Speech Communication 86 (2017) 97–106

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Multichannel acoustic echo cancellation exploiting effective fundamental frequency estimation

Laura Romoli[1,*], Stefania Cecchi[1], Francesco Piazza[1]

Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 19 October 2016
Accepted 21 November 2016
Available online 28 November 2016

Keywords:
Fundamental frequency estimation
Multichannel acoustic echo cancellation
Multichannel decorrelation
Multichannel double-talk detection
Multichannel voice activity detection

ABSTRACT

Multichannel teleconferencing systems exploit multichannel acoustic echo cancellers to weaken the echo replicas due to the acoustic coupling among loudspeakers and microphones. Many issues have to be dealt with to ensure an effective echo cancellation in such systems, including the reduction of the interchannel coherence among channels, the detection of an active remote human speaker, and the identification of the well-known double-talk scenario. In this paper, a comprehensive solution for dealing with the aforementioned aspects is discussed. The core of this solution is the estimation of the fundamental frequency of the audio signal. It is exploited for weakening the linear relation among channels and for tracking the presence of an active human speaker both in the remote room and in the local room of a teleconferencing system. The computational complexity of the presented approach is also reported to support its feasibility in a real scenario. Then, its real-time implementation is presented and validated on the NU-Tech framework, showing its fundamental frequency tracking capability and echo reduction performance both in simulated and real scenarios.

1. Introduction

Multiple microphones and loudspeakers are exploited in audio/video teleconferencing systems to improve the listening experience of its participants placing them inside the acoustic scenario (Benesty et al., 1998; Cecchi et al., 2016; Romoli et al., 2015, 2014b, 2014a). However, the presence of multiple channels implies the need for multichannel acoustic echo cancellers (MAECs) to reduce the echo replicas on the microphone signals arising from the coupling between each loudspeaker and each microphone. In the context of acoustic echo cancellation, several issues have to be dealt with. More specifically, an effective identification of multiple echo paths requires the reduction of the correlation among channels since convergence problems could arise from the linear relationship existing among them. This aspect is well-known in the literature as “non-uniqueness” problem (Benesty et al., 1998). Then, remote human speaker activity tracking and the identification of the well-known double-talk scenario are other important aspects to deal with, based on the discrimination of vocal frames, silence segments, and noisy periods (Romoli et al., 2015). Indeed, voice activity detection (VAD) procedures and double-talk detection (DTD) algorithms assume an important role in practical implementation of such systems since filter coefficients update is typically suppressed when there is no active human speaker in the remote room and/or when there are active human speakers both in the remote room and in the local room (Benesty et al., 2001; Tashev, 2009). A comprehensive work combining all these aspects in a unique solution is not present in the state of the art up to authors’ knowledge, while multiple approaches focusing on each aspect have been discussed in the literature.

Regarding channel decorrelation, some approaches, previously focused on the stereophonic scenario, have been generalized to the multichannel case, to reduce the interchannel coherence and thus, to improve the performance of multichannel systems (Cecchi et al., 2012). Among these, the phase modulation approach for surround sound systems (Herre et al., 2007) and the half-wave rectifier distortion (Benesty et al., 1998) were proposed. Recently, a novel solution for multichannel decorrelation that is suitable for both speech and music signals has been proposed. It is based on fundamental frequency estimation and removal in order to exploit psychoacoustic criteria and to avoid audio quality worsening (Romoli et al., 2014b, 2014a; Cecchi et al., 2011; Romoli et al., 2012, 2010).

Regarding VAD procedures, several approaches have been proposed in the literature as presented in Ma and Nishihara (2013) and Babu and Vanathi (2009) based on signal energy and

* Corresponding author.
E-mail addresses: l.romoli@univpm.it (L. Romoli), stefania.cecchi@gmail.com (S. Cecchi).
[1] Fax Number: +390712204453

http://dx.doi.org/10.1016/j.specom.2016.11.009
0167-6393/

zero-crossing rate (Atal and Rabiner, 1976) and on more robust acoustic features (Davis et al., 2006; Shuyin et al., 2009; Marzinzik and Kollmeier, 2002) or their combination with pitch analysis in the presence of higher levels of noise (Atal and Rabiner, 1976). Recently, a new method for detecting voiced segments of the far-end signal has been proposed exploiting the fundamental frequency tracking through second-order adaptive notch filters in a multichannel scenario (Romoli et al., 2015).

Regarding DTD algorithms, multichannel procedures are typically based on cross-correlation (Benesty and Gansler, 2002; Iqbal et al., 2009). More specifically, the approach described in Benesty and Gansler (2002) is based on the cross-correlation between the far-end signal and the microphone signal while the technique described in Iqbal et al. (2009) is based on the cross-correlation between the residual echo signal and the microphone signal. In both cases, the detection is made according to the value assumed by a decision variable with respect to a predefined threshold. Recently, a new DTD method has been proposed based on human speaker activity tracking through second-order adaptive notch filters (Cecchi et al., 2016).

In this paper, starting from the obtained promising results (Cecchi et al., 2016; Romoli et al., 2015, 2014b), a low-complexity comprehensive solution for multichannel acoustic echo cancellation based on fundamental frequency estimation is investigated. Although the exploitation of pitch analysis for active human speaker detection is a known topic in the literature, the proposed approaches are intended to be combined in an exhaustive solution for MAEC systems, since decorrelation, VAD, and DTD procedures are performed based on the same variables. Several results are reported to show the performance of pitch tracking and echo reduction. Moreover, the real-time implementation of the solution is discussed with the aim of applying the approach in real scenarios. The computation of the number of multiplications required for multichannel decorrelation, VAD procedure, and DTD algorithm is also reported in order to underline the advantage of using the same parameters to face different problems. Indeed, in this way, a self contained solution can be derived.

The paper is organized as follows. The role of fundamental frequency estimation in MAEC system is presented in Section 2 discussing the exploitation of the fundamental frequency estimation for decorrelation (Section 2.2), VAD algorithm (Section 2.3), and DTD procedure (Section 2.4). Then, the computational complexity of the aforementioned procedures (Section 4) is reported to prove its feasibility in real scenarios. The real-time implementation of the approach and its validation are reported in Section 5 in terms of effective tracking of fundamental frequency for remote and local human speakers and of satisfactory echo reduction. Finally, concluding remarks are presented in Section 6.

2. The role of fundamental frequency estimation

The proposed solution for multichannel acoustic echo cancellation that includes multichannel decorrelation, voice activity detection, and double-talk detection is based on fundamental frequency estimation. The adopted scenario is reported in Fig. 1, with L loudspeakers and P microphones. The remote signal is the residual error $r_l(n)$ (being $l = 1, \ldots, L$ and n the time index) and it is processed by the decorrelation block providing the decorrelated signals $x_l(n)$ and the estimated frequency $f_l^*(n)$ (Romoli et al., 2014b, 2014a). Then, the estimated echo signal $y_p(n)$ ($p = 1, \ldots, P$) is calculated by the system identification procedure and the error signal $e_p(n)$ is computed by subtracting it from the microphone signal $d_p(n)$. The system identification procedure is controlled by the VAD and DTD blocks, that can enable or disable the filter coefficients update. The former makes its decision according to the remote fundamental frequency $f_l^*(n)$, whereas the latter detects

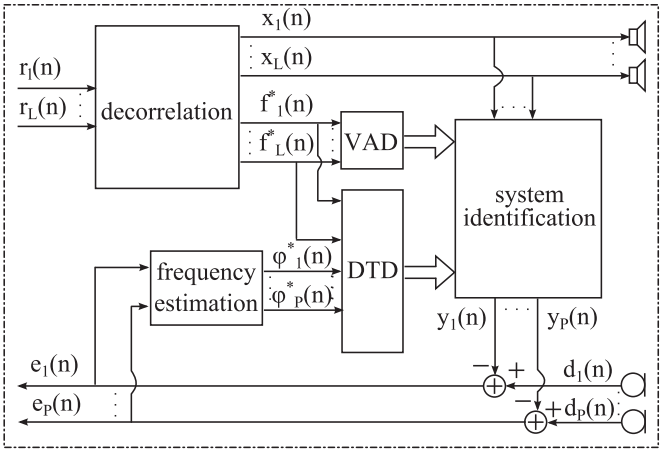


Fig. 1. Block diagram of multichannel audio teleconferencing system.

the double-talk situation considering both the remote fundamental frequency $f_l^*(n)$ and the local fundamental frequency $\varphi_p^*(n)$. In the following sections, a brief review of the algorithm adopted for estimating and removing the fundamental frequency is reported as previously discussed in Cecchi et al. (2016) and Romoli et al. (2015, 2014b). Then, its application to multichannel decorrelation, VAD procedure, and DTD algorithm is presented.

