A real-time speech enhancement framework in noisy and reverberated acoustic scenarios

Rudy Rotili, Emanuele Principi, Stefano Squartini, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Rotili, Rudy, Emanuele Principi, Stefano Squartini, and Björn Schuller. 2013. "A real-time speech enhancement framework in noisy and reverberated acoustic scenarios." *Cognitive Computation* 5 (4): 504–16. https://doi.org/10.1007/s12559-012-9176-x.

A Real-Time Speech Enhancement Framework in Noisy and Reverberated Acoustic Scenarios

Rudy Rotili · Emanuele Principi · Stefano Squartini · Björn Schuller

Abstract This paper deals with speech enhancement in noisy reverberated environments where multiple speakers are active. The authors propose an advanced real-time speech processing front-end aimed at automatically reducing the distortions introduced by room reverberation in distant speech signals, also considering the presence of background noise, and thus to achieve a significant improvement in speech quality for each speaker. The overall framework is composed of three cooperating blocks, each one fulfilling a specific task: speaker diarization, room impulse responses identification and speech dereverberation. In particular, the speaker diarization algorithm pilots the operations performed in the other two algorithmic stages, which have been suitably designed and parametrized to operate with noisy speech observations. Extensive computer simulations have been performed by using a subset of the AMI database under different realistic noisy and reverberated conditions. Obtained results show the effectiveness of the approach.

R. Rotili · E. Principi · S. Squartini
3MediaLabs, Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 1, 60131 Ancona, Italy
e-mail: sts@deit.univpm.it

R. Rotili
e-mail: r.rotili@univpm.it

E. Principi
e-mail: e.principi@univpm.it

B. Schuller
Institute for Human-Machine Communication, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany
e-mail: schuller@tum.de

Introduction

Speech-based Human–Machine interfaces have a large variety of application possibilities, as confirmed by the increasing scientific and commercial interest worldwide. A remarkable one is represented by multiparty meetings: Here, the speech signals have to be captured and processed in order to extract and likely interpret the information contained therein. The acoustic conditions in this kind of scenario are generally characterized by the presence of multiple active speakers (sometimes also simultaneously) in addition to the reverberation effect, due to convolution with room impulse responses (IRs) and the background noise. This results in a certain quality degradation of the acquired speech signals, and a strong signal processing intervention is required on purpose [39]. Moreover, another important issue in this type of systems is represented by the real-time constraints: The speech information often needs to be processed while the audio stream becomes available, making the complete task even more challenging.

Several solutions based on multiple-input multiple-output (MIMO) systems have been proposed in the literature to address the dereverberation problem under blind conditions [29]. The issue in multiparty meetings consists in coordinating the blind estimation of room IRs with the speech activity of different speakers, also taking the real-time constraints into account. That is why in this work a real-time speaker diarization algorithm has been implemented to inform when and how the blind channel estimation

algorithm has to operate. Once the IRs are estimated, the dereverberation algorithm can finalize the process and allows to yield speech signals of significantly improved quality. Furthermore, the information provided by the speaker diarizer allows the adaptive filters in the dereverberation algorithm to work only when speech segments of the same speaker occur at the same channel.

The authors [35, 37] have recently developed a real-time framework able to jointly separate and dereverberate signals in multi-talker environments, but the speaker diarization stage has been used at most as an oracle and not as a real algorithm. In [1, 36], the speaker diarization system has been included, but it is not able to work in blind mode, since it needs the knowledge of microphone position. The present contribution is aimed to face this issue and represents an additional step in the automatization process of the overall speech enhancement framework in real meeting scenarios.

Another important aspect that makes this contribution different from previous ones is the system capability to work in the presence of noise. There is a florid literature on noise reduction [2, 22, 25] and dereverberation [29] techniques in speech processing applications. Some scientists have recently developed speech enhancement algorithms able to jointly face both problems [8, 24, 42], but the issue of realizing a real-time framework operating with multiple speakers active in such realistic acoustic scenario has not been adequately addressed so far, up to the authors' knowledge. This justifies the paper objective, which is mainly targeted to suitably design the algorithms operating within the proposed speech enhancement framework to robustly behave also in the presence of background noise, always present in multiparty meeting scenarios.

In order to evaluate the achievable performance, several computer simulations under realistic noisy and reverberated acoustic conditions have been performed employing a subset of the AMI corpus [6]. The speech quality improvement, assessed by means of two different objective indexes, allowed the authors to positively conclude about the approach effectiveness.

The paper outline is the following. In "The proposed Speech Enhancement Framework," the overall speech enhancement framework, aimed at separating and dereverberating the speech sources, even in the presence of background noise, is described. "Computer Simulations" is targeted to discuss the experimental setup and performed computer simulations. Conclusions are drawn and future work ideas highlighted in "Conclusions".

The Proposed Speech Enhancement Framework

Assuming M independent speech sources and N microphones, the relationship between them is described by an $M \times N$ MIMO FIR (Finite Impulse Response) system. According to such a model and denoting $(\cdot)^T$ as the transpose operator, the following equations (in the time and z domain) for the n-th microphone signal hold:

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k, L_h) \tag{1}$$

$$X_n(z) = \sum_{m=1}^{M} H_{nm}(z) S_m(z), \tag{2}$$

where $\mathbf{h}_{nm} = [h_{nm,0}\ h_{nm,1}\ \ldots\ h_{nm,L_h-1}]^T$ is the L_h-taps IR between the n-th microphone and m-th source $\mathbf{s}_m(k, L_h) = [s_m(k)\ s_m(k-1)\ \ldots\ s_m(k-L_h+1)]^T$, with $(m = 1, 2, \ldots, M,\ n = 1, 2, \ldots, N)$. The objective is to recover the original clean speech sources by means of a "context-aware" speech dereverberation approach: Indeed, such a technique has to automatically identify who is speaking, accordingly estimate the unknown room IRs and then apply a knowledgeable dereverberation process to restore the original speech quality. To achieve such a goal, the proposed framework consists of three main stages: speaker diarization (SDiar), blind channel identification (BCI) and speech dereverberation (SDer). As aforementioned, the proposed approach recalls what already published by the same authors of this contribution in the recent past [35, 37], but with two noteworthy differences:

- A real speaker diarization algorithm has never been included into the speech enhancement framework operating in multiparty meetings: Indeed in [37], the SDiar has been assumed to operate according to an oracle fashion. Here, SDiar takes as input the microphone observables and for each frame, the output $\mathcal{P}_m$ is 1 if the m-th source is the only active, and 0 otherwise. In such a way, the framework is able to detect when to perform or not to perform the required operation. Both the BCI and the SDer take advantage of this information, activating the estimation and the dereverberation process, respectively, only when the right speaker is present in the right channel. It is important to point out that the usage of speaker diarization algorithm allows to consider the system composed by the only active source and the N microphones as a single-input multiple-output (SIMO), which can be blindly identified in order to perform the dereverberation process;
- The presence of noise has never been addressed in previous publications: The authors want to show in this work that the proposed framework is able to efficiently operate also under such acoustic conditions, making the simulated scenario even more realistic. Some modifications have been applied to the original BCI and SDer algorithms in order to suitably deal with noise presence, as it will be clearer later on.

