Real-time activity detection in a multi-talker reverberated environment

Emanuele Principi, Rudy Rotili, Martin Wöllmer, Florian Eyben, Stefano Squartini, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Principi, Emanuele, Rudy Rotili, Martin Wöllmer, Florian Eyben, Stefano Squartini, and Björn Schuller. 2012. "Real-time activity detection in a multi-talker reverberated environment." *Cognitive Computation* 4 (4): 386–97. https://doi.org/10.1007/s12559-012-9133-8.

Real-Time Activity Detection in a Multi-Talker Reverberated Environment

Emanuele Principi · Rudy Rotili · Martin Wöllmer · Florian Eyben · Stefano Squartini · Björn Schuller

Abstract This paper proposes a real-time person activity detection framework operating in presence of multiple sources in reverberated environments. Such a framework is composed by two main parts: The speech enhancement front-end and the activity detector. The aim of the former is to automatically reduce the distortions introduced by room reverberation in the available distant speech signals and thus to achieve a significant improvement of speech quality for each speaker. The overall front-end is composed by three cooperating blocks, each one fulfilling a specific task: Speaker diarization, room impulse responses identification, and speech dereverberation. In particular, the speaker diarization algorithm is essential to pilot the operations performed in the other two stages in accordance with speakers' activity in the room. The activity estimation algorithm is based on bidirectional Long Short-Term Memory networks which allow for context-sensitive activity classification from audio feature functionals extracted via the real-time speech feature extraction toolkit openSMILE. Extensive computer simulations have been performed by using a subset of the AMI database for activity evaluation in meetings: Obtained results confirm the effectiveness of the approach.

Keywords Speech enhancement · Blind channel identification · Speech dereverberation · Speaker diarization · Real-time signal processing · Activity detection

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

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

S. Squartini
e-mail: s.squartini@univpm.it

M. Wöllmer · F. Eyben · B. Schuller
Institute for Human-Machine Communication, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany
e-mail: woellmer@tum.de

F. Eyben
e-mail: eyben@tum.de

B. Schuller
e-mail: schuller@tum.de

Introduction

Recently, a certain attention has been paid by the scientific community to the design of machines with "cognitive" skills similar to those of humans. Taylor [36] suggests that cognitive machines should be able to "discern and empathize with the mental state of others with which it is in interaction, both machines and humans". In such a context, systems for the automatic analysis of social interaction in small groups have been recently developed [8]. Approaches have been proposed to manage group conversational interaction (e.g., addressing [22] and turn-taking [30]), persons' internal states (e.g., interest [9]), social relations (e.g., roles [46]), and personality (e.g., extroversion [28] and dominance [20]).

Participants' activity level often plays a central role in such analysis. Persons with higher vocal and visual activity (e.g., body movement and gestures correlated with speaking activity) are often perceived as more *dominant* [20]. This has been widely exploited in the dominance detection literature: For example in [2, 19, 20], dominance is estimated calculating the speaking lengths of each participant,

which are directly related to the speaking activities. In particular, the method in [19] uses audio cues only and obtains the speaking lengths through the ICSI speaker diarization system. The works in [2, 20] calculate the speaking lengths using the speech signal energy and augment the feature set with visual cues. Visual activity is measured by the motion vector magnitude and the residual coding bitrate.

Interest and *turn-taking* detection also benefit from the estimation of participants' activity level. For example in [23], the automatic detection of interest segments is performed using HMMs with features comprising both the speech activity measured with SRP-PHAT (steered power response phase transform), and the visual activity measured estimating head and hands motion paths. The same approach is undertaken in [47] for turn-taking detection, where a partially unsupervised framework made of a two-layer HMM discovers group activity patterns.

In this contribution, participant activity detection is addressed using nonverbal vocalic cues in a reverberated acoustic scenario. The meeting participants' voices are recorded by means of multiple distant microphones, suitably located in the room where the meeting takes place. This results in a multi-talker reverberated acoustic scenario which, up to the authors' knowledge, has never been addressed so far. The employed architecture is composed by two main algorithmic stages, both operating in real-time: The speech enhancement front-end and the activity detector (Fig. 1).

The former has been thought to suitably act in the addressed acoustic scenario where multiple speakers are active in a reverberated enclosure. The presence of the reverberation effect due to convolution with room impulse responses (RIRs) strongly degrades the speech quality and a signal processing intervention is required [26]. 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.

In order to perform a suitable dereverberation processing, the main issue to solve consists in coordinating the blind estimation of RIRs with the speech activity of different speakers. In addition, blind multiple-input multiple-output (MIMO) identification is a difficult task even if a short channel impulse responses are considered. To overcome this obstacle, in [18] is proposed to decompose the problem into several subproblems in which single-input multiple-output (SIMO) systems are blindly identified and the estimated RIRs used for source separation and speech deconvolution during double or multiple talk periods. In [32, 33], a real-time implementation of previous mentioned framework and its application as front-end for an automatic speech recognition engine have been proposed by some of the authors.

In this work, a real-time speaker diarization algorithm has been implemented to inform the blind channel identification (BCI) and the dereverberation stages when they have to operate. Only when speech segments of the same speaker occur at the same channel, the BCI stage provides the RIRs estimation to the dereverberation stage in order to accomplish the inversion of such an estimate and apply the inverse filters to the microphone signals.

The second main stage of the proposed architecture is the activity detector. As in [16], the detector operates on low-level features: First, a set of speech feature functionals is extracted from 10 s segments of speech data that has been enhanced by the proposed speech enhancement front-end. Then, the feature vector is used as input for a bidirectional long short-term memory (BLSTM) network that has been trained to map from speech features to four levels of activity. BLSTM networks have advantages over other classification frameworks such as support vector machines (SVM), hidden Markov models (HMM), or conventional (recurrent) neural networks since they efficiently incorporate long-range contextual information into the decoding process. As a result, classifiers based on the long short-term memory technique [10, 15] have shown excellent performance for various human behavior recognition tasks [38, 42].