2.1. Fundamental frequency estimation and removal

A block diagram summarizing the main steps for fundamental frequency estimation and removal is reported in the upper branch of Fig. 2. The algorithm is reported in this section for M input signals but this procedure will be exploited in the following sections for estimating and refining the fundamental frequency of both the L loudspeaker signals ($M = L$) and the P microphones signals ($M = P$).

The M input signals $s_m(n)$ ($m = 1, \ldots, M$) are decimated by a factor D to improve spectral resolution at low frequencies and then filtered by pre-processing second-order shelving filters H_{pre} with gain G_s and cut-off frequency f_{cuts} (Zölzer, 2002). Assuming that the fundamental frequency is typically bounded between $f_{\text{inf}} = 60$ Hz and $f_{\text{sup}} = 600$ Hz (Cecchi et al., 2016), G_s and f_{cuts} have to be chosen to apply a spectral decay that ensures a correct tracking and reduces possible mighty harmonic components. Then, the M signals are filtered by the following adaptive notch filters:

$$H_m(z, \widetilde{n}) = \frac{1 + 2k_m(\widetilde{n})z^{-1} + z^{-2}}{1 + k_m(\widetilde{n})[1 + \alpha_m(\widetilde{n})]z^{-1} + \alpha_m(\widetilde{n})z^{-2}}, \tag{1}$$

where $\widetilde{n}$ is the index of the decimated time instant, the contraction factor $\alpha_m(\widetilde{n})$ controls the m-th filter bandwidth, and $k_m(\widetilde{n})$ is an adaptive coefficient related to the fundamental frequency to be tracked (Cecchi et al., 2012). More specifically, the coefficient $k_m(\widetilde{n})$ is bounded to the range $(-1, +1)$ according to the following sigmoid function:

$$k_m(\widetilde{n}) = \frac{2}{1 + e^{-g_m(\widetilde{n})}} - 1, \tag{2}$$

where $g_m(\widetilde{n}) \in R$ has to be adapted in order to minimize the time average of the notch filter output as fully described in Cecchi et al. (2012) and Romoli et al. (2012). In this way, the signals $x_m(\widetilde{n})$ without the fundamental frequency are obtained. The signals $x_m(n)$ are then derived applying post-processing second-order shelving filters H_{post} with the same gain G_s and cut-off frequency f_{cuts} and upsampling by a factor D. Differently, the mth fundamental frequency is a function of the adaptive parameter $k_m(\widetilde{n})$ as given

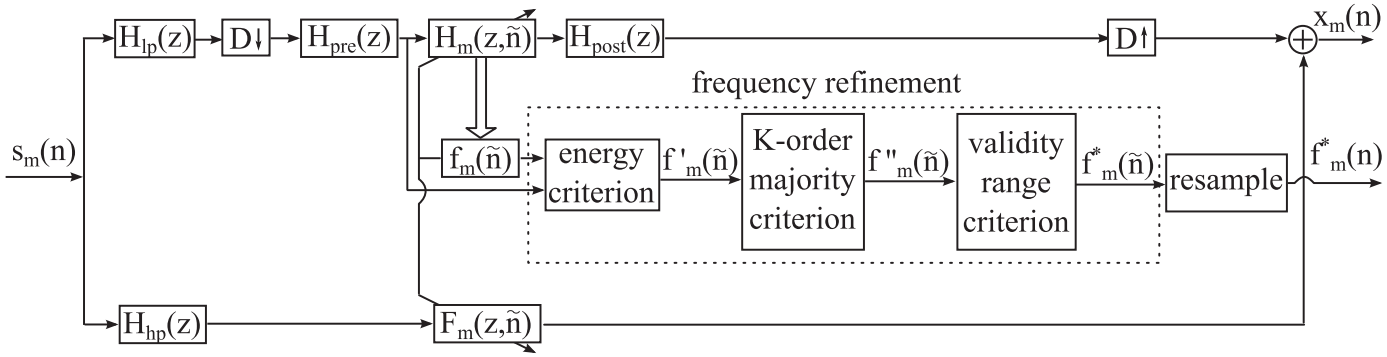


Fig. 2. Block diagram of multichannel decorrelation. The fundamental frequency estimation and removal is performed in the upper branch, while time-varying allpass filtering is performed in the lower branch.

in the following expression:

$$f_m(\tilde{n}) = \frac{f_s}{D} \cdot \frac{1}{2\pi} \cos^{-1}[-k_m(\tilde{n})], \tag{3}$$

where f_s is the sampling frequency. The tracking performance can be improved according to a refinement procedure based on the following three steps (Fig. 2):

1. *Energy criterion*

$$f'_m(\tilde{n}) = \begin{cases} f_m(\tilde{n}) & E_{s_m}(\tilde{n}) > \gamma_E \\ f_{\min} & E_{s_m}(\tilde{n}) \le \gamma_E, \end{cases} \tag{4}$$

where $E_{s_m}(\tilde{n}) = \beta E_{s_m}(\tilde{n}-1) + (1-\beta)s_{mD}(\tilde{n})$ is the energy of the mth decimated signal $s_{mD}(\tilde{n})$, β is the energy forgetting factor, γ_E is a properly chosen energy threshold, and $f_{\min} < f_{\inf}$.

2. *K-order majority criterion*

$$f''_m(\tilde{n}) = \begin{cases} \min\{F_{K,m}(\tilde{n})\} & n_K > K/2 \\ f_{\min} & n_K \le K/2, \end{cases} \tag{5}$$

where $F_{K,m}(\tilde{n}) = [f'_m(\tilde{n}), f'_m(\tilde{n}-1), \ldots, f'_m(\tilde{n}-K+1)]$ is a set of K consecutive frequencies. The fundamental frequency is refined to the lowest frequency belonging to $F_{K,m}(\tilde{n})$ if n_K consecutive frequencies are bounded in the same frequency band, i.e., a half-tone centered to $f'_m(\tilde{n})$.

3. *Validity range criterion*

$$f^*_m(\tilde{n}) = \begin{cases} f''_m(\tilde{n}) & f_{\inf} \le f''_m(\tilde{n}) \le f_{\sup} \\ f_{\min} & f''_m(\tilde{n}) < f_{\inf} \vee f''_m(\tilde{n}) > f_{\sup}, \end{cases} \tag{6}$$

where $f_{\inf}$ and $f_{\sup}$ have been previously described.

The assignment of $f_{\min}$ value in (4)–(6) is performed to set the refined fundamental frequency out of the validity range. This aspect will be exploited for VAD and DTD decisions in the following sections.

2.2. Multichannel decorrelation

The decorrelation block of Fig. 1 is based on the algorithm described in Section 2.1 and on the application of second-order time-varying allpass filters (Fig. 2). The former provides decorrelation in the validity range of the fundamental frequency while the latter is applied out of this range to guarantee decorrelation on the whole spectrum. Therefore, the input signals are low-pass filtered and high-pass filtered to apply the two aforementioned algorithms. The cut-off frequency is suitably designed according to the upper limit of the validity range of the fundamental frequency and to the decimation factor D as presented in Section 2.1.

In particular, the interchannel coherence is weakened by estimating and removing the fundamental frequency of the L input signals $r_l(n)$ through L adaptive notch filters with the same transfer function as in (1). This procedure is based on the "missing-fundamental" phenomenon (Cecchi et al., 2011; Romoli et al., 2012, 2010) and does not substantially alter the spatial perception and the audio quality of the input signal if the frequency is effectively tracked (Larsen and Aarts, 2004). In this way, the solution is suitable for speech and high quality audio signals. In order to supply the desired disparity among channels also when the same fundamental frequency is tracked on all channels, the adaptive notch filters described in Section 2.1 are used with different slowly varying values of the contraction factor $\alpha_l(n)$. The vector $\boldsymbol{\alpha}_l(n) = [\alpha_1(n), \ldots, \alpha_L(n)]$ is modified as reported in the following equation:

$$\boldsymbol{\alpha}(n) = \begin{cases} s[\boldsymbol{\alpha}(n-1), 1] & \text{if } \left(n - Q\left\lfloor \frac{n}{Q} \right\rfloor\right) = 0 \\ \boldsymbol{\alpha}(n-1) & \text{otherwise,} \end{cases} \tag{7}$$

where $s[\cdot,1]$ represents a right circular shift of one position within vector $\boldsymbol{\alpha}_l(n)$ performed every Q samples, being Q the block length for time variation, and $\lfloor\cdot\rfloor$ denotes the lowest integer value. Therefore, assuming $\boldsymbol{\alpha}(n) = [0.95, 0.75, 0.55]$, the right circular shift provides $\boldsymbol{\alpha}(n+Q) = [0.55, 0.95, 0.75]$.