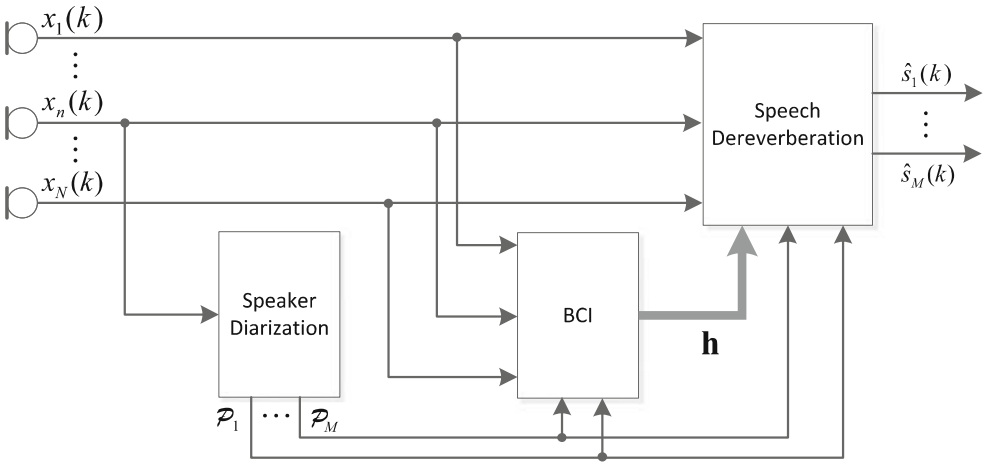

Fig. 1 Block diagram of the proposed framework

The block diagram of the proposed framework is shown in Fig. 1. The three aforementioned algorithmic stages are now briefly described.

Blind Channel Identification Stage

Considering a SIMO system for a specific source s_{m^*}, a BCI algorithm aims to find the IRs vector $\mathbf{h}_{nm^*} = [\mathbf{h}_{1m^*}^T \ \mathbf{h}_{2m^*}^T \ \cdots \mathbf{h}_{Nm^*}^T]^T$ by using only the microphone signals $x_n(k)$. In order to ensure this, two identifiability conditions are assumed satisfied [46]:

1. The polynomial formed from $\mathbf{h}_{nm^*}$ are co-prime, that is, the room transfer functions (RTFs) $H_{nm^*}(z)$ do not share any common zeros (channel diversity);
2. $\mathcal{C}\{s(k)\} \geq 2L_h + 1$, where $\mathcal{C}\{s(k)\}$ denotes the linear complexity of the sequence $s(k)$.

This stage performs the BCI through the robust normalized multichannel frequency-domain least mean square (RNMCFLMS) algorithm [15] that is a noise robust version of the popular UNMCFLMS [18]. In addition, it is well suited to satisfy the real-time constraints imposed by the case study since it offers a good compromise between fast convergence, adaptivity and low computational complexity.

Here, a brief review of the UNMCFLMS and of the RNMCFLMS algorithms is reported in order to understand the motivation of this choice in the proposed front-end. Refer to [18] and [15] for details.

The derivation of UNMCFLMS is based on cross-relation criteria [46] using the overlap-save technique [30]. The frequency-domain cost function for the q-th frame is defined as

$$J_f(q) = \sum_{n=1}^{N-1} \sum_{i=n+1}^{N} \underline{\mathbf{e}}_{ni}^H(q)\underline{\mathbf{e}}_{ni}(q) \tag{3}$$

where $\underline{\mathbf{e}}_{ni}(q)$ is the frequency-domain block error signal between the n-th and i-th channels, and $(\cdot)^H$ denotes the Hermitian transpose operator. The update equation of the UNMCFLMS is expressed as

$$\begin{aligned}\widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q+1) = \widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q) - \rho[\mathbf{P}_{nm^*}(q) + \delta\mathbf{I}_{2L_h\times L_h}]^{-1} \\ \times \sum_{n=1}^{N} \mathbf{D}_{x_n}^H(q)\underline{\mathbf{e}}_{ni}^{10}(q), \quad i = 1,\ldots,N\end{aligned} \tag{4}$$

where $0 < \rho < 2$ is the step-size, δ is a small positive number, $\widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q) = \mathbf{F}_{2L_h\times 2L_h}\left[\widehat{\mathbf{h}}_{nm^*}(q)\ \mathbf{0}_{1\times L_h}\right]^T$, $\underline{\mathbf{e}}_{ni}^{10}(q) = \mathbf{F}_{2L_h\times 2L_h}\left[\mathbf{0}_{1\times L_h}\left\{\mathbf{F}_{L_h\times L_h}^{-1}\underline{\mathbf{e}}_{ni}(q)\right\}^T\right]^T$, $\mathbf{P}_{nm^*}(q) = \sum_{n=1,n\neq i}^{N} \mathbf{D}_{x_n}^H(q)\mathbf{D}_{x_n}(q)$, and $\mathbf{F}$ denotes the discrete Fourier transform (DFT) matrix. The frequency-domain error function $\underline{\mathbf{e}}_{ni}(q)$ is given by

$$\underline{\mathbf{e}}_{ni}(q) = \mathbf{D}_{x_n}(q)\widehat{\underline{\mathbf{h}}}_{nm^*}(q) - \mathbf{D}_{x_i}(q)\widehat{\underline{\mathbf{h}}}_{im^*}(q) \tag{5}$$

where the diagonal matrix

$$\begin{aligned}\mathbf{D}_{x_n}(q) = \mathrm{diag}\big(\mathbf{F}\{[x_n(qL_h - L_h)\ x_n(qL_h - L_h + 1)\cdots \\ x_n(qL_h + L_h - 1)]^T\}\big)\end{aligned} \tag{6}$$

is the DFT of the q-th frame input signal block for the n-th channel.

From a computational point of view, the UNMCFLMS algorithm ensures an efficient execution of the circular convolution by means of the fast Fourier transform (FFT). In addition, it can be easily implemented for a real-time application since the normalization matrix $\mathbf{P}_{nm^*}(q) + \delta\mathbf{I}_{2L_h\times L_h}$ is diagonal, and it is straightforward to compute its inverse.

Though UNMCFLMS allows the estimation of long IRs, it requires a high input signal-to-noise ratio (SNR). When additive noise is present, it is possible to see that the UNMCFLMS rapidly diverges from the optimal solution. Such a misconvergence is associated with the non-uniform

spectral attenuation of the estimated impulse response [14, 16]. In order to avoid this problem in [15], it is proposed to use a modified cost function defined as

$$J_{\mathrm{mod}}(q) = J_f(q) + \beta(q) J_p(q) \tag{7}$$

where $J_f(q)$ and $J_p(q)$ are the original and penalty cost functions, respectively, and $\beta(q)$ is the Lagrange multiplier. The penalty cost function is formulated as

$$\text{maximize} \quad J_p(q) = \prod_{i=1}^{NL_h} \left| \underline{\widehat{h}}_i(q) \right|^2 \tag{8}$$

$$\text{subject to} \quad \left| \underline{\widehat{h}}_1(q) \right|^2 + \left| \underline{\widehat{h}}_2(q) \right|^2 + \cdots + \left| \underline{\widehat{h}}_{NL_h}(q) \right|^2 = \frac{1}{NL_h} \tag{9}$$

where Eq. 9 comes from the unit norm constraint imposed in the previous update equation. The coupling factor, $\beta(q)$, is estimated such that the gradient of $J_{\mathrm{mod}}(q)$ becomes zero in the steady-state condition.

This gives $\nabla J_f(q) = \beta(q) \nabla J_p(q)$, and premultiplying both sides by $J_p^H(q)$, it is possible to obtain $\beta(q)$ as

$$\beta(q) = \left| \frac{\nabla J_p^H(q) \nabla J_f(q)}{\left\| \nabla J_p(q) \right\|^2} \right|. \tag{10}$$

Considering the modified cost function of Eq. 7, the update equation for the RNMCFLMS is then

$$\begin{aligned} \underline{\widehat{\mathbf{h}}}_{nm^*}^{10}(q+1) = \underline{\widehat{\mathbf{h}}}_{nm^*}^{10}(q) - \rho [\mathbf{P}_{nm^*}(q) + \delta \mathbf{I}_{2L_h \times L_h}]^{-1} \\ \times \sum_{n=1}^{N} \mathbf{D}_{x_n}^H(q) \underline{\mathbf{e}}_{ni}^{10}(q) \\ + \rho \beta_k(q) \nabla J_{p_{nm^*}}(q), \quad i = 1, \ldots, N \end{aligned} \tag{11}$$

where $\beta_k(q)$ is estimated similar to Eq. 10 but using the UNMCFLMS algorithm update parameters.

