceived is almost unchanged, i.e. the pitch does not change but there is slight alteration of the sound timbre due to the number of harmonics reproduced.

Thus, it is possible to iteratively estimate and remove the pitch of one of the two channel, obtaining a great decorrelation in the lower part of the spectrum without affecting the signal quality [11]. The algorithm is characterized by two main blocks: the first one is able to track the fundamental frequency of the signal while the second one is capable to remove the pitch, by using an adaptive notch filter. Fig. 1 shows a block diagram of the entire proposed algorithm, which can be summarized as follows:

1) the signal is filtered by a low-pass filter in order to act just on the desired frequency range;
2) a down-sampling by a factor M is applied to the signal in order to increase spectral resolution;
3) the pitch is estimated and removed using the adaptive notch filter, controlling the adaptation through a constraint function;
4) an up-sampling by a factor M is applied to the signal;
5) the up-sampled signal is added to the delayed high-pass filtered signal.

As the stereo perception preservation is related to the correct pitch estimation, a novelty has been introduced in the adaptive coefficient $k_0$, which controls the algorithm, i.e. clipping has been replaced by a constraint function, based on a sigmoid function. Thus, new equations have to be derived for calculating $k_0$, as fully described in Section II-A.

A. Review of the proposed approach with the introduction of the constraint function

The main part of the adaptive algorithm in [11] is the notch filter, which depends on a coefficient able to track the fundamental frequency and adapted at any new sample of the input signal. This filter is given by second order lattice form; considering the structure in Fig. 2, we can derive the following equations:

$y(n) = x(n) - k_0 y(n - 1)(\alpha + 1) - \alpha y(n - 2) \tag{1}$

$w(n) = y(n) + 2k_0 y(n - 1) + y(n - 2) \tag{2}$

Substituting Eq.(1) into Eq.(2), it is possible to obtain the transfer function used to filter the signal for removing the fundamental frequency:

$H(z) = \frac{1 + 2k_0 z^{-1} + z^{-2}}{1 + k_0 (1 + \alpha) z^{-1} + \alpha z^{-2}} \tag{3}$

It is described by an adaptive coefficient $k_0$, related to the tracked frequency, and a pole-zero contraction factor $\alpha$, which controls the filter’s bandwidth [12]. In order to prevent the filter from diverging, the coefficient $k_0$ must be bounded in the range $(-1; 1)$. Instead of clipping $k_0$ as proposed in [12], a constraint function, based on a sigmoid function, can be used in order to bound $k_0$ in the specified range, as follows:

$k_0 = \frac{2}{1 + e^{-g_0}} - 1 \tag{4}$

Deriving the adaptive equation [14], the cost function for the Eq. (2) is represented by

$E(n) = \sum_{k=0}^{n} w^2 (k) = \sum_{k=0}^{n} [A(k)k_0 (g_0) + B(k)]^2 \tag{5}$

$A(n) = 2y(n - 1) \tag{6}$

$B(n) = y(n) + y(n - 2) \tag{7}$

The value of $g_0$ is derived minimizing Eq. (5):

$\frac{d(E)}{d(g_0)} = 0 \tag{8}$

where it results

$\frac{d(E)}{d(g_0)} = \frac{d[B(k)k_0 (g_0) + B(k)]^2}{d(g_0)} \tag{9}$

$= \left[2A^2(k)k_0 (g_0) + 2A(k)B(k)\right] \frac{d(k_0 (g_0))}{d(g_0)} \tag{10}$

$= \left[2A^2(k) \left[\frac{2}{1 + e^{-g_0}} - 1\right] + 2A(k)B(k)\right] \frac{2e^{-g_0}}{(1 + e^{-g_0})^2} \tag{11}$

It follows:

$g_0(n) = -\ln \left| \frac{C(n)}{D(n)} \right| \tag{12}$

where the absolute value is introduced to prevent $g_0$ from assuming complex values and $D(n)$ and $C(n)$ are adapted as follows:

$C(n) = \lambda C(n - 1) + (1 - \lambda)A(n)[A(n) + B(n)] \tag{13}$

$D(n) = \lambda D(n - 1) + (1 - \lambda)A(n)[A(n) - B(n)] \tag{14}$

Fig. 1. Overall scheme of the improved proposed approach.

Fig. 2. Structure of the adaptive notch filter.
Substituting Eq. (6) and Eq. (7), we can obtain:

\[ C(n) = \lambda C(n-1) + (1 - \lambda)2y(n-1)[2y(n-1) + y(n) + y(n-2)] \]
\[ D(n) = \lambda D(n-1) + (1 - \lambda)2y(n-1)[2y(n-1) - y(n) - y(n-2)] \]

where \( y(n) \) is the all-pole filtered version of the input signal \( x(n) \) as shown in Fig. 2 and \( \lambda \) is a forgetting factor close to 1. The adaptive coefficient \( k_0 \) is calculated from \( g_0 \) using Eq. (4). The estimated frequency can be computed at any step from the knowledge of \( k_0(n) \), since

\[ f_{est}(n) = \frac{f_s}{M} \sin^{-1}(-k_0(n)) \]

where \( f_s \) is the sampling frequency and \( M \) is the down-sampling factor.