To ensure decorrelation on the whole spectrum, the high-pass filtered L input signals $r_l(n)$ are processed by L time-varying allpass filters as discussed in Romoli et al. (2014b). The pole (multiplicity 2) of the lth time-varying allpass filters is controlled by the adaptive coefficient $k_l(n)$ related to the tracked fundamental frequency in (1) (Oppenheim et al., 1999):

$$F_l(z,n) = \frac{k_l^2(n) - 2k_l(n)z^{-1} + z^{-2}}{1 - 2k_l(n)z^{-1} + k_l^2(n)z^{-2}}. \tag{8}$$

The filter is causal and stable since $|k_l(n)| < 1$.

2.3. Multichannel voice activity detection

The VAD block of Fig. 1 takes advantage of the algorithm described in Section 2.1 to identify the presence of voice activity. In particular, the VAD decision is based on the L refined fundamental frequencies $f^*_l(n)$ provided by the decorrelation block through (1)–(6).

The remote human speaker is classified as active if voice activity is detected for at least $L/2+1$ far-end signals $r_l(n)$. To this aim, the decision variable ζ is set to 0. Then, for each channel, the mean value $\tilde{f}^*_l$ of the current N refined fundamental frequencies is computed as follows:

$$\tilde{f}^*_l = \frac{1}{N} \sum_{n=1}^{N} f^*_l(n), \tag{9}$$

where N is the input signal block length. The decision variable ζ is incremented if $\tilde{f}^*_l \in [f_{\inf}, f_{\sup}]$, i.e., if the channel is characterized by a valid fundamental frequency. Finally, the current far-end

Table 1
Computational complexity computed at the receiving side of the communication system with L loudspeakers and P microphones for the multichannel decorrelation algorithm, the VAD decision procedure, and the DTD decision procedure. It has been considered that N is the block size, N_f is the filter length of the low-pass filter and the anti-aliasing filter, and D is the decimation factor.

	Number of multiplications		
	Decorrelation	**VAD**	**DTD**
Proposed approach	$LN\left(2N_f + 4 + \frac{26}{D}\right)$	L	$P\left[1 + N\left(N_f + \frac{22}{D}\right)\right]$
Approach of Herre et al. (2007)	$2LN[2 + \log(2N)]$		
Approach of Atal and Rabiner (1976)		LN	
Approach of Iqbal et al. (2009)			$P[2(4N + 5) + 2N\log((2N)]$

Table 2
Configuration for MAEC system.

Parameter	Assigned value
Fundamental frequency estimation	
Downsampling D	32
Shelving filters gain G_s	−30 dB
Shelving filters cut-off f_{cuts}	350 Hz
K-order in majority criterion K	5
Energy threshold γ_E	0.015
Energy forgetting factor β	0.999
Lower bound for valid fundamental frequency f_{inf}	60 Hz
Upper bound for valid fundamental frequency f_{sup}	600 Hz
Multichannel decorrelation	
Forgetting factor β_{ff} for $g_m(\tilde{n})$ computation	0.98
Lower bound for contraction factor α	0.55
Upper bound for contraction factor α	0.95
Circular shift interval Q	1 s
DTD algorithm	
Forgetting factor β_{ff} for $g_m(\tilde{n})$ computation	0.9
Fixed contraction factor α	0.95
Multichannel system identification	
Adaptive filter length N_F	4096
Convergence rate μ	0.01

signal frame is detected as voiced if $\zeta > \left\lceil \frac{L}{2} \right\rceil$, where $\lceil \cdot \rceil$ denotes the greatest integer value.

2.4. Multichannel double-talk detection

Also the DTD block of Fig. 1 takes advantage of the algorithm described in Section 2.1 to identify the situation of double-talk, i.e., when there is voice activity detected both in the remote room and in the local room. P refined fundamental frequency $\varphi_p^*(n)$ are obtained applying (1)–(6) to the P residual echo signals $e_p(n)$. Then, the algorithm takes into consideration the L mean values $\tilde{f}_l^*$ computed in (9) and the P refined frequencies $\varphi_p^*(n)$. In particular, if double-talk is detected on at least $P/2 + 1$ microphones, then the scenario is classified as double-talk. To this aim, the control variable ζ_{dtd} is initialized to 0 and the mean value $\tilde{\varphi}_p^*$ of the current N refined fundamental frequencies is computed as follows

$$\tilde{\varphi}_p^* = \frac{1}{N}\sum_{n=1}^{N} \varphi_p^*(n). \tag{10}$$

Then, the control variable ζ_{dtd} is incremented if $\tilde{\varphi}_p^* > f_{min}$, i.e., if a valid frequency has been tracked. Eventually, double-talk is detected if the condition $\zeta_{dtd} > \left\lfloor \frac{P}{2} \right\rfloor$ is met.

3. System identification

The estimation of the echo paths is performed according to a block normalized least mean square (BNLMS) optimization approach (Romoli et al., 2012). It is assumed that the adaptive filter coefficients vector $\hat{\boldsymbol{h}}_{lp}(m)$ between loudspeaker l and microphone p at each block index m has length N_F. Given the lth input signal

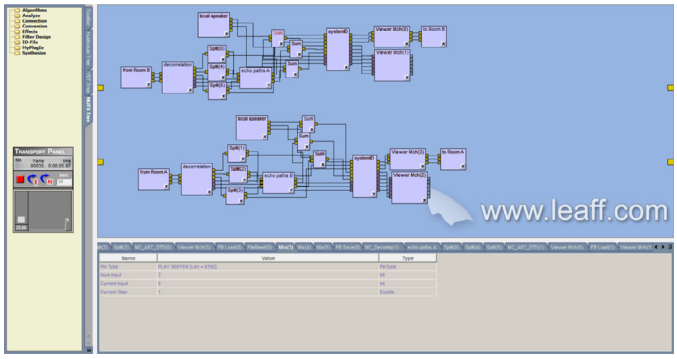


Fig. 3. NU-Tech board for validating the proposed solution for multichannel acoustic echo cancellation.

matrix $\boldsymbol{X}_l(m) = [\boldsymbol{x}_l(mN) \cdots \boldsymbol{x}_l(mN - N_F + 1)]$, where N is the input block length, $\boldsymbol{x}_l(mN) = [x_l(mN) \cdots x_l(mN - N + 1)]^T$ is the lth input signal column vector, and $(\cdot)^T$ is the vector transpose operator, the pth estimated echo signal is obtained through the following equation:

$$\boldsymbol{y}_p(m) = \sum_{l=1}^{L} \boldsymbol{X}_l(m)\hat{\boldsymbol{h}}_{lp}(m-1). \tag{11}$$

Therefore, the pth residual error $\boldsymbol{e}_p(m)$ is obtained from the knowledge of the pth desired signal $\boldsymbol{d}_p(m)$ as follows:

$$\boldsymbol{e}_p(m) = \boldsymbol{d}_p(m) - \boldsymbol{y}_p(m). \tag{12}$$

The minimization of the mean square error provides the following equation for the filter coefficients vector update:

$$\hat{\boldsymbol{h}}_{lp}(m) = \hat{\boldsymbol{h}}_{lp}(m-1) + \mu \frac{\nabla_{lp}(m)}{\boldsymbol{x}_l^T(mN)\boldsymbol{x}_l(mN) + \delta}, \tag{13}$$

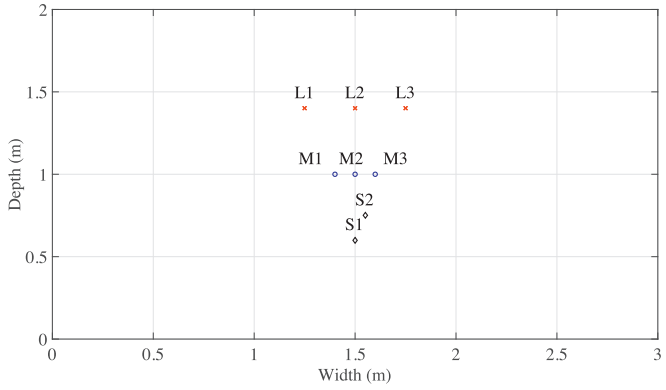
where δ is the regularization factor and $\nabla_{lp}(m)$ is the block gradient column vector estimated for each channel l and microphone p as follows:

$$\nabla_{lp}(m) = \boldsymbol{X}_l^T(m)\boldsymbol{e}_p(m). \tag{14}$$

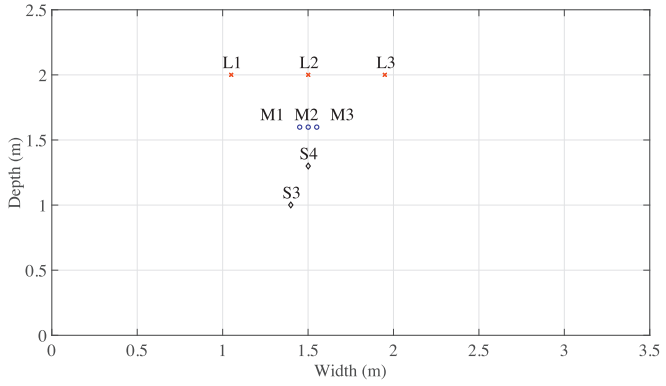
It is worth noting that Eqs. (11),(13), and (14) can be efficiently calculated in the frequency domain. Considering Fig. 1, adaptation is suppressed when the VAD block does not reveal the presence of an active remote speaker or when the DTD block reveals the presence of both an active remote speaker and an active local speaker.