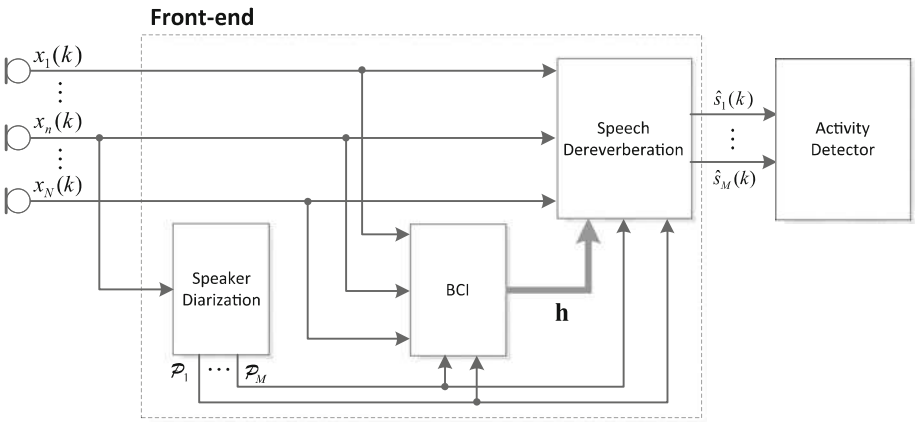


Fig. 1 Block diagram of the proposed framework

In order to evaluate the feasibility and the performance of the proposed approach, experiments have been conducted on a subset of the AMI corpus suitably annotated with activity levels [16]. First, the BCI, dereverberation, and speaker diarization stages have been separately evaluated by means of appropriate quality indexes. Then, the activity detection capabilities of the entire framework have been assessed. Obtained results showed that introducing the proposed front-end gives an average detection improvement of 24.17 %.

The paper outline is the following. In sect. "Speech Enhancement Front-End", the speech enhancement front-end aimed at dereverberating the speech sources is described, whereas section "Activity Detector" details the algorithm developed for activity estimation. Section "Experiments" is targeted to discuss the experimental setup and performed experiments. Conclusions are drawn in section "Conclusion".

Speech Enhancement Front-End

Let M be the number of independent speech sources and N the number of microphones. The relationship between them is described by an $M \times N$ MIMO FIR (Finite Impulse Response) system. According to such a model, the n-th microphone signal at k-th sample time is:

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k, L_h), \quad k = 1, 2, \ldots, K, \quad n = 1, 2, \ldots, N \tag{1}$$

where $(\cdot)^T$ denotes the transpose operator and

$$\mathbf{s}_m(k, L_h) = [s_m(k)\, s_m(k-1) \ldots s_m(k - L_h + 1)]^T. \tag{2}$$

is the m-th source. The term

$$\mathbf{h}_{nm} = [h_{nm,0}\, h_{nm,1} \ldots h_{nm,L_h-1}]^T, \quad n = 1, 2, \ldots, N, \quad m = 1, 2, \ldots, M \tag{3}$$

is the L_h-taps RIR between the n-th microphone and the m-th source. Applying the z transform, Eq. (1) can be rewritten as:

$$X_n(z) = \sum_{m=1}^{M} H_{nm}(z) S_m(z), \qquad n = 1, 2, \ldots, N \tag{4}$$

where

$$H_{nm}(z) = \sum_{l=0}^{L_h-1} h_{nm,l} z^{-l}. \tag{5}$$

The objective is recovering the original clean speech sources s_m by means of a "context-aware" speech dereverberation approach: Indeed, it is necessary to automatically identify who is speaking, accordingly estimating the unknown RIRs and then apply a dereverberation process to restore the original speech quality. To achieve such a goal, the proposed front-end consists of three main stages (Fig. 1): speaker diarization, blind channel identification, and speech dereverberation. The speaker diarization stage takes as input one microphone signal and for each sample, 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 dereverberation stages 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 the speaker diarization algorithm allows to consider the system as composed by the only active source and the N microphones as a SIMO which can be blindly identified in order to perform the dereverberation process.

Speaker Diarization

The speaker diarization stage drives the BCI and dereverberation blocks so that they can operate into speaker-homogeneous regions. Current state-of-the-art speaker diarization systems are based on clustering approaches, usually combining HMMs and the Bayesian Information Criterion metric [7, 43]. 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.

The approach taken here as reference has been proposed in [37], and its block scheme is shown in Fig. 2. The algorithm operation is divided in two phases, training and recognition. In the first, the acquired signals are divided into frames 25 ms long and overlapped by 60 %. These are then transformed in feature vectors composed of 19 Mel-frequency cepstral coefficients (MFCC) plus their first and second derivatives, and cepstral mean normalization is finally applied to deal with stationary channel effects. Speaker models are represented by mixture of Gaussians trained by means of the expectation maximization algorithm. The number of Gaussians and the end accuracy at convergence have been empirically determined and set to 100 and 10^{-4}, respectively.

In the recognition phase, the input signal is divided into non-overlapping chunks, and the same feature extraction pipeline of the training phase extracts feature vectors. The decision is then taken using majority vote on the likelihoods: Every feature vector in the current segment is assigned to one of the known speaker's model based on the maximum likelihood criterion. The model which has the majority of vectors assigned determines the speaker

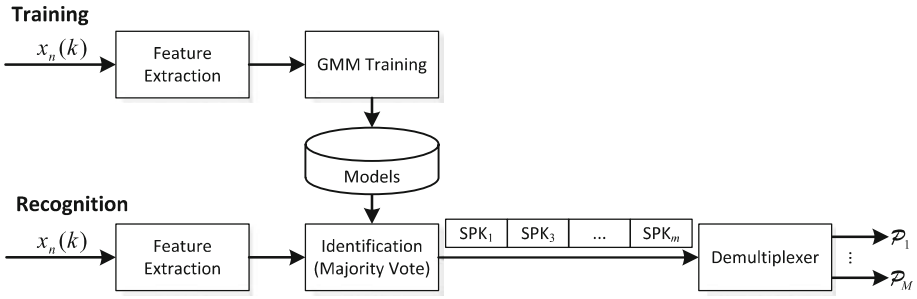


Fig. 2 The speaker diarization block scheme: "SPK$_m$" are the speaker identities labels assigned to each chunk

identity on the current segment. The "Demultiplexer" block associates each speaker label to a distinct output and sets it to "1" if the speaker is the only active, and "0" otherwise.