The tracking capability of the adaptive notch filter has been evaluated considering a test signal composed by different tones; the results confirm its effectiveness, as also demonstrated in [12]. Fig. 3 compares the pitch tracking performance for a female speech, using [11] and the novel control of \( k_0 \); it is evident that the latter overcomes the former in term of pitch estimation accuracy as underlined in the figure. It is worth noting that in both cases the algorithm does not diverge in the presence of unvoiced segments, taking advantage of the low pass filtering; even if a pitch is not actually present the decorrelation keeps working using a time-varying notch filter as it happens with time-varying comb filters [15].

### III. EXPERIMENTAL RESULTS

Several experimental results will be shown in this section in order to prove the effectiveness of the improved proposed approach, focusing on the two main aspects that have to be considered for a decorrelation technique:

- improvement of the adaptive filters convergence speed (Section III-A); it has been evaluated in terms of misalignment, calculated as reported in [6], but only in the frequencies range of interest;
- subjective quality assessment (Section III-B); it has been evaluated through listening tests conforming to the ITU-R BS.1534 (MUSHRA - Multiple Stimulus with Hidden Reference and Anchors) Recommendation [16].

#### A. Performance evaluation

Tests have been carried out on a speech signal sampled at 16 kHz, using simulated impulse responses at the transmission and receiving rooms. The adaptive filters used for the identification of the echo paths have the same length as the simulated impulse responses (i.e. 1024 samples). The improved Normalized Least Mean Square (NLMS) algorithm [17] has been used, deriving the performance of the approaches in terms of misalignment. For the implementation of the proposed method, a crossover network, based on a FIR solution [18], has been used in order to split the input signal into two sub-bands. In order to strongly assess its effectiveness, a comparison with a non linear approach [19] (\( \alpha = 0.4 \)) and the masked noise injection approach [10] has been done, considering the frequencies range of interest. Fig. 4 shows the results obtained for the aforementioned decorrelation techniques, proving as the improved proposed approach achieves better results in the same number of iterations, considering a frequency band of \([0 : 500]\) Hz. As expected, no speed convergence improvement is visible between [11] and the same approach based on the novel control, because it affects the audio quality, guaranteeing a more correct pitch estimation. Consistent results have been obtained even in terms of Echo Return Loss Enhancement (ERLE), as reported in Fig. 5 for the whole spectrum.

#### B. Subjective quality evaluation

To assess the stereophonic perception and signal quality of the improved proposed approach, we also performed listening tests, according to the ITU-R BS.1534 (MUSHRA) Recommendation [16]. MUSHRA is a double-blind multi-stimulus test method with hidden reference and one or more hidden anchors. Following the MUSHRA guidelines, the subjects are required to score the stimuli according to a continuous quality scale divided in five equal intervals (i.e. Bad,
prove the effectiveness of the approach. Tracking. Several experimental results have been showed in order to fundamental frequency estimation, allowing even its more correct without the corresponding frequency actually being contained in based on the psychoacoustic effect of the missing fundamental, has by the same authors for the stereophonic decorrelation problem, to the excellent intervals, but the best results have been obtained when the subject correctly identifies the hidden reference.

Each subject has been asked to grade these aspects on a 100-point scale as mentioned above. A default grade of 100 is assigned to the stimulus identified by the subject as the hidden reference, as suggested by the Recommendation. The number of involved subjects was 10 (8 males and 2 females), as suggested by [16], with ages ranging from 21 to 35. Regarding the used equipment, two pair of loudspeakers (Genelec 6010A) have been used connected to an Intel Centrino 2 Laptop with a professional sound card MOTU traveler connected to a Intel Centrino 2 Laptop.

A subjective difference grade (SDG) is computed by subtracting the score assigned to the actual hidden reference from the score assigned to the actual modified speech signal [20]. Positive values can be obtained when the subject correctly identifies the hidden reference and negative value if the subject misidentifies the hidden reference. The mean scores are calculated considering all the selected subjects, and these results have been used to evaluate the performance of the improved proposed approach. Fig. 6 shows the results obtained in terms of subjective difference grade for Test 1: according to [20], it is evident that all the values assumed by this difference are limited to the excellent intervals, but the best results have been obtained through the improved proposed approach, that shows values close to 0%. With regard to Test 2, all the evaluated techniques properly preserve the stereo perception.

IV. Conclusion

In this paper an improvement of the novel approach proposed by the same authors for the stereophonic decorrelation problem, based on the psychoacoustic effect of the missing fundamental, has been presented. This phenomenon implies the perception of a pitch without the corresponding frequency actually being contained in the signal. Starting from this idea, a new control function for the adaptive algorithm has been introduced, in order to speed up the fundamental frequency estimation, allowing even its more correct tracking. Several experimental results have been shown in order to prove the effectiveness of the approach.

However, since our method produces good results for low frequencies, future works will be oriented towards a subband architecture, which combines the improved proposed approach for low frequencies with other techniques, taking advantage of other psychoacoustic phenomena, to cover the entire frequency spectrum.

REFERENCES