Speech Dereverberation Stage

Given the SIMO system $H_{nm^*}(z)$ corresponding to the specific source s_{m^*}, a set of inverse filters $G_{nm^*}(z)$ can be found by using the multiple-input/output inverse theorem (MINT) [27] such that

$$\sum_{n=1}^{N} H_{nm^*}(z) G_{nm^*}(z) = 1, \tag{12}$$

assuming that the RTFs do not have any common zeros. In the time-domain, the inverse filter vector denoted as $\mathbf{g}_{m^*}$ is calculated by minimizing the following cost function:

$$C = \left\| \mathbf{H}_{m^*} \mathbf{g}_{m^*} - \mathbf{v} \right\|^2, \tag{13}$$

where $\| \cdot \|$ denote the l_2-norm operator and

$$\mathbf{g}_{m^*} = [g_{1m^*}(1), \cdots, g_{1m^*}(L_i), \cdots, g_{Nm^*}(1), \cdots, g_{Nm^*}(L_i)]^T, \tag{14}$$

$$\mathbf{v} = [\underbrace{0, \cdots, 0}_{d}, 1, \cdots, 0]^T. \tag{15}$$

The vector $\mathbf{v}$ is the target vector, that is, the Kronecker delta shifted by an appropriate modeling delay ($0 \le d \le NL_i$) while $\mathbf{H}_{m^*} = [\mathbf{H}_{1m^*}, \cdots, \mathbf{H}_{Nm^*}]$ where $\mathbf{H}_{nm^*}$ is the convolution matrix of the IR between the source s_m^* and n-th microphone. When the matrix $\mathbf{H}_{m^*}$ is given, the inverse filter set can be calculated as

$$\mathbf{g}_{m^*} = \mathbf{H}_{m^*}^{\dagger} \mathbf{v} \tag{16}$$

where $(\cdot)^{\dagger}$ denotes the Moore–Penrose pseudoinverse. By setting L_i so that matrix $\mathbf{H}_{m^*}$ is square, a filter set with the minimum length is obtained.

Considering the presence of disturbances, that is, additive noise or RTFs fluctuations, the cost function Eq. 13 is modified as follows [17]:

$$C = \left\| \mathbf{H}_{m^*} \mathbf{g}_{m^*} - \mathbf{v} \right\|^2 + \gamma \left\| \mathbf{g}_{m^*} \right\|^2, \tag{17}$$

where the parameter $\gamma (\ge 0)$, called regularization parameter, is a scalar coefficient representing the weight assigned to the disturbance term. It should be noticed that Eq. 17 has the same form of Tikhonov regularization for ill-posed problems [9]. Its value has been set to 0.1 in all simulations described in the following.

Let the RTF for the fluctuation case be given by the sum of two terms, the mean RTF ($\overline{\mathbf{H}}_{m^*}$) and the fluctuation from the mean RTF ($\widetilde{\mathbf{H}}_{m^*}$), and let $E\langle \widetilde{\mathbf{H}}_{m^*}^T \widetilde{\mathbf{H}}_{m^*} \rangle = \gamma \mathbf{I}$. In this case, a general cost function, embedding noise and fluctuation case, can be derived:

$$C = \mathbf{g}_{m^*}^T \mathcal{H}^T \mathcal{H} \mathbf{g}_{m^*} - \mathbf{g}_{m^*}^T \mathcal{H}^T \mathbf{v} - \mathbf{v}^T \mathcal{H} \mathbf{g}_{m^*} + \mathbf{v}^T \mathbf{v} + \gamma \mathbf{g}_{m^*}^T \mathbf{g}_{m^*} \tag{18}$$

where

$$\mathcal{H} = \begin{cases} \mathbf{H}_{m^*} & (\text{noise case}) \\ \overline{\mathbf{H}}_{m^*} & (\text{fluctuation case}). \end{cases} \tag{19}$$

The filter that minimizes the cost function in Eq. 18 is obtained by taking derivatives with respect to $\mathbf{g}_{m^*}$ and setting them equal to zero. The required solution is

$$\mathbf{g}_{m^*} = \left(\mathcal{H}^T \mathcal{H} + \gamma \mathbf{I} \right)^{-1} \mathcal{H}^T \mathbf{v}. \tag{20}$$

The usage of Eq. 20 to calculate the inverse filters requires a matrix inversion that, in the case of long IRs, can result in a high computational burden. Instead, an adaptive algorithm [34] has been adopted to satisfy the real-time constraints. It is based on the well-known steepest-descent technique, whose recursive estimator has the form

$$\mathbf{g}_{m^*}(q+1) = \mathbf{g}_{m^*}(q) - \frac{\mu(q)}{2}\nabla C. \tag{21}$$

Moving from Eq. 18 through simple algebraic calculations, the following expression is obtained:

$$\nabla C = -2[\mathcal{H}^T(\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)) - \gamma\,\mathbf{g}_{m^*}(q)]. \tag{22}$$

Substituting Eq. 22 into Eq. 21, the following holds

$$\mathbf{g}_{m^*}(q+1) = \mathbf{g}_{m^*}(q) + \mu(q)[\mathcal{H}^T(\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)) - \gamma\mathbf{g}_{m^*}(q)] \tag{23}$$

where $\mu(q)$ is the step-size.

The convergence of the algorithm to the optimal solution is guaranteed if the usual conditions for the step-size in terms of autocorrelation matrix $\mathcal{H}^T\mathcal{H}$ hold. However, the achievement of the optimum can be slow if a fixed step-size value is chosen. The algorithm convergence speed can be increased following the approach in [12], where the step-size is chosen in order to minimize the cost function at the next iteration. The analytical expression obtained for the step-size is the following:

$$\mu(q) = \frac{\mathbf{e}^T(q)\mathbf{e}(q)}{\mathbf{e}^T(q)\left(\mathcal{H}^T\mathcal{H} + \gamma I\right)\mathbf{e}(q)} \tag{24}$$

where

$$\mathbf{e}(q) = \mathcal{H}^T[\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)] - \gamma\mathbf{g}_{m^*}(q).$$

In using the previously illustrated algorithm, different advantages are obtained: The regularization parameter, which takes into account the presence of disturbances, makes the dereverberation process more robust to noise and to estimation errors due to the BCI algorithm [17]; the real-time constraint can be met also in the case of long IRs since no matrix inversion is required. Finally, the complexity of the algorithm has been decreased computing the required operation in the frequency-domain using FFTs.

Speaker Diarization Stage

The algorithm taken here as reference is the one proposed in [41], which consists in segmenting live recorded audio into speaker-homogeneous regions with the goal of answering the question "who is speaking now?". Current state-of-the-art speaker diarization systems are based on clustering approaches, usually combining hidden Markov models (HMMs) and the Bayesian Information Criterion metric [11, 45]. Despite their state-of-the-art performance, such systems have the drawback of operating on the entire signals, making them unsuitable to work online as required by the proposed framework. For the system to work online, the question has to be answered on small chunks of the recorded audio data, and the decisions must not take longer than real time. In order to do that, two distinct operating modes are foreseen for the SDiar system (Fig. 2): the training and the online recognition one.

In training mode, the user is asked to speak for one minute. The voice is recorded and transformed in the mel-frequency cepstral coefficient (MFCC) features space. The speech segments detected by means of a ground-truth Voice Activity Detector (acting as SDiar entry-algorithm in both operating modes) are then used to train a Gaussian mixture model (GMM), by means of the expectation–maximization (EM) algorithm. The number of Gaussians is 100 and the accuracy threshold value (to stop EM iterations) equal to 10^{-4}. Such values have been empirically determined on meetings IS1004a-d of the AMI corpus.