It is worth pointing out that the speaker diarization algorithm is not able to detect overlapped speech, that is, segments in which more than one speaker talks. Section "Speech Enhancement Front-End Operation" will describe how these errors affect the front-end operation.

Blind Channel Identification

Considering a SIMO system for a specific source s_{m^*}, a BCI algorithm aims to find the RIRs 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 [44]:

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 unconstrained normalized multi-channel frequency-domain least mean square (UNMCFLMS) algorithm [17]. It is an adaptive technique well suited to satisfy the real-time constraints imposed by the case study since it offers a good compromise among fast convergence, adaptivity, and low computational complexity.

Here, a brief review of the UNMCFLMS is reported in order to understand the motivation of its choice in the proposed front-end. Refer to [17] for details. The derivation of UNMCFLMS is based on cross relation criteria [44] using the overlap and save technique [27].

The frequency-domain cost function for the q-th frame is defined as

$$J_f = \sum_{n=1}^{N-1} \sum_{i=i+1}^{N} \underline{\mathbf{e}}_{ni}^H(q)\underline{\mathbf{e}}_{ni}(q) \tag{6}$$

where $\mathbf{e}_{ni}(q)$ 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

$$\begin{aligned}\widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q+1) = \widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q) - \rho[\mathbf{P}_{nm^*}(q) + \delta\mathbf{I}_{2L_h\times L_h}]^{-1} \\ \times \sum_{n=1}^{N} \mathbf{D}_{x_n}^H(q)\underline{\mathbf{e}}_{ni}^{10}(q), \quad i = 1,\ldots,N\end{aligned} \tag{7}$$

where $0 < \rho < 2$ is the step-size, δ is a small positive number and

$$\widehat{\underline{\mathbf{h}}}_{nm^*}^{10}(q) = \mathbf{F}_{2L_h\times 2L_h}[\widehat{\mathbf{h}}_{nm^*}(q)\,\mathbf{0}_{1\times L_h}]^T, \tag{8}$$

$$\underline{\mathbf{e}}_{ni}^{10}(q) = \mathbf{F}_{2L_h\times 2L_h}[\mathbf{0}_{1\times L_h}\,\{\mathbf{F}_{L_h\times L_h}^{-1}\underline{\mathbf{e}}_{ni}(q)\}^T]^T, \tag{9}$$

$$\mathbf{P}_{nm^*}(q) = \sum_{n=1,n\neq i}^{N} \mathbf{D}_{x_n}^H(q)\mathbf{D}_{x_n}(q) \tag{10}$$

The term $\mathbf{F}$ denotes the discrete fourier transform (DFT) matrix. The frequency-domain error function $\underline{\mathbf{e}}_{ni}(q)$ is given by

$$\underline{\mathbf{e}}_{ni}(q) = \mathbf{D}_{x_n}(q)\widehat{\underline{\mathbf{h}}}_{nm^*}(q) - \mathbf{D}_{x_i}(q)\widehat{\underline{\mathbf{h}}}_{im^*}(q) \tag{11}$$

(11) where the diagonal matrix

$$\begin{aligned}\mathbf{D}_{x_n}(q) = \mathrm{diag}(\mathbf{F}\{[x_n(qL_h - L_h)\,x_n(qL_h - L_h \\ + 1)\cdots x_n(qL_h + L_h - 1)]^T\})\end{aligned} \tag{12}$$

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 RIRs, it requires a high input signal-to-noise ratio. In this paper, the presence of noise has not been taken into account, and therefore, the UNMCFLMS still remain an appropriate choice. Different solutions have been proposed in literature in order to alleviate the misconvergence problem of the UNMCFLMS in presence of noise. Among them, the algorithms presented in [12, 13, 45] guarantee a significant robustness against noise and they could be used to improve the proposed front-end.

Speech Dereverberation

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) [24] such that

$$\sum_{n=1}^{N} H_{nm^*}(z)G_{nm^*}(z) = 1, \tag{13}$$

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:

$$C = \|\mathbf{H}_{m^*}\mathbf{g}_{m^*} - \mathbf{v}\|^2, \tag{14}$$

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

$$\mathbf{g}_{m^*} = [g_{1m^*}(1), \cdots, g_{1m^*}(L_i), \cdots, g_{Nm^*}(1), \cdots, g_{Nm^*}(L_i)]^T, \tag{15}$$

$$\mathbf{v} = [\underbrace{0, \cdots, 0}_{d}, 1, \cdots, 0]^T. \tag{16}$$

The vector $\mathbf{v}$ is the target vector, that is, the Kronecker delta shifted by an appropriate modeling delay ($0 \leq d \leq NL_i$) while $\mathbf{H}_{m^*} = [\mathbf{H}_{1m^*}, \cdots, \mathbf{H}_{Nm^*}]$ where $\mathbf{H}_{nm^*}$ is the convolution matrix of the RIR 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

$$\mathbf{g}_{m^*} = \mathbf{H}_{m^*}^{\dagger}\mathbf{v} \tag{17}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudoinverse. By setting the 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. (14) is modified as follows [14]:

$$C = \|\mathbf{H}_{m^*}\mathbf{g}_{m^*} - \mathbf{v}\|^2 + \gamma\|\mathbf{g}_{\mathbf{m}^*}\|^2, \tag{18}$$

where the parameter $\gamma(\geq 0)$, called regularization parameter, is a scalar coefficient representing the weight assigned to the disturbance term. It should be noticed that Eq. (18) has the same form to that of Tikhonov regularization for ill-posed problems [5].