4. Computational complexity

The computational complexity required for channel decorrelation, VAD decision, and DTD decision is provided in this section in terms of number of multiplications C, considering L loudspeakers, P microphones, block size N, and decimation factor D. Regarding the multichannel decorrelation algorithm (Section 2.2), the calculation of the computational complexity considers low-pass filters of



(a)



(b)

Fig. 4. Simulated rooms. (a) Room A. (b) Room B.

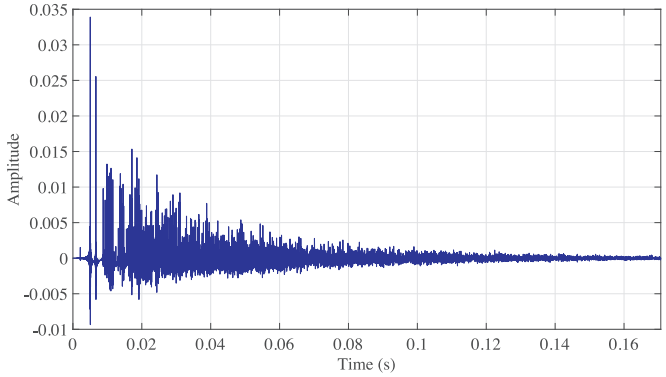
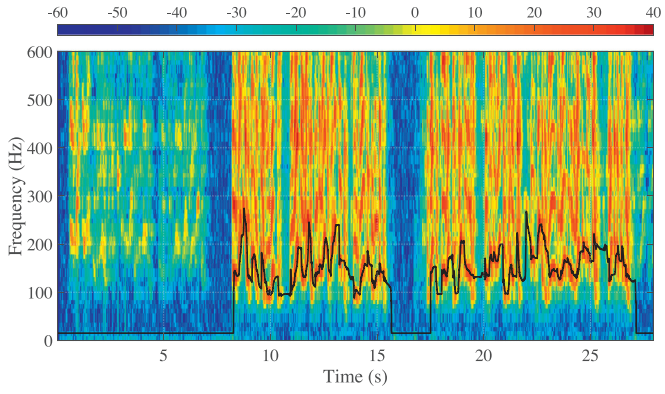
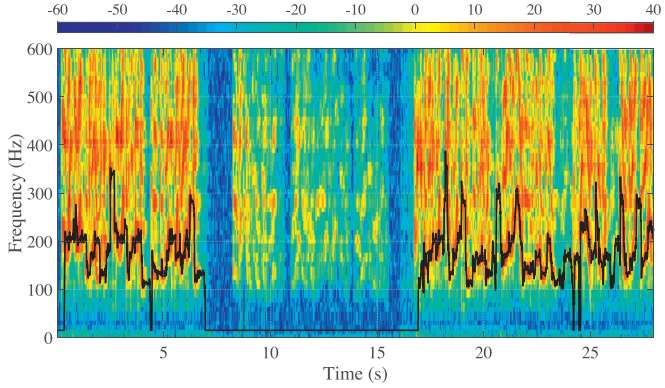


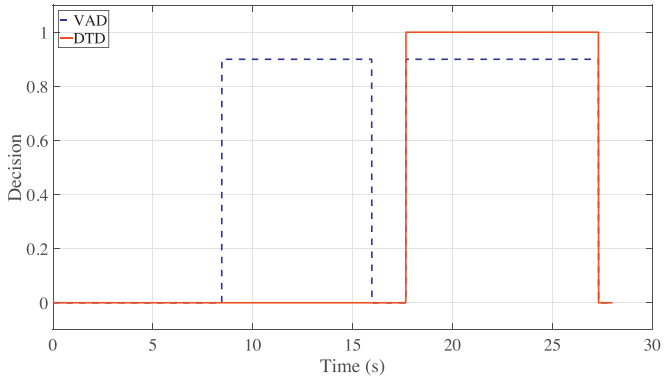
Fig. 5. Impulse response modelling the acoustic path between the loudspeaker L2 and the microphone M1 in Fig. 4(a).



(a)



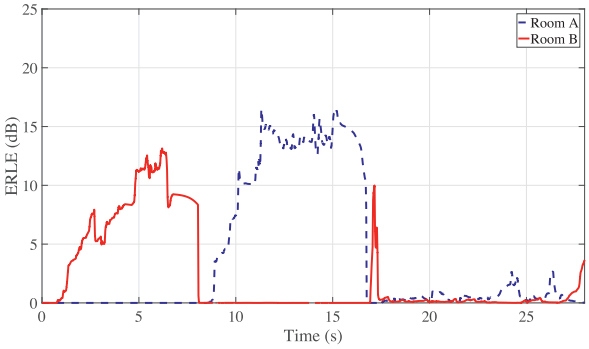
(b)



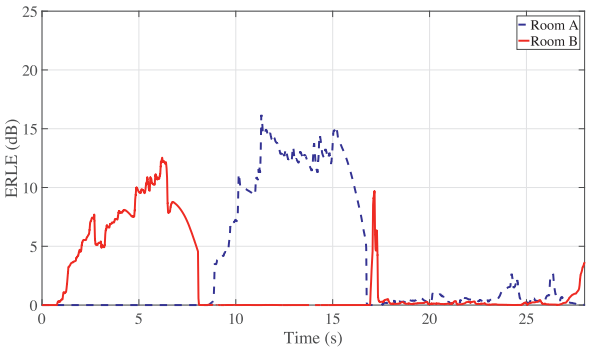
(c)

Fig. 6. Simulated test 1: tracking performance considering Room A. (a) Remote human speaker. (b) Local human speaker. (c) VAD and DTD decisions.

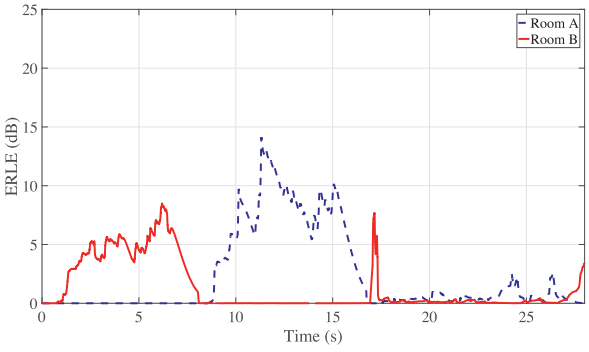
length N_f ($C = LNN_f$), second-order shelving filters for pre/post processing ($C = 2 \cdot 4LN/D$), second-order notch filters and their adaptive equations fully described in Cecchi et al. (2012) and Romoli et al. (2012) ($C = 13LN/D$), energy update ($C = 3LN/D$), K-order majority criterion ($C = 2LN/D$), anti-aliasing filters of length N_f for upsampling ($C = LNN_f$), and second-order all-pass filters ($C = 4LN$). Regarding the multichannel VAD algorithm (Section 2.3), it mainly exploits the results of channel decorrelation, thus, the calculation of the computational complexity considers only the averaging of the fundamental frequency vectors ($C = L$). Finally, regarding the multichannel DTD algorithm (Section 2.4), the calculation of the computational complexity takes into consideration low-pass filters of length N_f ($C = PNN_f$), second-order shelving filters for pre processing ($C = 4PN/D$), second-order notch filters and their adaptive equations fully described in Cecchi et al. (2012) and Romoli et al. (2012) ($C = 13PN/D$), energy update ($C = 3PN/D$), K-order majority criterion ($C = 2PN/D$), and the averaging of the fundamental frequency vectors ($C = P$). Therefore, the total number of multiplications needed for processing the audio streaming is reported in Table 1 for each procedure, also showing a comparison with state-