In the actual recognition mode, the system records and processes chunks of audio as follows: in the first stage, MFCC features are extracted and cepstral mean subtraction (CMS) is applied, to deal with stationary channel effects. In the subsequent classification step, the likelihood for each set of features is computed against each set of Gaussian Mixtures obtained in the training step. As stated in [41], 2 s chunks of audio and a frame-length of 25 ms (with frame-shift equal to 10 ms) have been used, meaning that a total of 200 frames are examined to determine if an audio segment belongs to a certain speaker in the non-speech model. The decision is reached using majority vote on the likelihoods: Every feature vector in the current segment is assigned to one of the known speaker model based on the maximum likelihood criterion. The model that has the

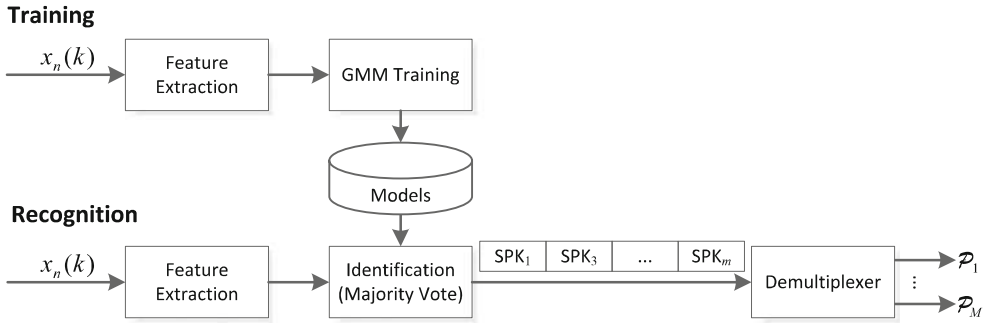

Fig. 2 The speaker diarization block scheme: "SPK$_m$" are the speaker identities labels assigned to each chunk

majority of vectors assigned determines the speaker identity on the current segment.

The "Demultiplexer" block shown in Fig. 2 associates each speaker label to a distinct output and sets it to "1" if the speaker is the only active, and "0" otherwise.

Computer Simulations

The overall framework depicted in Fig. 1 has been developed on a freeware software platform, namely NU-Tech [3], suitable for real-time audio processing.[1]

The acoustic scenario under study consists of an array of five microphones placed on the meeting table (located in a small office) and four speakers around them, as depicted in Fig. 3. A similar setup is used in the AMI [6] sub-corpus addressed in simulations described later on. Such a sub-corpus contains the "IS" meetings, well suited for the evaluation of algorithms working in multiparty conversational speech scenarios: Indeed, they have been used in [41] to test the performance of the speaker diarization system.

The headset recordings of this database have been used as original speech sources and then convolved with IRs generated using the RIR Generator tool [13], thus synthetically generating the microphone signals. As background noise added at the microphone level, two options have been chosen: white and colored (pink), both uncorrelated over the different channels. Three different reverberation conditions have been taken into account corresponding to T_{60} = 120, 240, 360 ms, respectively, with IRs 1024 taps long. Four different (segmental) SNR values have been considered, that is, SNR = 10, 20, 30, 40 dB. SNR = ∞ stands for the no-additive noise case study. FIR filters used in BCI and SDer stages are as long as the simulated IRs: The real-time factor corresponding to this parametrization is equal to 0.6, split into 0.15 for SDiar and 0.45 for both BCI and SDer. It must be noted that the extra-computational burden due to the employment of noise robust algorithms is negligible [15], and therefore, the real-time factor does not depend on the choices made above for the speech enhancement framework stages.

Two quality indexes have been considered for evaluation purposes. The first one is the normalized segmental signal-to-reverberation ratio (NSegSRR), which is defined as follows [29]:

[1] NU-Tech allows the developer to focus on the algorithm implementation without worrying about the interface with the sound card. The ASIO protocol is supported to guarantee low latency times. NU-Tech architecture is plug-in based: An algorithm can be implemented in C++ language to create a NUTS (NU-Tech Satellite) that can be plugged in the graphical user interface.

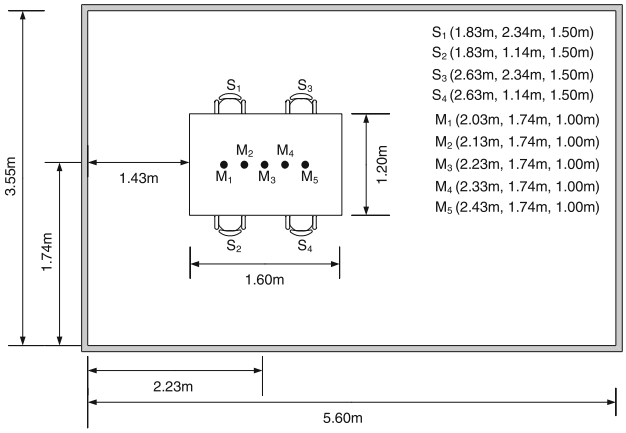

Fig. 3 Room setup

Table 1 Diarization Error Rates (in percentage) for non-processed audio files of meeting *IS1009b* in all addressed acoustic conditions in the presence of additive white and colored noise

SNR (dB)						
T_{60} (ms)	10	20	30	40	∞	Average
White noise						
120	28.21	16.50	11.67	8.36	6.36	14.22
240	32.20	17.29	11.32	9.38	6.61	15.36
360	34.38	19.77	12.68	9.52	7.16	16.70
Average	31.60	17.85	11.89	9.09	6.71	15.43
Colored noise						
120	30.02	16.82	11.17	8.08	6.36	14.49
240	35.05	19.02	11.09	8.57	6.61	16.07
360	34.53	20.39	11.38	7.7	7.16	16.23
Average	33.20	18.74	11.21	8.12	6.71	15.60

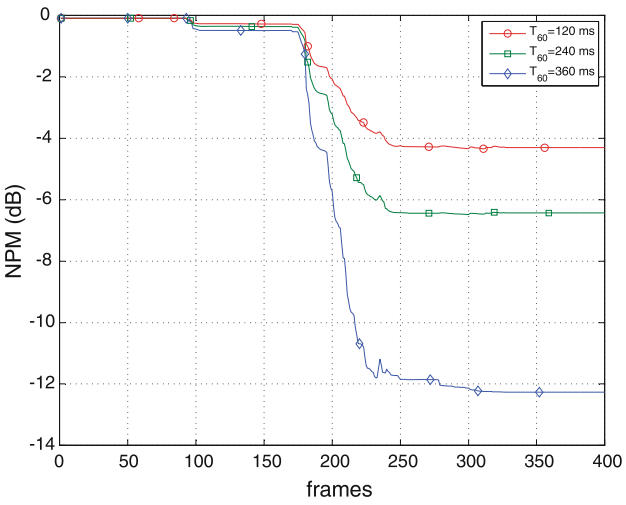

Fig. 4 NPM convergence performance over three different reverberation case studies. The SNR is equal to ∞