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:

$$C = \mathbf{g}_{m^*}^T\mathcal{H}^T\mathcal{H}\mathbf{g}_{m^*} - \mathbf{g}_{m^*}^T\mathcal{H}^T\mathbf{v} - \mathbf{v}^T\mathcal{H}\mathbf{g}_{m^*} + \mathbf{v}^T\mathbf{v} + \gamma\mathbf{g}_{m^*}^T\mathbf{g}_{m^*} \tag{19}$$

where

$$\mathcal{H} = \begin{cases} \mathbf{H}_{m^*} & \text{(noise case)} \\ \overline{\mathbf{H}}_{m^*} & \text{(fluctuation case).} \end{cases} \tag{20}$$

The filter that minimizes the cost function in Eq. (19) is obtained by taking derivatives with respect to $\mathbf{g}_{m^*}$ and setting them equal to zero. The required solution is

$$\mathbf{g}_{m^*} = (\mathcal{H}^T\mathcal{H} + \gamma\mathbf{I})^{-1}\mathcal{H}^T\mathbf{v}. \tag{21}$$

The usage of Eq. (21) to calculate the inverse filters requires a matrix inversion that, in the case of long RIRs, can result in a high computational burden. Instead, an adaptive algorithm [31] has been here adopted to satisfy the real-time constraints. It is based on the well-known steepest-descent technique, whose recursive estimator has the form

$$\mathbf{g}_{m^*}(q+1) = \mathbf{g}_{m^*}(q) - \frac{\mu(q)}{2}\nabla C. \tag{22}$$

Moving from Eq. (19) through simple algebraic calculations, the following expression is obtained:

$$\nabla C = -2[\mathcal{H}^T(\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)) - \gamma\mathbf{g}_{m^*}(q)]. \tag{23}$$

Substituting Eq. (23) into Eq. (22) is

$$\mathbf{g}_{m^*}(q+1) = \mathbf{g}_{m^*}(q) + \mu(q)[\mathcal{H}^T(\mathbf{v} - \mathcal{H}\mathbf{g}_{m^*}(q)) - \gamma\mathbf{g}_{m^*}(q)]. \tag{24}$$

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 $\mathcal{H}^T\mathcal{H}$ eigenvalues 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 [11], 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:

$$\mu(q) = \frac{\mathbf{e}^T(q)\mathbf{e}(q)}{\mathbf{e}^T(q)(\mathcal{H}^T\mathcal{H} + \gamma I)\mathbf{e}(q)} \tag{25}$$

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 estimation errors due to the BCI algorithm [14]; the real-time constraint can be met also in the case of long RIRs since no matrix inversion is required. Finally, the complexity of the algorithm has been decreased computing the required operation in the frequency-domain by using FFTs.

Speech Enhancement Front-End Operation

The proposed front-end operates in two distinct modalities: Training and testing. In the training phase, each speaker is asked to talk for 60 s. During this period, the speaker diarization stage trains the speakers' models, whereas the BCI and the dereverberation stages perform, respectively, the estimation of the RIRs and the computation of the inverse filters. This strategy avoids a mismatch between training and testing conditions, since speakers' models are always trained under the same acoustic condition of the testing phase.

In the testing phase, the input signal is divided into non-overlapping chunks of 2 s, the speaker diarization stage provides as output the speakers' activity, while in the BCI and dereverberation stages, no adaptation is performed. However, the dereverberation stage still uses the information coming from the speaker diarizer by applying the inverse filter calculated in the training phase if for the m-th speaker the corresponding $\mathcal{P}_m$ is "1". If this is not the case, it sets the output sample to zero. This choice has been made since it is assumed that the speakers do not change their position throughout the meeting. The front-end could handle the situation where the speakers change their position if the adaptation process in the BCI and dereverberation stages were not blocked. However, preliminary simulations showed that the front-end is not able to work properly due to the speaker diarization errors. In particular, the BCI stage is sensitive to false alarms (speaker in hypothesis but not in reference) and speaker errors (mapped reference is not the same as hypothesis speaker). If one of these occurs, the BCI performs the adaptation of the RIRs using an inappropriate input frame providing as output an incorrect estimation. An additional error which produces the previously highlighted behavior is the miss speaker overlap detection.

The sensitivity to false alarms and speaker errors could be reduced imposing a constraint in the estimation procedure and updating the RIR only when a decrease in the cost function occurs. A solution to miss overlap error would be to add an overlap detector and not to perform the estimation if multiple speakers are simultaneously active. On the other hand, missed speaker errors (speaker in reference but not in hypothesis) do not negatively affect the RIRs estimation procedure, since the BCI stage does not perform the adaptation in such frames. Only a reduced convergence rate can be noticed in this case.

Activity Detector

Speech Feature Extraction

For speech feature extraction, the online audio analysis toolkit openSMILE [6] is employed. The audio feature set consists of 1 941 features, composed of 25 energy and spectral related low-level descriptors (LLD) × 42 functionals, 6 voicing related LLD × 32 functionals, 25 delta coefficients of the energy/spectral LLD × 23 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features. Details for the LLD and functionals are given in Tables 1 and 2, respectively. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. The functional set has been based on similar sets, such as the one used for the Interspeech 2011 Speaker State Challenge [35], but has been carefully reduced to avoid LLD/functional combinations that produce values which are constant, contain very little information and/or a high amount of noise.

Long Short-Term Memory

Building on recent studies in the field of context-sensitive affective computing and human behavior analysis [38, 41, 42], an activity classification framework that is based on bidirectional Long Short-Term Memory has been designed. The basic concept of Long Short-Term Memory (LSTM) networks was introduced in [15] and can be seen as an extension of conventional recurrent neural networks that enables the modeling of long-range temporal context for improved sequence labeling. LSTM networks are able to store information in linear memory cells over a longer period of time and can learn the optimal amount of contextual information relevant for the classification task. An LSTM hidden layer is composed of multiple recurrently connected subnets (so-called *memory blocks*). Every memory block consists of self-connected *memory cells* and

Table 1 31 low-level descriptors

Energy and Spectral (25)
Loudness (auditory model based)
Zero crossing rate
Energy in bands 250–650 Hz, 1–4 kHz
25, 50, 75, and 90 % Spectral roll-off points
Spectral flux, entropy
Spectral variance, skewness, kurtosis
Psychoacoustic sharpness, harmonicity
MFCC 1-10
Voicing related (6)
F_0 (Sub-harmonic summation (SHS) followed by Viterbi smoothing)
Probability of voicing
Jitter, shimmer (local)
Jitter (delta: "jitter of jitter")
Logarithmic harmonics-to-noise ratio (logHNR)