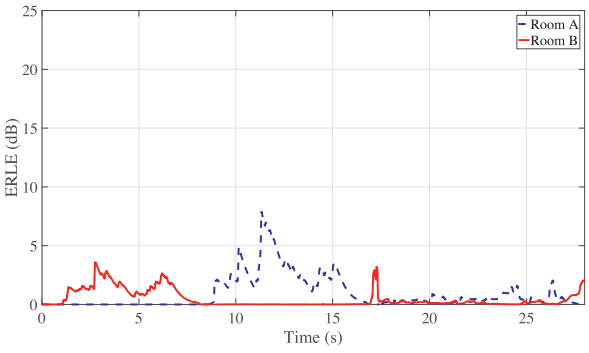
(a)



(b)

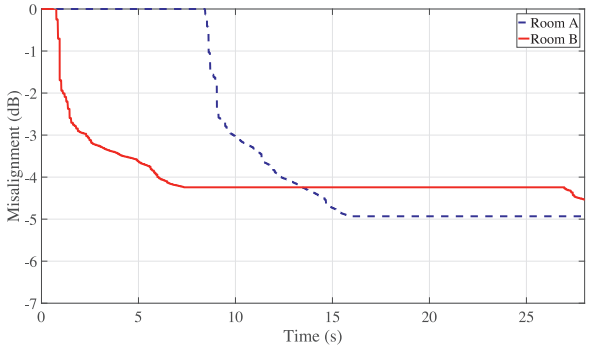


(c)

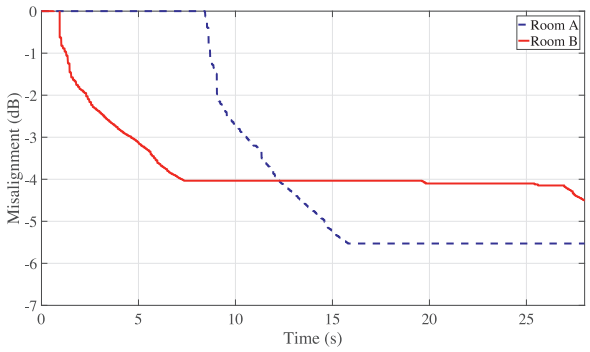


(d)

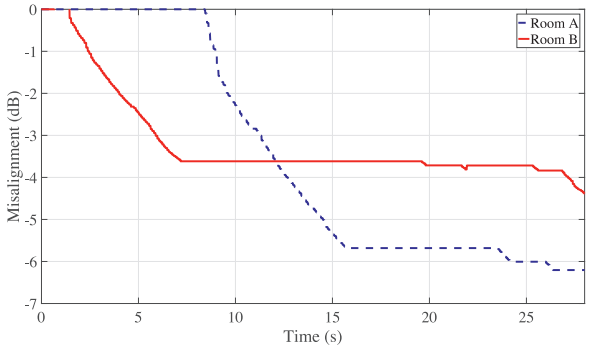
Fig. 7. Simulated test 1: ERLE. (a) SNR = 40 dB. (b) SNR = 30 dB. (c) SNR = 20 dB. (d) SNR = 10 dB.



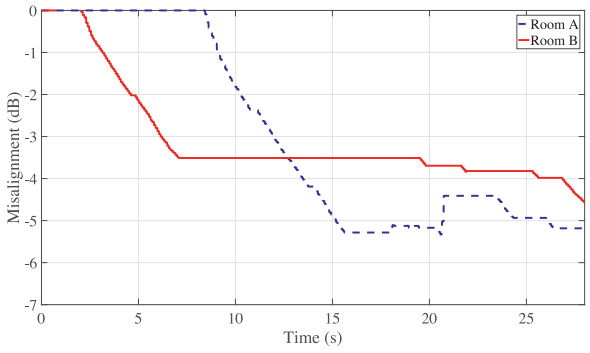
(a)



(b)



(c)



(d)

Fig. 8. Simulated test 1: Misalignment. (a) SNR = 40 dB. (b) SNR = 30 dB. (c) SNR = 20 dB. (d) SNR = 10 dB.

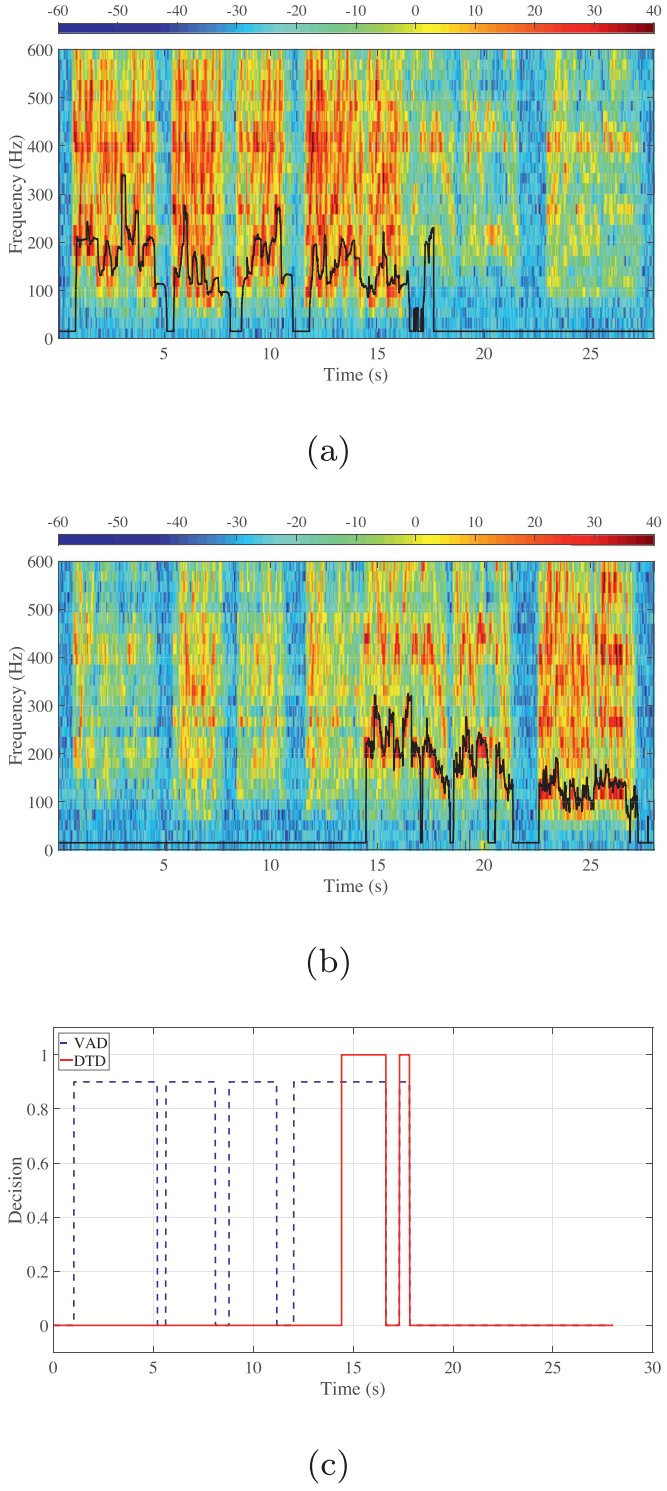


Fig. 9. Simulated test 2: tracking performance considering Room B. (a) Remote human speaker. (b) Local human speaker. (c) VAD and DTD decisions.

of-the-art approaches for decorrelation (Herre et al., 2007), VAD (Atal and Rabiner, 1976), and DTD (Iqbal et al., 2009). These techniques were chosen as reference approaches for each considered aspect since a multichannel comprehensive solution has not been discussed in the literature up to authors' knowledge. The number of multiplications of the aforementioned approaches are reported as an example, assuming $N = 4096$, $N_f = 1024$, $D = 32$, $L = 3$, and $P = 3$. In particular, focusing on the decorrelation technique, the proposed approach and the approach of Herre et al. (2007) require

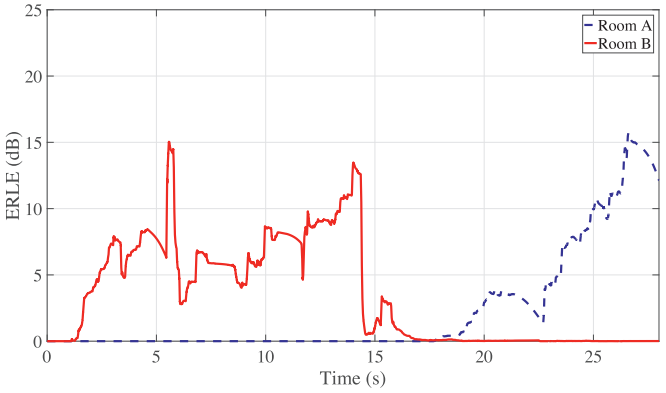


Fig. 10. Simulated test 2: ERLE.