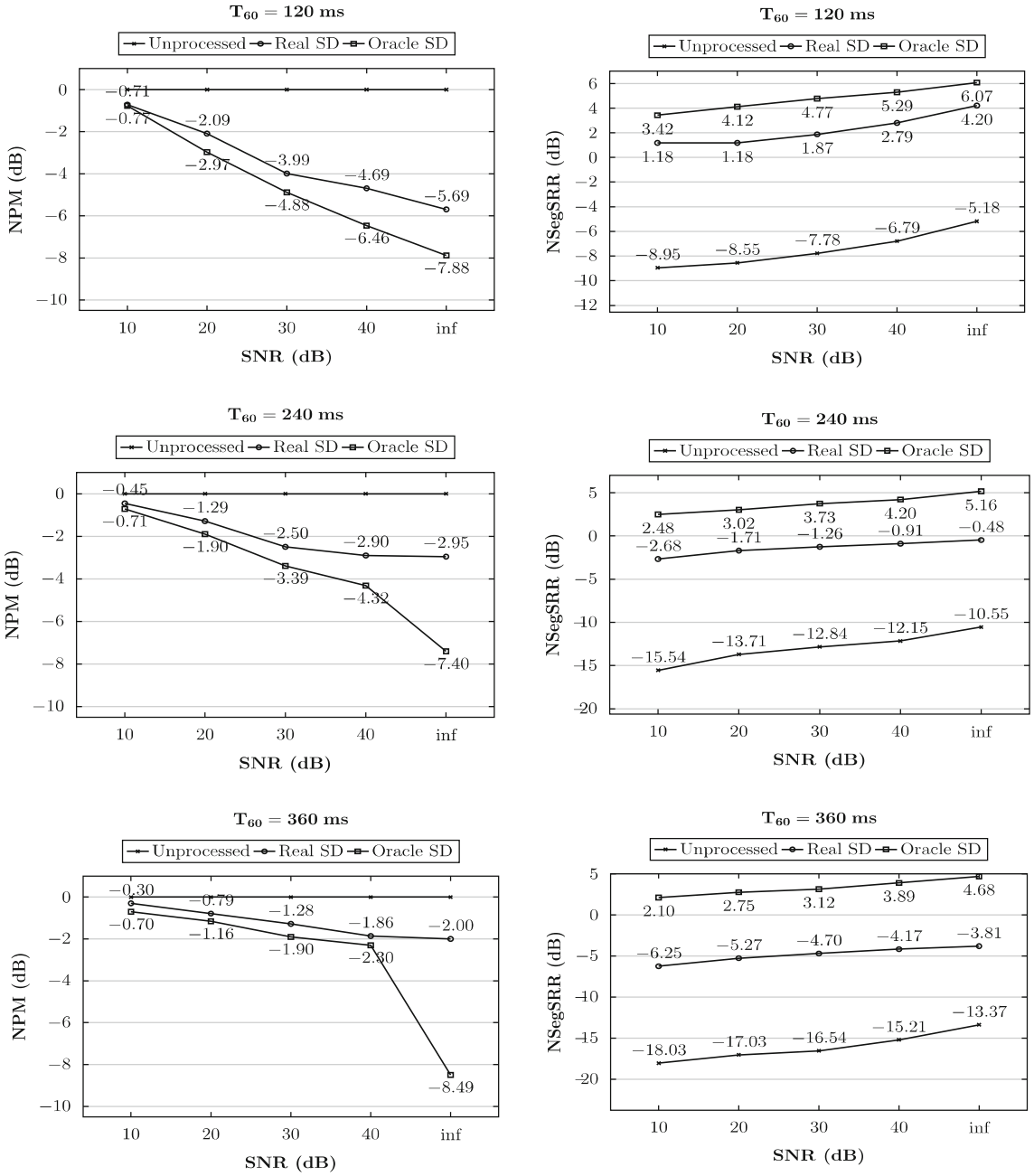

Fig. 5 White noise case study. NPM (*left column*) and NSegSRR (*right column*) performance of all addressed speech enhancement framework configurations under the different acoustic conditions. Averaged values over all speakers are considered

$$\text{NSegSRR} = 10\log_{10}\left(\frac{\|\mathbf{s}_m\|_2}{\|(1/\alpha)\hat{\mathbf{s}}_m - \mathbf{s}_m\|_2}\right), \quad m = 1,\ldots,M \tag{25}$$

where, $\mathbf{s}_m$ and $\hat{\mathbf{s}}_m$ are the desired direct-path signal and recovered speech signal, respectively, and α is a scalar assumed stationary over the duration of the measurement. Of course, in calculating the NSegSRR value, the involved signals are assumed to be time-aligned. The higher the NSegSRR value, the better it is.

For the BCI stage, a channel-based measure called normalized projection misalignment (NPM) [28] is employed:

$$\text{NPM}(q) = 20\log_{10}\left(\frac{\|\epsilon(q)\|}{\|\mathbf{h}\|}\right), \tag{26}$$

where

$$\epsilon(q) = \mathbf{h} - \frac{\mathbf{h}^T\widehat{\mathbf{h}}(q)}{\widehat{\mathbf{h}}^T(q)\widehat{\mathbf{h}}(q)}\widehat{\mathbf{h}}(q) \tag{27}$$

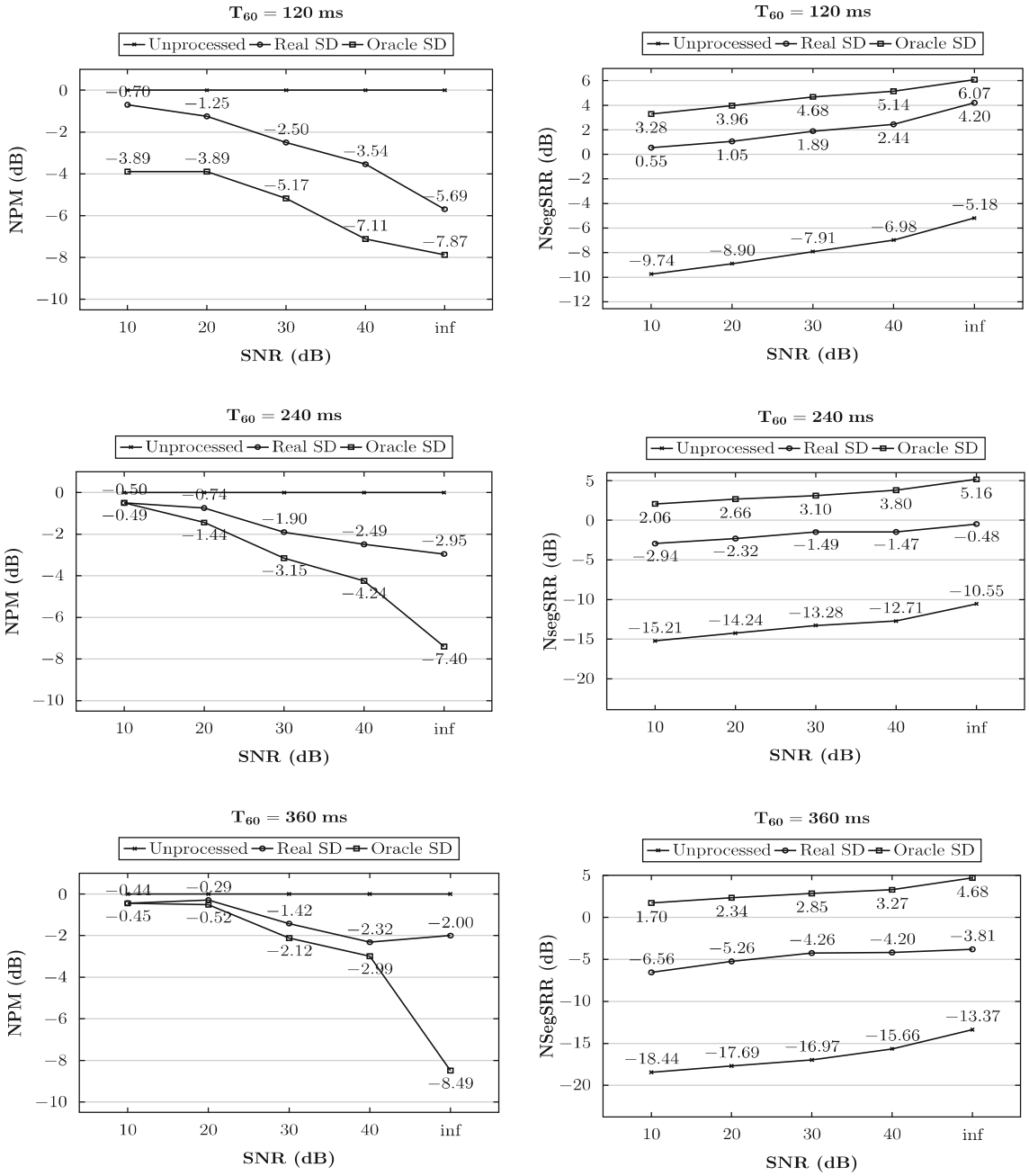

Fig. 6 Colored noise case study. NPM (left column) and NSegSRR (right column) performance of all addressed speech enhancement framework configurations under the different acoustic conditions. Averaged values over all speakers are considered

is the projection misalignment vector, $\mathbf{h}$ is the real IR vector, whereas $\widehat{\mathbf{h}}(q)$ is the estimated one at the q-th iteration, that is, the frame index. In this case, the lower the NPM value, the better it is.