Table 2 Set of all 42 functionals

Statistical functionals (23)
(Positive[b]) arithmetic mean, root quadratic mean
Standard deviation, flatness
Skewness, kurtosis
Quartiles, and inter-quartile ranges
1, 99 % percentile
Percentile range 1–99 %
Percentage of frames contour is above: min + 25, 50, and 90 % of the range
Percentage of frames contour is rising
Max, mean, min segment length[c]
Standard deviation of segment length[c]
Regression functionals[a] *(4)*
Linear regression slope, and corresponding approximation error (linear)
Quadratic regression coefficient *a*, and approximation error (linear)
Local minima/maxima related functionals[a] *(9)*
Mean and standard deviation of rising and falling slopes (minimum to maximum)
Mean and standard deviation of inter maxima distances
Amplitude mean of maxima
Amplitude mean of minima
Amplitude range of maxima
Other[a,c] (6)
LP gain, LPC 1-5

[a] Not applied to delta coefficient contours

[b] For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied

[c] Not applied to voicing related LLD

three multiplicative *gate* units (input, output, and forget gates). Since these gates allow for write, read, and reset operations within a memory block, an LSTM block can be interpreted as (differentiable) memory chip in a digital computer (see Fig. 3). Further details on the LSTM principle can be found in [10].

Experiments

Corpus Description

Experiments have been conducted on a subset of the AMI corpus [3, 16]. The subset has a total duration of 180 min and is separated into 36 parts each having a duration of 5 min. Parts are extracted from the scenario-based data recorded with headset microphones at the IDIAP meeting room. Activity annotations are performed every 10 s ranking each participant from 0 to 5, with 0 representing

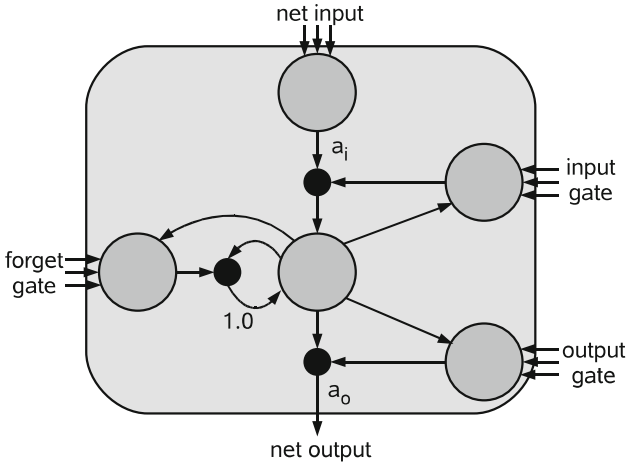


Fig. 3 LSTM memory block consisting of one memory cell: The input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as *small circles*); input, output, and forget gate scale input, output, and internal state, respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state

the lowest activity level and 5 the highest [16]. Table 3 lists the labels with their respective description: As pointed out in the table, here participant activity is annotated considering the amount of speaking time, body movements, gestures and verbal cues. Note, finally, that unlike the experiments described in [16], here the focus is on four levels of activity, since activity level 0 does not occur and level 5 almost never occurs. Thus, the activity level 0 is not considered and levels 4 and 5 were simply clustered together.

The acoustic scenario under study is composed of an array of microphones placed at the center of the meeting table with four speakers sitting around it (Fig. 4). The number of microphones is five, since it must be greater than the number of speakers [17] and the inter-microphone distance is 10 cm. This choice represents a good comprise between impulse response diversification, which increases with the inter-microphone distance, and the need for a reasonably sized array. It is worth highlighting that the UNMCFLMS and MINT algorithms do not suffer from the spatial aliasing problem as delay and sum beamformer [21]. Microphone signals have been created by manually removing cross-talk from the headset sources and convolving them with impulse responses 1,024 taps long. RIRs have been generated by means of the image method [1] using Habets' RIR Generator tool[1], and they represent three different reverberation conditions (T_{60}): 120, 240 and 360 ms. Cross-talk free headset sources will be denoted as "Clean" in the following sections.

[1] http://home.tiscali.nl/ehabets/rir_generator.html.

Table 3 Participant activity labels with their respective descriptions

Label	Description
0—Absent	The participant does not move and does not speak
1—Not active	The participant does not speak and movement or gestures are not associated with any information for activity (e.g., scratching, changing the position in the seat, using the computer, moving the arms, playing with a pen, moving the notepad)
2—Little active	The person is listening, the vocal activity is low (e.g., he said "yes"). He takes note, stands up, sits down, goes to the presentation screen or white board
3—Active	The participant is talking, makes important gestures (e.g., he shakes the head or nodding), he is pointing. He uses additional devices (e.g., "the new remote-control", a prototype), stands in front of the white board or the presentation screen
4—High active	The participant gives a presentation, and/or is the person who talks more then the others, he writes onto the white board, gives some ideas about features, design or architecture of the remote control
5—Most active	The participant makes a decision, defines what the group is doing next, defines what the product should look like