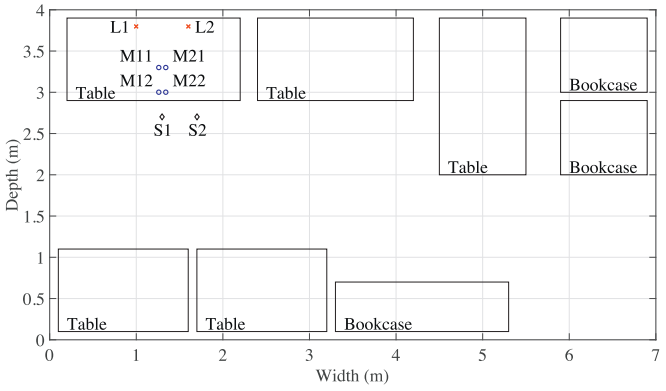


Fig. 11. Real room where real tests were performed.

25224960 and 145327 multiplications, respectively. For VAD decision, the proposed approach requires only 3 multiplications taking advantage of the results provided by decorrelation, while the approach of Atal and Rabiner (1976) requires 12288 multiplications. Finally, for DTD decision, the proposed approach and the approach of Iqbal et al. (2009) require 12591363 and 194509 multiplications, respectively. It is worth noting that the obtained total number of multiplications does not prevent the real-time performance as it will be shown in the following section in terms of CPU load percentage. Moreover, the proposed techniques guarantees a good trade-off between performance and audio quality preservation, exceeding many state-of-the-art approaches, as proved in previous works (Cecchi et al., 2016; Romoli et al., 2015, 2014b).

5. Performance evaluation

The real-time performance of the proposed solution for MAEC application in a teleconferencing audio system is presented in this section. The overall system has been validated on the NU-Tech framework (Lattanzi et al., 2008) in order to demonstrate its real-time performance. NU-Tech is a software platform for real-time testing digital audio processing (Lattanzi et al., 2008) exploiting its Plug-In architecture based on NU-Tech Satellites (NUTSs). A set of NUTSs performing the aforementioned procedures for multichannel acoustic echo cancellation has been realized and suitably connected on the graphical interface reported in Fig. 3, considering the communication between two different rooms, denoted as Room A and Room B. The main NUTSs shown in Fig. 3 are described in the following list:

- *from Room A/B* and *to Room A/B*: these NUTSs implement the communication between the two rooms.
- *local human speaker*: this NUTS loads the audio streaming relative to the local human speaker in the room.

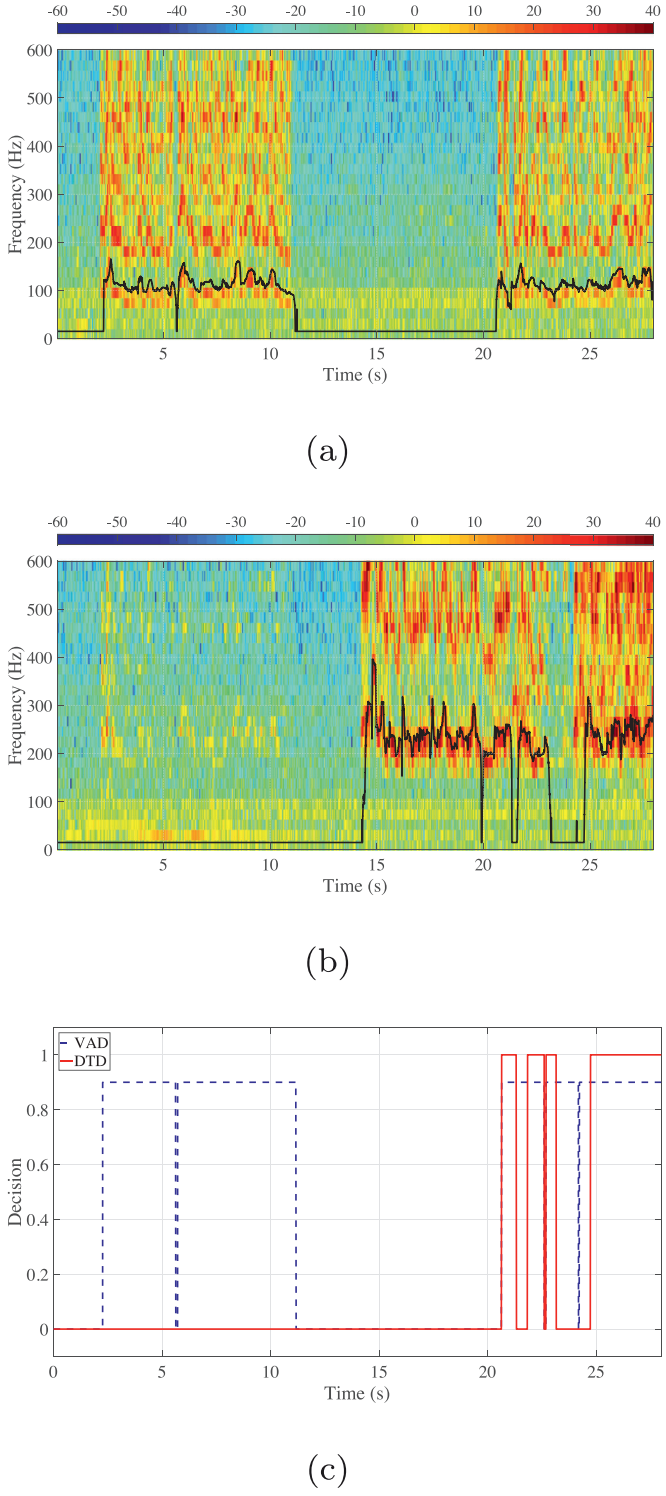


Fig. 12. Real test 1: tracking performance. (a) Remote human speaker. (b) Local human speaker. (c) VAD and DTD decisions.

- *decorrelation*: this NUTS implements the algorithm for multichannel decorrelation as described in Section 2.2.
- *echo paths A/B*: this NUTS performs the filtering to simulate the echo paths among loudspeakers and microphones.
- *systemID*: this NUTS implements system identification, also including VAD and DTD procedures as described in Sections 2.3 and 2.4, respectively.

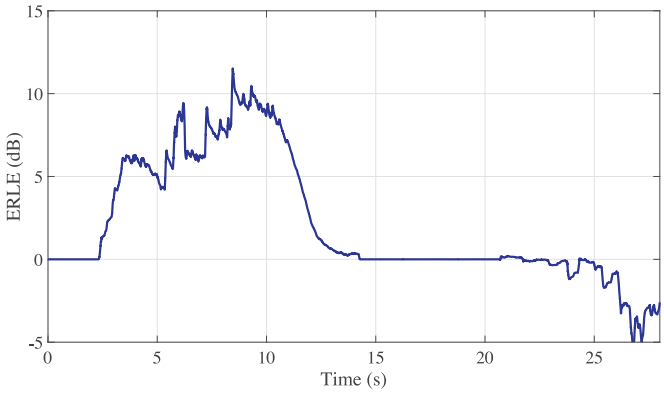


Fig. 13. Real test 1: ERLE.