Experimental Results

Computer simulations discussed in this section are related to the meeting *IS1009b* of the corpus [6]. It has a total length of 33 m and 15 s, and all the four participants are female speakers. The amount of speaking time for each speaker, including overlap, is 7 m and 47 s, 5 m and 10 s, 7 m and 20 s, 9 m and 0 s for speaker s_1, s_2, s_3 and s_4, respectively, whereas the total overlap is 3 m and 5 s.

As stated in previous section, twelve distinct acoustic scenarios have been addressed for each noise type, corresponding to the combination of the three aforementioned T_{60} and SNR values: For each of them, the non-processed and processed cases have been evaluated. Moreover, two

Table 2 NSegSRR values for non-processed audio files of meeting *IS1009b* (containing four female speakers) in all reverberation and SNR conditions–White noise case study

T_{60} (ms)	s_1	s_2	s_3	s_4
SNR = 10 dB				
120	−8.18	−7.23	−12.44	−7.93
240	−10.24	−8.78	−26.60	−16.53
360	−11.06	−9.42	−32.42	−19.24
SNR = 20 dB				
120	−7.97	−6.95	−11.86	−7.43
240	−8.91	−8.62	−21.31	−16.01
360	−10.46	−9.01	−30.48	−18.16
SNR = 30 dB				
120	−7.91	−6.48	−10.10	−6.62
240	−8.38	−7.96	−20.64	−14.38
360	−10.21	−8.46	−29.62	−17.87
SNR = 40 dB				
120	−6.84	−5.88	−9.04	−5.39
240	−8.26	−7.12	−19.74	−13.46
360	−9.06	−7.91	−28.45	−15.41
SNR = ∞				
120	−4.98	−4.77	−6.78	−4.18
240	−6.11	−6.45	−20.06	−9.59
360	−6.61	−7.56	−27.55	−11.76

operating modes for the SDiar system have been considered: *oracle* (diarization coincides with manual AMI annotations) and *real* (speakers' activity is detected by means of the algorithm described in "Speaker Diarization Stage")

The SDiar performance has been measured by the diarization error rate[2] (DER). DER is defined by the following expression:

$$\mathrm{DER} = \frac{\sum_{s=1}^{S} \mathrm{dur(s)}(\max(\mathrm{N_{ref}(s)}, \mathrm{N_{hyp}(s)}) - \mathrm{N_{correct}(s)})}{\sum_{s=1}^{S} \mathrm{dur}(s) N_{\mathrm{ref}}(s)} \tag{28}$$

where S is the total number of segments in which no speaker change occurs, $N_{\mathrm{ref}}(s)$ and $N_{\mathrm{hyp}}(s)$ indicate the number of speakers in the reference and in the hypothesis, and $N_{\mathrm{correct}}(s)$ indicates the number of speakers that speak in the segment s and have been correctly matched between the reference and the hypothesis. As recommended by the National Institute for Standards and Technology (NIST), evaluation has been performed by means of the "md-eval" tool with a collar of 0.25 s around each segment to take into account timing errors in the reference.

Simulations have been accomplished in all different acoustic scenarios, and related results are shown in Table 1: They clearly show a strong dependence on the presence of noise (both for white and colored case studies) and a lower but still clear dependence on the reverberation. The error obtained in the clean speech case study is equal to 6.51%.

The behavior of the BCI algorithm has been preliminarily addressed on an audio file containing speech from a single speaker. Figure 4 shows its rate of convergence in the three different reverberation cases when SNR = ∞. NPM values have to be referred to an initial value of about 0 dB, obtained by initializing the overall channel IRs vector to satisfy the unit norm constraint [18]. Curves related to other SNR case studies have not been depicted for the sake of conciseness: However, NPM values obtained at convergence are reported for all acoustic scenarios in tables below. BCI algorithm convergence is assumed to be reached in the last two seconds of speaking activity.

Then, the overall system performance has been finally evaluated and related results are reported in Figs. 5, 6. They compare the performance of the speech enhancement framework (both *oracle* and *real* SDiar operating modes) and the non-processing option (i.e., when the proposed

Table 3 *Oracle Speaker Diarization case study:* NPM and NSegSRR values for dereverberated audio files of meeting *IS1009b* (containing four female speakers) in all reverberation and SNR conditions–White noise case study

T_{60} (ms)	NPM (dB)				NSegSRR (dB)			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
SNR = 10 dB								
120	−0.94	−0.58	−0.61	−0.98	3.73	4.14	3.51	2.29
240	−0.88	−0.60	−0.58	−0.79	3.44	0.92	3.38	2.19
360	−0.79	−0.77	−0.56	−0.66	3.11	0.33	2.96	2.02
SNR = 20 dB								
120	−5.24	−2.42	−1.31	−2.89	4.34	4.87	3.95	3.31
240	−3.15	−1.47	−1.12	−1.87	3.97	1.06	3.78	3.27
360	−1.91	−0.81	−0.69	−1.23	3.70	0.53	3.24	3.54
SNR = 30 dB								
120	−7.40	−1.77	−4.55	−5.8	5.78	4.91	4.23	4.16
240	−5.48	−1.44	−2.7	−3.93	5.58	1.15	4.17	4.03
360	−3.25	−1.13	−1.12	−2.08	4.36	0.78	3.36	3.97
SNR = 40 dB								
120	−8.9	−2.97	−6.04	−7.92	6.21	5.41	4.64	4.89
240	−6.08	−2.04	−3.78	−5.37	6.28	1.22	4.61	4.69
360	−3.75	−1.45	−1.21	−2.80	6.07	0.84	4.01	4.62
SNR = ∞								
120	−13.23	−3.09	−6.16	−9.02	6.65	5.83	5.11	6.67
240	−10.96	−1.70	−6.74	−10.19	7.00	1.29	5.68	6.69
360	−11.52	−1.90	−7.83	−12.69	6.87	1.07	5.25	5.54

[2] http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/

Table 4 *Real Speaker Diarization case study:* NPM and NSegSRR values for dereverberated audio files of meeting *IS1009b* (containing four female speakers) in all reverberation and SNR conditions–White noise case study

T_{60} (ms)	NPM (dB)				NSegSRR (dB)			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
SNR = 10 dB								
120	−0.75	−0.88	−0.60	−0.60	0.84	1.17	0.68	0.46
240	−0.64	−0.4	−0.25	−0.49	−2.49	−2.83	−2.32	−3.09
360	−0.17	−0.41	−0.19	−0.43	−4.97	−5.51	−7.85	−6.67
SNR = 20 dB								
120	−1.24	−2.30	−2.60	−2.23	1.74	1.88	0.73	0.38
240	−1.07	−1.02	−1.58	−1.48	−1.15	−2.02	−1.38	−2.32
360	−1.17	−0.94	−0.41	−0.65	−4.22	−4.74	−6.96	−5.17
SNR = 30 dB								
120	−5.52	−1.23	−3.51	−5.69	2.46	2.60	1.24	1.16
240	−3.69	−0.47	−2.4	−3.44	−0.77	−1.61	−1.18	−1.47
360	−2.97	−0.23	−0.57	−1.33	−3.94	−3.36	−6.61	−4.88
SNR = 40 dB								
120	−8.21	−1.36	−1.40	−7.80	2.71	2.68	1.87	3.89
240	−6.07	−0.6	−0.64	−4.27	−0.52	−1.14	−1.00	−0.96
360	−4.36	−0.51	−0.41	−2.15	−3.36	−3.03	−5.88	−4.41
SNR = ∞								
120	−12.27	−0.69	−1.99	−7.80	3.52	2.97	2.21	8.11
240	−6.47	−0.20	−1.08	−4.05	−0.17	−1.13	−0.84	0.25
360	−4.48	−0.11	−0.67	−2.75	−3.06	−4.23	−5.04	−2.90

speech enhancement front-end is not used) in terms of NPM and NSegSRR, respectively, for white and colored noise case studies. Averaged values over all speakers are reported in these graphs. Note that in the unprocessed case, there is no IR estimation, and therefore, the NPM value is always equal to 0 dB.