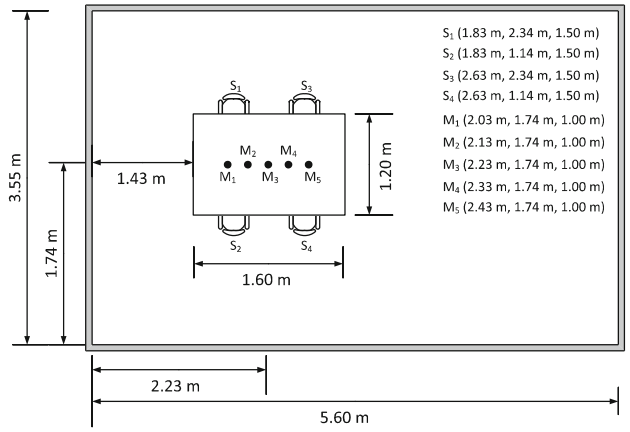


Fig. 4 Room setup

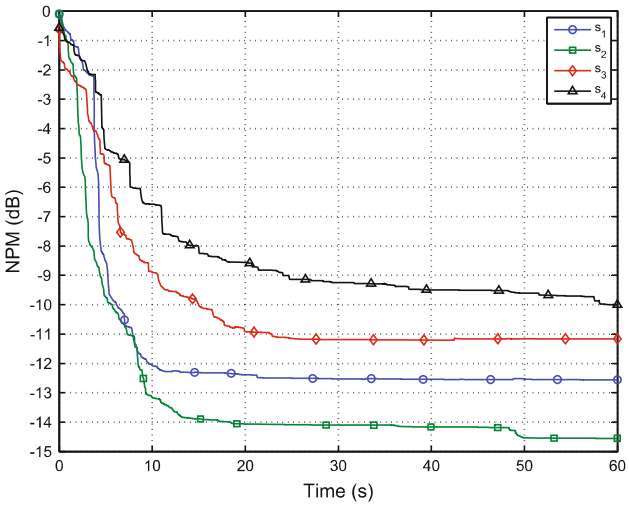


Fig. 5 NPM for all the RIRs relative to each source

Speech Enhancement Front-End Evaluation

As stated in section "Speech Enhancement Front-End", the proposed speech enhancement front-end consists in three different stages. In order to evaluate the performance of each stage, three different quality indexes have been used.

For the BCI stage, a channel-based measure called normalized projection misalignment (NPM) [25] is employed:

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

where

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

is the projection misalignment vector, $\mathbf{h}$ is the real RIR vector, whereas $\widehat{\mathbf{h}}(q)$ is the estimated one at the q-th iteration, that is, the frame index. Figure 5 shows the NPM curve for the identification of each SIMO system relative to each source at $T_{60} = 120$ ms. The curves are obtained considering the signal used in the front-end training phase. It is worth pointing out that different convergence values are achieved since four different SIMO systems are identified but also since the speech quality is different among all the sources.

The performances of the dereverberation stage have been assessed using Normalized Segmental Signal-to-Reverberation Ratio (NSegSRR) that is a signal-based measure defined as follows [26]:

$$\mathrm{NSegSRR} = 10\log_{10}\left(\frac{\|\mathbf{s}_m\|_2}{\|(1/\alpha)\hat{\mathbf{s}}_m - \mathbf{s}_m\|_2}\right), \quad m = 1,\ldots,M \tag{28}$$

where, $\mathbf{s}_m$ and $\hat{\mathbf{s}}_m$ are the desired direct-path signal and recovered speech signals, respectively, and α is a scalar assumed stationary over the duration of the measurement. In calculating the NSegSRR value, the involved signals are assumed to be time-aligned. In Table 4 are reported the NSegSRR values for processed audio files of meeting IS1009b, for each source and all different reverberation time. In order to provide a comparison, the NSegSRR for

Table 4 NSegSRR values for processed audio files of meeting IS1009b

NSegSRR (dB)				
T_{60} (ms)	s_1	s_2	s_3	s_4
120	6.89	12.16	9.54	3.78
240	5.40	8.36	6.84	2.30
360	4.77	8.26	5.74	2.12

Table 5 NSegSRR values for non-processed audio files of meeting IS1009b

NSegSRR (dB)				
T_{60} (ms)	s_1	s_2	s_3	s_4
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

non-processed audio files has been evaluated as well. The obtained values are shown in Table 5.

The increase in terms of NSegSRR confirms the effectiveness of the dereverberation process. However, it is important to remark that the performance of the dereverberation stage are strictly related to the quality of the RIRs estimation obtained through the BCI block.

The performance of speaker diarization algorithms is measured by the diarization error rate[2] (DER). DER is defined by the following expression:

$$\mathrm{DER} = \frac{\sum_{s=1}^{S} \mathrm{dur}(s)(\max(N_{\mathrm{ref}}(s), N_{\mathrm{hyp}}(s)) - N_{\mathrm{correct}}(s))}{\sum_{s=1}^{S} \mathrm{dur}(s) N_{\mathrm{ref}}(s)} \tag{29}$$

where S is the total number of segments in which no speaker change occurs, dur(s) is the duration of segment s, $N_{\mathrm{ref}}(s)$ and $N_{\mathrm{hyp}}(s)$ indicate, respectively, 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 an example, consider one segment ($S = 1$) where for the first half talks SPK_1, and for the second SPK_2 (thus $N_{\mathrm{ref}}(s) = 2$). If the diarization output is SPK_1 for the first half and SPK_3 for the second, then $N_{\mathrm{hyp}}(s) = 2$, but $N_{\mathrm{correct}}(s) = 1$. The diarization error rate is therefore 50 %. 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.

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

Table 6 Speaker diarization performance on the "Clean" and reverberated acoustic scenarios

	Clean	$T_{60} = 120$ ms	$T_{60} = 240$ ms	$T_{60} = 360$ ms
DER (%)	13.02	12.26	11.89	12.03

Table 6 shows the results obtained testing the speaker diarization algorithm on the reverberation-free signals, as well as on the three reverberated scenarios. The performance across the four scenarios are similar due to the matching of the training and testing conditions, and are consistent with [37].

The real-time capabilities of the proposed front-end have been evaluated calculating the real-time factor on a Intel® Core™ i7 machine running at 3 GHz with 4 GB of RAM. The obtained value for the speaker diarization stage is 0.03, meaning that a new result is output every 2.06 s. The real-time factor for the RIRs estimation and dereverberation procedure is 0.04 resulting in a total value of 0.07 for the entire front-end.

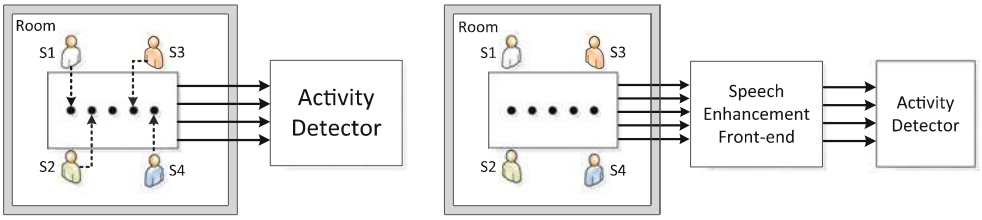
Activity Detector Training and Evaluation Procedure

The networks used for the experiments consist of 1,941 input nodes (one for each speech feature extracted from 10 s of speech), 128 memory blocks containing one memory cell each, and four output nodes that represent the likelihoods of the four activity classes.