Other minor NUTSs available from NU-Tech framework are also present in Fig. 3 for splitting the signal into multiple identical copies (*Split*), for performing the sum of the echo replicas and the local human speaker at each microphone (*Sum*), and for viewing the audio streaming (*Viewer*). Table 2 summarizes the values of the main system parameters assigned to optimize the trade-off between obtained performance and audio quality preservation. It is worth noting that the same values were considered for fundamental frequency estimation of remote human speaker and local human speaker, except for the contraction factor. In particular, a vector of contraction factor values linearly spaced between a lower bound and an upper bound (Table 2) was considered for multichannel decorrelation, whereas the same contraction factor value was used for all channels in the estimation of the fundamental frequency of the local human speaker. Regarding the sensitivity of the system performance to the parameters listed in Table 2, it is strictly related to the forgetting factor in fundamental frequency estimation. In this context, it is important to suitably set this parameter to avoid possible annoying distortion on the perceived signals and DTD dropouts due to very short silence periods during speech. The adopted algorithm to estimate the echo paths is the frequency-domain adaptive filtering algorithm as described in Cecchi et al. (2013). Different tests were carried out assuming both a simulated scenario and a real room. The obtained results are reported in terms of tracking of the remote and local human speakers to validate the VAD and DTD decision and echo return loss enhancement (ERLE) and misalignment between true and estimated echo paths to show echo cancellation performance. In particular, ERLE is calculated for each sample n as described by the following equation:

$$\mathrm{ERLE}(n) = 10\log_{10}\sum_{p=1}^{P}\frac{\mathrm{E}[d_p^2(n)]}{\mathrm{E}[e_p^2(n)]}, \tag{15}$$

where the expectations $\mathrm{E}[\cdot]$ are computed as follows

$$\mathrm{E}[d_p^2(n)] = \eta\mathrm{E}[d_p^2(n-1)] + (1-\eta)d_p^2(n) \tag{16}$$

$$\mathrm{E}[e_p^2(n)] = \eta\mathrm{E}[e_p^2(n-1)] + (1-\eta)e_p^2(n), \tag{17}$$

and $e_p(n)$ and $d_p(n)$ are the residual echo and microphone signals at microphone p and sample n, and $0 < \eta < 1$ is a forgetting factor. The misalignment is calculated for each time index n as (Herre et al., 2007):

$$\varepsilon(n) = 10\log_{10}\sum_{p=1}^{P}\frac{\sum_{l=1}^{L}\left\|h_{pl} - \widehat{h}_{pl}(n)\right\|^2}{\sum_{l=1}^{L}\left\|h_{pl}\right\|^2} \tag{18}$$

where $\widehat{h}_{pl}$ is the estimated coefficients vector of the true echo path h_{pl} between the lth loudspeaker and the pth microphone.

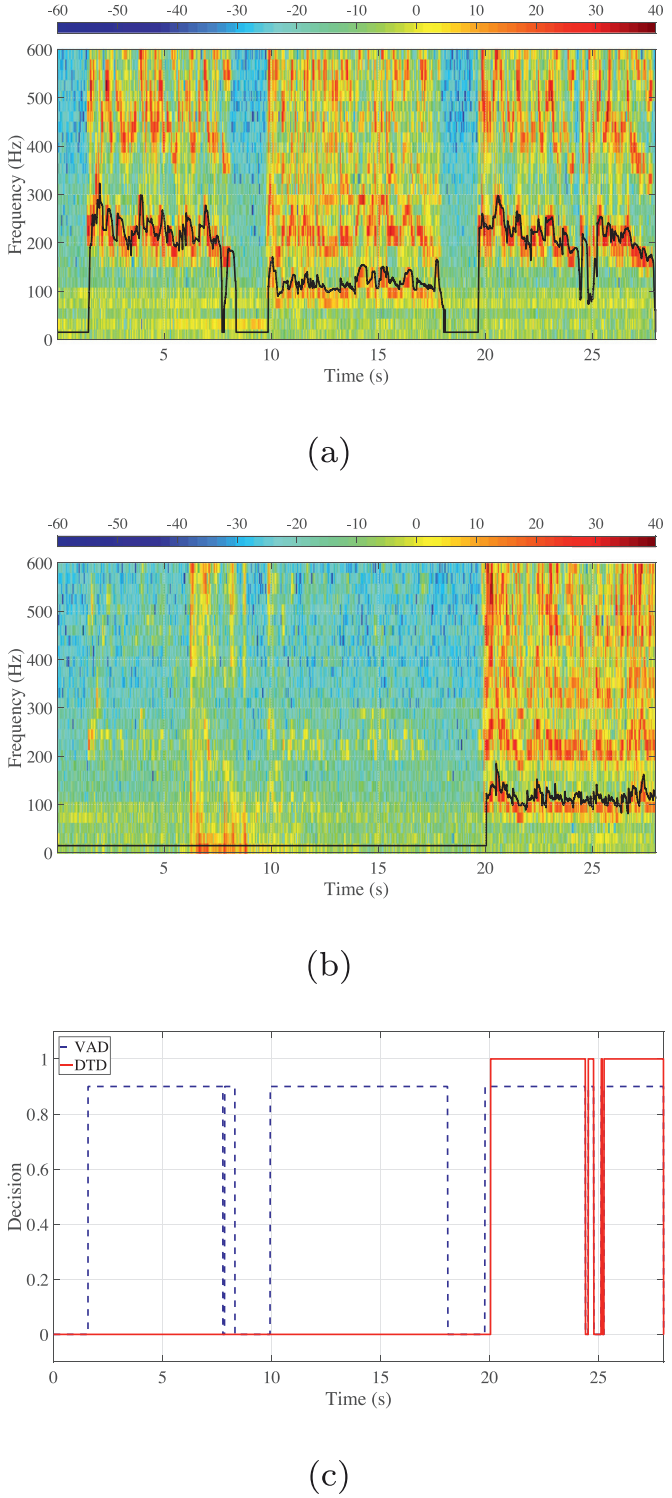


Fig. 14. Real Test 2: tracking performance. (a) Remote human speaker. (b) Local human speaker. (c) VAD and DTD decisions.

5.1. Simulated setup

The results reported in this section were obtained assuming a simulated scenario. Room A and Room B can be modelled as in Fig. 1 according to the image method (Allen and Berkeley, 1979). Room A has dimensions 3 m × 2 m × 2.5 m and reverberation time $RT_{60} = 0.35$ s and Room B has dimensions 3.5 m × 2.5 m × 2.5 m and $RT_{60} = 0.25$ s. Positions of microphones, loudspeakers, and human speakers are reported in Figs. 4(a) and (b). The acoustic coupling among 3 loudspeakers and 3 microphones was modelled by impulse responses truncated to the length of 171 ms. For example, the impulse response modelling the acoustic path between the loudspeaker L2 and the microphone M1 in Fig. 4(a) is shown in Fig. 5.

A first test session was carried out, considering different noise levels conditions. White Gaussian noise has been adopted during tests. A 28-second female speech (SQAM disk (European Broadcast Union, 1988)) has been used for the local human speaker in Room A, i.e., S1 in Fig. 4(a), whereas a 28-second male speech (SQAM disk (European Broadcast Union, 1988)) has been used for the local human speaker in Room B, i.e., S3 in Fig. 4(b), assuming a sampling frequency $f_s = 48$ kHz. The tracking capability is reported in Fig. 6 to validate the VAD and DTD decision, assuming a signal-to-noise ratio (SNR) of 40 dB. More specifically, the fundamental frequency estimated for the remote human speaker and the local human speaker in Room A is shown in Figs. 6(a) and (b), respectively. The estimated frequencies provide the correct VAD and DTD decisions reported in Fig. 6(c), where VAD decision was multiplied by a factor 0.9 for the sake of clarity. Then, the ERLE curves are reported in Figs. 7(a)–(d), lowering the SNR value between the echo replicas and the noise signal at microphones. As expected, the maximum obtainable ERLE decreases for higher level of noise, since the order of magnitude of the echo replicas becomes comparable with that of background noise. Figs. 8(a)–(d) show the obtained misalignment among true and estimated echo paths in order to provide the order of magnitude of the expected performance. In particular, it can be noticed that echo paths are effectively updated during periods of active remote human speakers while they are preserved during double-talk scenario. For the sake of completeness, it is worth noting that the CPU load measured in the NU-Tech framework is around 20% confirming the real-time performance of the proposed solution.

A second test session was carried out to show the performance of the proposed solution in particular recurring situations in teleconferencing systems. More specifically, an abrupt change in the echo paths occurs at time instant 6 s in Room B. Moreover, like during a conference, the far-end human speakers, i.e., S1 and S2 in Fig. 4(a), and the near-end human speakers, i.e., S3 and S4 in Fig. 4(b), keep on switching, considering female and male speech signals (SQAM disk (European Broadcast Union, 1988)) at a sampling frequency $f_s = 48$ kHz. The tracking capability is reported in Fig. 9 together with the VAD and DTD decisions in Fig. 9(c), where VAD decision was multiplied by a factor 0.9 for the sake of clarity, proving that the proposed algorithm for double-talk is able to distinguish between double-talk scenario and echo path change. In the presence of an abrupt change, the adaptive filters have to re-converge as reported in Fig. 10 in terms of obtained ERLE.