Looking at these experimental results, it can be easily concluded that consistent NPM and NSegSRR improvements are registered in processed audio files due to the use of the proposed algorithmic framework. When the real speaker diarization system is employed, the speech enhancement framework performance decreases: This is mainly due to the occurrence of speaker errors (i.e., the confusion of one speaker identity with another one), which makes the BCI algorithm convergence problematic, thus reducing the dereverberation capabilities of the SDer stage. Nevertheless, still significant improvements are obtained with respect to the results attained in the non-processed case study (Figs. 5, 6). That said, there is space for improvements and some refinements are foreseen in the near future to increase the framework robustness to the speaker diarization errors. Moreover, it must also be underlined that IRs could be estimated during the SDiar training phase (performed using 60s of speech for each speaker), thus accelerating the overall system convergence fulfillment in the real testing phase. However, it must be stressed the fact that the IRs can be estimated continuously even if some changes, such as speaker movements, occur in the room.

With regard to the background noise impact, it is evident that performance degrades as soon as the SNR decreases, but always a significant improvement with respect to the no-processing solution can be registered in all acoustic conditions and for both white and colored noise case studies. It must be underlined that this behavior is guaranteed by the robustness to noise of employed adaptive algorithms within BCI and SDer stages. For instance, in contrast to the RNMCFLMS solution adopted, the UNMCFLMS algorithm diverges even at high SNR values, and therefore, its usage would not lead to the overall results obtained in this work.

It is also worth mentioning that the system performance dependence on SNR is much more evident than the one on T_{60}, as expected. Indeed, the reverberation effect is compensated by the SDer stage, whereas the noise impact is just tolerated: Future efforts will be targeted to face this aspect.

The dependence of NPM and NSegSRR on the speakers active in the meeting has been also evaluated, and related results are reported in Tables 2, 3 and 4. In particular, Table 2 regards the NSegSRR results obtained when the proposed speech enhancement front-end is not used (as mentioned above, the NPM value is always equal to 0 dB in this case). In Tables 3, 4, the NPM and NSegSRR results

for each meeting participant and all addressed acoustic conditions, both in *oracle* and *real* operating modes, are detailed. A certain variability of the evaluation indexes with speaker activity is registered which likely depends on the properties of the recorded speech of the database and on the related speaking duration. However, performance varies coherently over the different T_{60} and SNR conditions for each speaker: This consequently yields the regular curves plotted in Figs. 5, 6. These results are relative to the white noise case study: same conclusions can be drawn in the colored one.

Further simulations have been conducted by considering two other meetings, precisely *IS1008b* and *IS1000b* that have a total duration of 25 m, 5 s and 32 m, 22 s, respectively. The DER percentages for these meetings and the *IS1009b*, subject of previous simulations, are reported in Table 5. The related NSegSRR results averaged over the three reverberation conditions, for the white noise case study, are shown in Table 6: A similar behavior is attainable when additive colored noise is considered in the simulated scenario. Looking at these results, it can be easily concluded that a relevant performance improvement with respect to the unprocessed case study is achievable for the two new meetings, confirming what observed in simulations related to *IS1009b*. Again, the system shows a certain robustness to the noise presence and its performance increase with the SNR value. It must be observed that a certain variability of absolute NSegSRR values are registered over the three meetings, in correspondence with the different behavior of the SDiar sub-system, as reported in Table 6.

Table 5 Diarization error rate (in percentage) values for three AMI meetings for all addressed SNR values and the unprocessed case

Meeting	SNR (dB)					Avg
	10	20	30	40	∞	
IS1009b	31.60	17.85	11.89	9.09	6.71	15.43
IS1008b	11.30	9.13	6.80	4.73	3.93	6.92
IS1000b	22.22	17.81	11.20	8.64	8.41	14.97

Table 6 NSegSRR values for three AMI meetings for all addressed SNR values and the unprocessed case

Meeting	SNR (dB)					Avg	Avg (Unproc.)
	10	20	30	40	∞		
IS1009b	−2.72	−1.94	−1.36	−0.76	−0.03	−1.36	−12.15
IS1008b	−1.57	−1.02	−0.55	−0.05	0.35	−0.56	−10.75
IS1000b	−2.13	−1.50	−0.85	−0.35	0.00	−0.96	−11.71

Conclusions

In this work, an advanced multichannel algorithmic framework to enhance the speech quality in multiparty meetings scenarios has been developed. The overall architecture is able to blindly identify the impulse responses and use them to dereverberate the distorted speech signals available at the microphone. A speaker diarization algorithm is part of the framework and is needed to detect the speakers' activity and provide the related information to steer the blind channel estimation and speech dereverberation operations to optimize the performance. All these algorithmic blocks are able to operate also in the presence of background noise. The algorithms work in real time, and a PC-based implementation of the overall system has been discussed in this contribution. Several computer simulations, based on a subset of the AMI corpus and considering different noisy and reverberated acoustic conditions, have been carried out. Related results have shown a significant improvement in performance with respect to the no-processing option, thus proving the effectiveness of the developed system and its suitability for applications in real-life human–machine interaction scenarios.

As future works, adequate procedures will be integrated in the current framework to reduce the impact of noise presence by means of noise reduction algorithms. Two possible intervention schemes can be foreseen within the speech enhancement front-end on purpose: at the input level, to maximize the performance of the BCI and the speaker diarization stages, and at the output level, to enhance the quality of the signal to be delivered out of the front-end.

Moreover, the speech separation unit will be also integrated in the future, in order to automatically recover the speech content coming from overlapping speakers. Note that the employment of suitable Speech Overlap Detection algorithms [4, 19] within the real-time speaker diarization block is needed on purpose. Some interesting investigations have been made by the authors in the past [35, 37], but never applied to the realistic scenarios addressed. This will likely allow a significant improvement in terms of speech intelligibility [21, 23, 26], which will be adequately evaluated [10].

Last but not least, the application of the proposed framework to automatic speech recognition will be analyzed: Some work has already been done by the authors [37], but more efforts are needed to take the noise presence into account and to suitably integrate the framework with the feature extraction front-end [31, 40]. Other relevant application scenarios to be investigated in the near future are the keyword spotting [43, 44], the activity detection [32], the dominance estimation [20, 33], the emotion understanding and recognition [5, 7, 38].