The 36 meetings contained in the database were split into three distinct sets: A training, a validation, and a test set. As a 9-fold cross-validation is used, the test set consisted of four meetings for every fold. Four further meetings were used as validation set for each fold, so that the training set was composed of 28 meetings.

All features were mean and variance normalized prior to processing via BLSTM networks. For each fold, means and variances were calculated from the training set only. During training, a learning rate of 10^{-5} and a momentum of 0.9 are used. Zero mean Gaussian noise with standard deviation 0.6 was added to the inputs in the training phase in order to improve generalization. Prior to training, all weights were randomly initialised in the range from −0.1 to 0.1. Input and output gates used tanh activation functions, while the forget gates had logistic activation functions. Training was aborted as soon as no improvement on the validation set could be observed for 25 epochs. Finally, the network that achieved the best classification performance on the validation set is used.

The real-time factor of the activity detector stage is 0.1, which combined with the speech enhancement front-end value confirms the real-time capabilities of the overall framework.



(a) *No front-end* configuration. **(b)** *Full-system* configuration.

Fig. 6 System configurations under study. In **a**, the central microphone is not used and the dashed arrows denote a logical link between speakers and microphones. **a** *No front-end* configuration. **b** *Full-system* configuration

Full-System Evaluation

Activity detection performance has been evaluated in terms of F_1-Measure. Note that due to the unbalanced class distribution, accuracy is a rather inappropriate performance measure. Thus, the F_1-Measure, calculated as the harmonic mean between unweighted recall and unweighted precision, has been used for performance comparison. F_1-Measure is defined as:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \tag{30}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n} \tag{31}$$

$$F_1\text{-Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{32}$$

The terms T_p and F_p and F_n are, respectively, the number of true positives, false positives and false negatives.

The system evaluation has been conducted considering three configurations:

- *No front-end*: The speech enhancement front-end is not present and the activity recognizer operates on four microphone signals (Fig. 6a). Each microphone is logically associated to a single speaker, meaning that the activity detector expects each signal to contain only one voice. The purpose of this experiment is to highlight the need for a front-end able to divide and dereverberate the inputs.
- *Full-system (oracle)*: In this configuration, the full-system as depicted in Figs. 1 and 6b is present, but the speaker diarization stage operates as an oracle. Here, the purpose is to assess the performance of the system without speaker diarization errors.
- *Full-system*: Here, the system is configured as shown in Fig. 6b with the speaker diarization stage operating as described in section "Speaker Diarization".

The baseline F_1-Measure obtained without speech enhancement front-end on "Clean" data is 64.29 %.

Table 7 Activity detection results in terms of F_1-Measure (%). The baseline value obtained on "Clean" data without speech enhancement front-end is 64.29 %

Configuration	$T_{60} = 120$ ms	$T_{60} = 240$ ms	$T_{60} = 360$ ms
No front-end	27.45	26.89	26.07
Full-system (oracle)	55.71	55.61	55.66
Full-system	50.76	50.95	51.19

Table 7 shows the results obtained in the three reverberated conditions.

The performance of the "No front-end" configuration is close to chance level (i.e., 25 %): This is due to the reverberation effect and the inability to isolate the speakers' voices. In fact, the four microphone signals on input of the activity detector contain the mixed voices of all speakers. The introduction of the speech enhancement front-end with oracle speaker diarization improves the system performance, with values close to 56 % across the three reverberated conditions and an average absolute improvement of 28.86 % over the "No front-end" configuration. The introduction of the real speaker diarization stage decrease the performance by 4.69 % on average, while maintaining an absolute increment of 23.31 % with $T_{60} = 120$ ms, 24.06 % with $T_{60} = 240$ ms and 25.12 % with $T_{60} = 360$ ms over the "No front-end" configuration. Note also that the full-system F_1-Measures are less sensitive to the T_{60} variations compared to the "No front-end" configuration. This is consistent with the NSegSRR values shown in section "Speech Enhancement Front-End Evaluation" in which the same behavior can be observed in the non-processed (Table 4) and processed (Table 5) results.

Conclusion

In this paper, an advanced multi-channel algorithmic framework to detect the participant activity levels in multi-talker acoustic reverberated scenarios has been developed. The overall architecture is composed by two main

elements: The speech enhancement front-end and the activity detector. The front-end is able to blindly identify the impulse responses and use them to dereverberate the distorted speech signals acquired by multiple distant microphones. A speaker diarization algorithm is also 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 stages. The developed activity detection algorithm is based on the speech feature extraction toolkit openSMILE. To exploit contextual information, a bidirectional Long Short-Term Memory network which produces the final estimate of the activity level for each speaker is employed.

The entire system is able to work in real-time, and the performed experiments, based on a subset of the AMI corpus, have shown the effectiveness of the developed system, making it appealing for applications in real-life human-computer interaction (HCI) scenarios.

As future works, distinct improvements are foreseen for both the activity estimator and the speech enhancement front-end. Starting from the former, the fusion with video features will be primarily addressed. Moreover, the evaluation of the so-called bottleneck network architectures for enhanced BLSTM modeling of a participant's activity in meetings is planned. A deeper integration of the speaker diarization stage and activity detector algorithm is also foreseen, for example, augmenting the activity detector feature set with the speaking lengths of each participant. With regard to the speech enhancement front-end, the presence of additive noise will be considered and suitable procedures will be taken into account to reduce its impact and maximize the output audio quality. Moreover, the speaker diarization stage will be featured with an overlap-detector algorithm, which also allows to include a source separation stage within the front-end and therefore use also the overlapped speech segments as useful information for meeting activity estimation. Finally, the proposed front-end will be applied in other relevant HCI tasks, for example, keyword spotting [29 39, 40] and emotion recognition [4, 34].