5.2. Real setup

The results reported in this section were obtained assuming the real scenario reported in Fig. 11 as local room, where the positions of loudspeakers, microphones, and human speakers are shown. Measurements were performed with professional equipment consisting of a professional M-AUDIO sound card connected to two M-Audio Studiophile AV 20 loudspeakers and to two C 400 BL microphones mounted on a board, that are managed by NU-Tech Framework. Multichannel acoustic echo cancellation was applied in the real room assuming a remote signal, preliminary recorded using the aforementioned equipment.

A first test session was carried out to validate VAD and DTD decisions and echo cancellation in the real room, considering loudspeakers L1 and L2 and microphones M11 and M21 in Fig. 11. A male speaker was first recorded in the room (S2 in Fig. 11). Then,

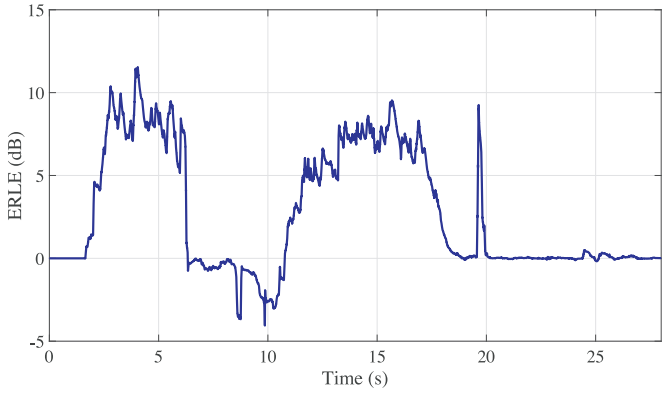


Fig. 15. Real test 2: ERLE.

the signal was reproduced in the room, where a female speaker (S1 in Fig. 11) started speaking after about 20 s. The tracking capability is reported in Figs. 12(a)–(c), where VAD decision was multiplied by a factor 0.9 for the sake of clarity. Then, ERLE is reported in Fig. 13. It can be noticed that filters update is halted for most of the time, thus, convergence is limited to a short interval.

A second test session was carried out to compare the response of the proposed solution to double-talk situation and echo path change. Switching male and female speakers (S2 and S1 in Fig. 11) were first recorded in the room. Then, the signal was reproduced in the room. Microphone positions were moved after about 6 s from M11 and M21 to M12 and M22 s and a male speaker (S2 in Fig. 11) started speaking at time instant 20 s. The tracking capability is reported in Figs. 14(a)–(c), where VAD decision was multiplied by a factor 0.9 for the sake of clarity, proving that the proposed DTD effectively distinguish echo paths change from double-talk. Filters have to re-converge after microphones movement as shown by ERLE curve in Fig. 15.

Finally, informal listening tests were carried out to provide a subjective evaluation of the proposed solution. Subjects were asked to judge the speech intelligibility and the naturalness and comfortableness of the conversation. Positive comments were provided by the involved listeners, thus, obtaining satisfactory results also from a subjective point of view.

6. Conclusion

A comprehensive solution for dealing with the main issues arising in multichannel acoustic echo cancellation has been presented in this paper. More specifically, starting from previous results on multichannel decorrelation approach, VAD algorithm, and DTD technique, the role of fundamental frequency in tracking of remote and local human speakers has been highlighted and a low-complexity self-contained solution for MAEC application has been obtained. The presented system focuses on the theory of pitch tracking through second-order adaptive notch filters for solving the aforementioned involved different aspects. The real-time performance of the approach was validated showing the correct fundamental frequency estimation and refinement and the convergence performance both in simulated and real setups. In particular, the reported ERLE proves the reduction of the undesired echoes derived from coupling among microphones and loudspeakers assuming different noise levels conditions and echo paths change. Informal listening tests were carried out to validate the proposed solution also from a subjective point of view, obtaining satisfactory results.

References

Allen, J.B., Berkeley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. 65, 943–950.
Atal, B., Rabiner, L., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE Trans. Acoust., Speech, Signal Process. 24 (3), 201–212.
Babu, C.G., Vanathi, P.T., 2009. Performance analysis of voice activity detection algorithms for robust speech recognition. Int. J. Comp. Science Comm. Tech. 2 (1), 288–293.
Benesty, J., Gaensler, T., Morgan, D.R., Sondhi, M.M., Gay, S.L., 2001. Advances in Network and Acoustic Echo Cancellation. Springer-Verlag.
Benesty, J., Gansler, T., 2002. A multichannel acoustic echo canceler double-talk detector based on a normalized cross-correlation matrix. Eur. Trans. Telecommun. 13 (2), 95–101.
Benesty, J., Morgan, D.R., Sondhi, M.M., 1998. A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation. IEEE Trans. Speech Audio Process. 6 (2), 156–165.
Cecchi, S., Romoli, L., Peretti, P., Piazza, F., 2011. A combined psychoacoustic approach for stereo acoustic echo cancellation. IEEE Trans. Audio, Speech Lang. Process. 19 (6), 1530–1539.
Cecchi, S., Romoli, L., Peretti, P., Piazza, F., 2012. Low-complexity implementation of a real-time decorrelation algorithm for stereophonic acoustic echo cancellation. Signal Process. 92 (11), 2668–2675.
Cecchi, S., Romoli, L., Piazza, F., 2016. Multichannel double-talk detector based on fundamental frequency estimation. IEEE Signal Process. Lett. 23 (1), 94–97.
Cecchi, S., Romoli, L., Piazza, F., Carini, A., 2013. A multichannel and multiple position adaptive room response equalizer in warped domain. Proceeding of 8th Int'l Symposium on Image and Signal Processing and Analysis. Trieste, Italy.
Davis, A., Nordholm, S., Togneri, R., 2006. Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold. IEEE Trans. Audio, Speech Lang. Proc. 14 (2), 412–424.
European Broadcast Union, 1988. Sound Quality Assessment Material Recordings for Subjective Tests. Tech. 3253-E.
Herre, J., Buchner, H., Kellermann, W., 2007. Acoustic Echo Cancellation for Surround Sound using Perceptually Motivated Convergence Enhancement. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 1, pp. 17–20.
Iqbal, M.A., Grant, S.L., Stokes, J.W., 2009. A frequency domain doubletalk detector based on cross-correlation and extension to multi-channel case. In: Proc. Proc. 43rd Asilomar Conference on Signals, Systems and Computers, pp. 638–641.
Larsen, E., Aarts, R.M., 2004. Audio Bandwidth Extension. J. Wiley & Sons.
Lattanzi, A., Bettarelli, F., Cecchi, S., 2008. NU-Tech: The Entry Tool of the hArtes Toolchain for Algorithms Design. In: Proc. 124th Audio Engineering Society Convention, Amsterdam, The Netherlands, pp. 1–8.
Ma, Y., Nishihara, A., 2013. Efficient voice activity detection algorithm using long-term spectral flatness measure. J. Audio, Speech, Music Proc. 21, 1–18.
Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech and Audio Proc. 10 (2), 109–118.
Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing. Prentice Hall International Inc. pp. 274–279.
Romoli, L., Cecchi, S., Comminiello, D., Piazza, F., Uncini, A., 2014. A novel decorrelation approach for an advanced multichannel acoustic echo cancellation system. In: Proc. 22nd European Signal Processing Conference, Lisbon, Portugal, pp. 651–655.
Romoli, L., Cecchi, S., Palestini, L., Peretti, P., Piazza, F., 2010. A Novel Approach to Channel Decorrelation for Stereo Acoustic Echo Cancellation based on Missing Fundamental Theory. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp. 329–332.
Romoli, L., Cecchi, S., Peretti, P., Piazza, F., 2012. A mixed decorrelation approach for stereo acoustic echo cancellation based on the estimation of the fundamental frequency. IEEE Trans. Audio, Speech Lang. Process. 20 (2), 690–698.
Romoli, L., Cecchi, S., Piazza, F., 2014. A novel decorrelation approach for multichannel system identification. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, pp. 6652–6656.
Romoli, L., Cecchi, S., Piazza, F., 2015. A voice activity detection algorithm for multichannel acoustic echo cancellation Exploiting Fundamental Frequency Estimation. In: Proc. 9th Int'l Symposium on Image and Signal Processing and Analysis. Zagreb, Croatia.
Shuyin, Z., Ying, G., Buhong, W., 2009. Auto-correlation property of speech and its application in voice activity detection. In: Proc. 1st Int. Workshop Education Technology and Computer Science, Wuhan, TBD, China, pp. 265–268.
Tashev, I.J., 2009. Sound Capture and Processing: Practical Approaches. J. Wiley & Sons.
Zölzer, U., 2002. Digital Audio Effects. J. Wiley & Sons.