References

1. Araki S, Hori T, Fujimoto M, Watanabe S, Yoshioka T, Nakatani T, Nakamura A. Online meeting recognizer with multichannel speaker diarization. In: Signals, systems and computers (ASILOMAR), 2010 conference record of the forty fourth asilomar conference on. 2010. p. 1697–701. IEEE
2. Benesty J, Chen J, Huang Y, Cohen I. Noise reduction in speech processing. 1st edn. Springer Publishing Company, Incorporated. 2009.
3. Bettarelli F, Ciavattini E, Lattanzi A, Zallocco D, Squartini S, Piazza F. NU-Tech: implementing DSP algorithms in a plug-in based software platform for real time audio applications. In: Proceedings of 118th convention of the AES; 2005. p. 1–12. Paper number 6389
4. Boakye K, Trueba-Hornero B, Vinyals O, Friedland G. Overlapped speech detection for improved speaker diarization in multiparty meetings. In: Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on; 2008. p. 4353–6. IEEE
5. Bourbakis N, Esposito A, Kavraki D. Extracting and associating meta-features for understanding peoples emotional behaviour: face and speech. Cognit Comput. 2011;3(3):436–48
6. Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, et al. The AMI meeting corpus: a pre-announcement. Machine Learning for Multimodal Interaction; 2006. p. 28–39
7. Chetouani M, Mahdhaoui A, Ringeval F. Time-scale feature extractions for emotional speech characterization. Cognit Comput. 2009;1(2):194–201
8. Cohen I, Benesty J, Gannot S. Speech processing in modern communication: challenges and perspectives. Springer Topics in Signal Processing: Springer; 2010
9. Egger H, Engl H. Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates. Inverse Probl. 2005;21(3):1027–45
10. Falk T, Zheng C, Chan W. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans Audio Speech Lang Processing. 2010;18(7):1766–1774
11. Fredouille C, Bozonnet S, Evans N. The LIA-EURECOM RT'09 speaker diarization system. In: RT'09, NIST rich transcription workshop. Melbourne, Florida; 2009. p. 1–10
12. Guillaume M, Grenier Y, Richard G. Iterative algorithms for multichannel equalization in sound reproduction systems. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing. 2005. vol 3, p. iii/269–iii/272
13. Habets E. Room impulse response (RIR) generator. 2008. http://home.tiscali.nl/ehabets/rirgenerator.html. Accessed 2 Oct 2011.
14. Haque M, Bashar M, Naylor P, Hirose K, Hasan M. Energy constrained frequency-domain normalized lms algorithm for blind channel identification. Signal Image Video Process. 2007;1:203–213
15. Haque M, Hasan M. Noise robust multichannel frequency-domain lms algorithms for blind channel identification. IEEE Signal Process Lett. 2008;15:305–8
16. Hasan M, Benesty J, Naylor P, Ward D. Improving robustness of blind adaptive multichannel identification algorithms using constraints. In: Proceedings of European signal processing conference (EUSIPCO), Antalya, Turkey; 2005. vol 1, p. 11–4
17. Hikichi T, Delcroix M, Miyoshi M. Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations. EURASIP J Adv Signal Process. 2007;1:1–12
18. Huang Y, Benesty J. A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans Speech Audio Process. 2003;51(1):11–24
19. Huijbregts M, van Leeuwen DA, de Jong FMG. Speech overlap detection in a two-pass speaker diarization system. In: INTERSPEECH'09; 2009. p. 1063–6
20. Hung H, Huang Y, Friedland G, Gatica-Perez D. Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans Audio Speech Lang Processing. 2011;19(4):847–60
21. Hussain A, Campbell D. Intelligibility improvements using binaural diverse sub-band processing applied to speech corrupted with automobile noise. In: Vision, image and signal processing, IEE proceedings-; 2001. vol 148, p. 127–32. IET
22. Hussain A, Chetouani M, Squartini S, Bastari A, Piazza F. Nonlinear speech enhancement: an overview. In: Progress in nonlinear speech processing, Lecture notes in computer science; 2007. vol 4391, p. 217–48. doi:10.1007/978-3-540-71505-4_12
23. Kocinski J. Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms. Speech Commun. 2008;50(1):29–37
24. Kokkinis EK, Tsilfidis A, Georganti E, Mourjopoulos J. Joint noise and reverberation suppression for speech applications. In: Proceedings of the 130th convention of the audio engineering society; 2011. vol 9, p. 10–62
25. Loizou P. Speech enhancement: theory and practice (Signal processing and communications). CRC; 2007.
26. Loizou P, Kim G. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans Audio Speech Lang Processing. 2011;19(1):47–56
27. Miyoshi M, Kaneda Y. Inverse filtering of room acoustics. IEEE Trans Signal Process. 1988;36(2):145–52
28. Morgan D, Benesty J, Sondhi M. On the evaluation of estimated impulse responses. IEEE Signal Process Lett. 1998;5(7):174–76
29. Naylor P, Gaubitch N. Speech dereverberation. Signals and communication technology. Heidelberg: Springer; 2010.
30. Oppenheim AV, Schafer RW, Buck JR. Discrete-time signal processing, 2 edn. Upper Saddle River: Prentice Hall; 1999.
31. Principi E, Cifani S, Rotili R, Squartini S, Piazza F. Comparative evaluation of single-channel mmse-based noise reduction schemes for speech recognition. J Electr Comput Eng. 2010; p. 1–7. doi:10.1155/2010/962103. http://www.hindawi.com/journals/jece/2010/962103.html
32. Principi E, Rotili R, Wöllmer M, Eyben F, Squartini S, Schuller B. Real-time activity detection in a multi-talker reverberated environment. Cognit Comput. p. 1–12. doi:10.1007/s12559-012-9133-8
33. Principi E, Rotili R, Wöllmer M, Squartini S, Schuller B. Dominance detection in a reverberated acoustic scenario. In: Advances in neural networks-ISNN2012, Lecture notes in computer science, vol 7368. Springer; 2012.
34. Rotili R, Cifani S, Principi E, Squartini S, Piazza F. A robust iterative inverse filtering approach for speech dereverberation in presence of disturbances. In: Proceedings of IEEE APCCAS; 2008. p. 434–7
35. Rotili R, De Simone C, Perelli A, Cifani A, Squartini S. Joint multichannel blind speech separation and dereverberation: a real-time algorithmic implementation. In: Proceedings of ICIC; 2010. p. 85–93
36. Rotili R, Principi E, Squartini S, Piazza F. Real-time joint blind speech separation and dereverberation in presence of overlapping speakers. In: Proceedings of ISNN. Berlin:Springer; 2011. p. 437–46.
37. Rotili R, Principi E, Squartini S, Schuller B Real-time speech recognition in a multi-talker reverberated acoustic scenario. In: Huang DS, Gan Y, Gupta P, Gromiha M, editors. Advanced intelligent computing theories and applications. With aspects of artificial intelligence, Lecture notes in computer science. Berlin: Springer; 2012. p. 379–86

38. Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. (2011);53(9/10): 1062–87
39. Solé-Casals J, Zaiats V, Monte-Moreno E. Non-linear and non-conventional speech processing: alternative techniques. Cognit Comput. 2010;2(3):133–4
40. Squartini S, Principi E, Rotili R, Piazza F. Environmental robust speech and speaker recognition through multi-channel histogram equalization. Neurocomputing. 2012;78(1):111–120
41. Vinyals O, Friedland G. Towards semantic analysis of conversations: a system for the live identification of speakers in meetings. In: Proceedings of IEEE international conference on semantic computing; 2008. p. 426 –31
42. Weninger F, Schuller B, Batliner A, Steidl S, Seppi D Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. EURASIP J Adv Signal Process. 2011;11:1–16
43. Wöllmer M, Eyben F, Graves A, Schuller B, Rigoll G. Bidirectional lstm networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cognit Comput. 2010;2(3): 180–90
44. Wöllmer M, Marchi E, Squartini S, Schuller B. Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting. Cogn Neurodyn. 2011;5(3):253–64
45. Wooters C, Huijbregts M. The ICSI RT07s speaker diarization system. In: Stiefelhagen R, Bowers R, Fiscus J, editors. Multimodal technologies for perception of humans, Lecture notes in computer science. Berlin: Springer; 2008. p. 509–19
46. Xu G, Liu H, Tong L, Kailath T. A least-squares approach to blind channel identification. IEEE Trans Signal Process. 1995; 43(12):2982–93