References

1. Allen J, Berkley D. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am. 1979; 65(4):943–50.
2. Aran O, Gatica-Perez D. Fusing audio-visual nonverbal cues to detect dominant people in group conversations. In: Proceedings of the international conference on pattern recognition. 2010. pp. 3687–90.
3. 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. pp. 28–39.
4. Chetouani M, Mahdhaoui A, Ringeval F. Time-scale feature extractions for emotional speech characterization. Cogn Comput. 2009; 1:194–201.
5. Egger H, Engl H. Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates. Inverse Prob. 2005;21(3):1027–45.
6. Eyben F, Wöllmer M, Schuller B. openSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the ACM Multimedia. Firenze, Italy; 2010. pp. 1459–62.
7. Fredouille C, Bozonnet S, Evans N. The LIA-EURECOM RT'09 speaker diarization system. In: RT'09, NIST rich transcription workshop. Melbourne, Florida, USA; 2009.
8. Gatica-Perez D. Automatic nonverbal analysis of social interaction in small groups: A review. Image Vis Comput. 2009; 27(12):1775–87.
9. Gatica-Perez D, McCowan I, Zhang D, Bengio S. Detecting group interest-level in meetings. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing. Philadelphia; 2005. pp. 489–92.
10. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005; 18(5–6):602–10.
11. 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, vol. 3. 2005. pp. iii/269–72.
12. Haque M, Bashar MS, Naylor P, Hirose K, Hasan MK. Energy constrained frequency-domain normalized LMS algorithm for blind channel identification. Signal Image Video Process. 2007; 1(3):203–13.
13. Haque M, Hasan MK. Noise robust multichannel frequency-domain LMS algorithms for blind channel identification. IEEE Signal Process Lett. 2008; 15:305–8.
14. 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;2007(1): 1–12.
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997; 9(8):1735–80.
16. Hörnler B, Rigoll G. Multi-modal activity and dominance detection in smart meeting rooms. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing. 2009. pp. 1777–80.
17. 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.
18. Huang Y, Benesty J, Chen J. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. IEEE Trans Speech Audio Process. 2005;13(5):882–95.
19. Hung H, Huang Y, Friedland G, Gatica-Perez D. Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans Audio Speech Lang Process 2011;19(4):847–60.
20. Jayagopi D, Hung H, Yeo C, Gatica-Perez D. Modeling dominance in group conversations using nonverbal activity cues. IEEE Trans Audio Speech Lang Process 2009;17(3):501–13.
21. Johnson DH, Dudgeon DE. Array signal processing. Englewood Cliffs, NJ: Prentice-Hall; 1993.
22. Jovanovic N. To whom it may concern: addressing in face-to-face meetings. Ph.D thesis, Department of Computer Science, University of Twente 2007.
23. McCowan L, Gatica-Perez D, Bengio S, Lathoud G, Barnard M, Zhang D. Automatic analysis of multimodal group actions in meetings. IEEE Trans Pattern Anal Mach Intell. 2005; 27(3): 305–17.
24. Miyoshi M, Kaneda Y. Inverse filtering of room acoustics. IEEE Trans Signal Process 1988;36(2):145–52.

25. Morgan D, Benesty J, Sondhi M. On the evaluation of estimated impulse responses. IEEE Signal Process Lett. 1998;5(7):174–76.
26. Naylor P, Gaubitch N. Speech dereverberation. Signals and communication technology. New York: Springer; 2010.
27. Oppenheim AV, Schafer RW, Buck JR. Discrete-time signal processing, 2 edn. Upper Saddle River, NJ: Prentice Hall; 1999.
28. Pianesi F, Mana N, Cappelletti A, Lepri B, Zancanaro M. Multimodal recognition of personality traits in social interactions. In: Proceedings of the international conference on multimodal interfaces. Chania, Greece; 2008. pp. 53–60.
29. Principi E, Cifani S, Rocchi C, Squartini S, Piazza F. Keyword spotting based system for conversation fostering in tabletop scenarios: preliminary evaluation. In: Proceedings of 2nd international conference on human system interaction, pp. 216–9. Catania 2009.
30. Reiter S, Schuller B, Rigoll G. Segmentation and recognition of meeting events using a two-layered HMM and a combined MLP-HMM approach. In: Proceedings of IEEE international conference on multimedia and expo, pp. 953–6. Toronto 2006.
31. 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 Asia Pacific conference on circuits and systems, pp. 434–7.
32. 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 6th international conference on intelligent computing, 2010; pp. 85–93.
33. 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, vol. 6839. Berlin, Heidelberg: Springer; 2012. pp. 379–86.
34. 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 Communication 2011. pp. 1062–87.
35. Schuller B, Steidl S, Batliner A, Schiel F, Krajewski J. The interspeech 2011 speaker state challenge. In: Proceedings of interspeech 2011. Florence, Italy 2011.
36. Taylor J. Cognitive computation. Cogn Comput 2009;1:4–16.
37. 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. pp. 426 –431.
38. Wöllmer M, Blaschke C, Schindl T, Schuller B, Färber B, Mayer S, Trefflich B. On-line driver distraction detection using long short-term memory. IEEE Trans Intell Trans Syst. 2011;12(2):574–82.
39. Wöllmer M, Eyben F, Graves A, Schuller B, Rigoll G. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cogn Comput. 2010;2:180–90.
40. Wöllmer M, Marchi E, Squartini S, Schuller B. Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting. Cogn Neurodynamics. 2011;5(3):253–64.
41. Wöllmer M, Metallinou A, Eyben F, Schuller B, Narayanan S. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Proceedings of interspeech. Makuhari, Japan; 2010. pp. 2362–5.
42. Wöllmer M, Schuller B, Eyben F, Rigoll G. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE J Sel Topics Signal Process. 2010;4(5):867–81.
43. 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, Heidelberg: Springer; 2008. pp. 509–19.
44. 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.
45. Yu Z, Er M. A robust adaptive blind multichannel identification algorithm for acoustic applications. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, vol. 2. 2004. pp. 25–8.
46. Zancanaro M, Lepri B, Pianesi F. Automatic detection of group functional roles in face to face interactions. In: Proceedings of the international conference on multimodal interfaces. Banff, Canada; 2006. pp. 28–34.
47. Zhang D, Gatica-Perez D, Bengio S, McCowan I, Lathoud G. Multimodal group action clustering in meetings. In: Proceedings of the ACM 2nd international workshop on video surveillance and sensor networks. New York, NY, USA; 2004. pp. 54–62.