A real-time speech enhancement framework for multi-party meetings

Rudy Rotili, Emanuele Principi, Stefano Squartini, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Rotili, Rudy, Emanuele Principi, Stefano Squartini, and Björn Schuller. 2011. "A real-time speech enhancement framework for multi-party meetings." *Lecture Notes in Computer Science* 7015: 80–87. https://doi.org/10.1007/978-3-642-25020-0_11.

A Real-Time Speech Enhancement Framework for Multi-party Meetings

Rudy Rotili[1], Emanuele Principi[1], Stefano Squartini[1], and Björn Schuller[2]

[1] A3LAB, Department of Biomedics, Electronics and Telecommunications,
Università Politecnica delle Marche, Via Brecce Bianche 1, 60131 Ancona, Italy
{r.rotili,e.principi,s.squartini}@univpm.it
http://www.a3lab.dibet.univpm.it
[2] Institute for Human-Machine Communication
Technische Universität München
Arcisstr. 21, 80333, Munich, Germany
Schuller@tum.de

Abstract. This paper proposes a real-time speech enhancement framework working in presence of multiple sources in reverberated environments. The aim 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 framework is composed by three cooperating blocks, each one fulfilling a specific task: speaker diarization, room-impulse response 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. Extensive computer simulations have been performed by using a subset of the AMI database: Obtained results show the effectiveness of the approach.

Keywords: Speech Enhancement, Blind Channel Identification, Speech Dereverberation, Speaker Diarization, Real-time Signal Processing.

1 Introduction

Multi-party meetings surely represent an interesting real-life acoustic scenario where speech-based Human-Machine interfaces, which have been gaining an increasing scientific and commercial interest worldwide, find application. In this kind of scenario, multiple speakers are active (sometimes also simultaneously) in a reverberated enclosure. The presence of overlapping speech sources and of the reverberation effect due to convolution with room Impulse Responses (IRs) strongly degrades the speech quality and a strong signal processing intervention is required on purpose. 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 [1]. However, up to the authors' knowledge, very few contributions are targeted to face the problem in multi-party meetings, also taking the real-time constraints into account. The main issue to solve consists in coordinating the blind estimation of room IRs with the speech activity of different speakers. In this work a real-time speaker diarization algorithm has been implemented for this purpose. Its aim is first 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. Also, at this level the information provided by the speaker diarizer allows the adaptive filter in the dereverberation algorithm to work only when speech segments of the same speaker occur at the same channel.

It must be observed that some of the authors [2,3] 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 [4,5], 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 these lacks and represents an additional step in the automatization process of the overall speech enhancement framework in real meeting scenarios.

In order to evaluate the achievable performances, several simulations 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. Nevertheless, there is space for improvements and some refinements are foreseen in the near future to increase the framework robustness to the speaker diarization errors.

The paper outline is the following. In Section 2 the overall speech enhancement framework, aimed at dereverberating the speech sources is described. Section 3 is targeted to discuss the experimental setup and performed computer simulations. Conclusions are drawn in Section 4.

2 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:

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

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

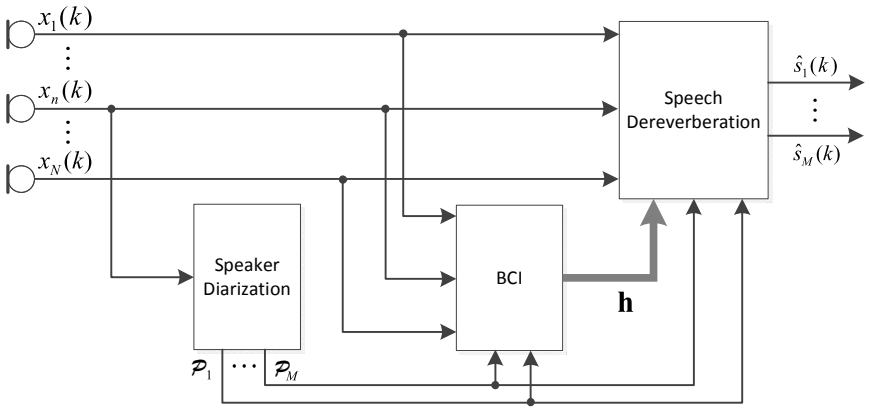

Fig. 1. Block diagram of the proposed framework

with $(m = 1, 2, ..., M,\ n = 1, 2, ..., N)$. The objective is recovering 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 estimating 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) speech dereverberation (SDer). As aforementioned, something close has been proposed by part of the authors of this contribution in the recent past [2,3], but with two noteworthy differences:

- A real speaker diarization algorithm has never been included into the speech enhancement framework operating in multi-party meetings: Indeed in [3], 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 $\mathcal{P}_i$ is 1 if the i-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.
- Here the separation stage has not been comprised: Indeed this stage fulfils its task when overlapping segments occur and these segments need to be automatically detected by means of a specific procedure within the SDiar block. Future works will thus be targeted to develop an overlap-detector algorithm in order to integrate the separation stage into the algorithmic architecture.

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 real-time scenario adaptive filtering techniques are the most suitable. In particular the so-called Unconstrained Normalized Multi-Channel Frequency-domain Least Mean Square algorithm (UNMCFLMS) [7] represents an appropriate choice in terms of estimation quality and computational cost. Though allowing the estimation of long IRs, the UNMCFLMS requires a high input signal-to-noise ratio. Here the noise free case have been assumed and future developments will consider some refinement to make the algorithm work also in presence of significant noise power.

Speech Dereverberation Stage. Given the SIMO system corresponding to source s_m, let us consider the polynomials $G_{s_m,n}(z), n = 1, 2, \ldots, N$ as the dereverberation filters to be applied to the SIMO outputs to provide the final estimation of the clean speech source s_m, according to the following:

$$\hat{S}_m(z) = \sum_{n=1}^{N} G_{s_m,n}(z) X_n(z). \tag{2}$$

The dereverberation filters can be obtained using the well known Bezout's Theorem. However, such a technique requires a matrix inversion that, in the case of long IRs, can be a heavy operation in terms of computational cost. Instead, here an adaptive solution, as presented in [8], is efficiently adopted in order to satisfy the real-time constraints.

Speaker Diarization Stage. The algorithm taken here as reference is the one proposed in [9], which consists in segmenting liverecorded audio into speaker-homogeneous regions with the goal of answering the question "who is speaking now?". 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: 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}.

In the actual recognition mode, the system records and processes chunks of audio as follows: At a 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 [9], 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.

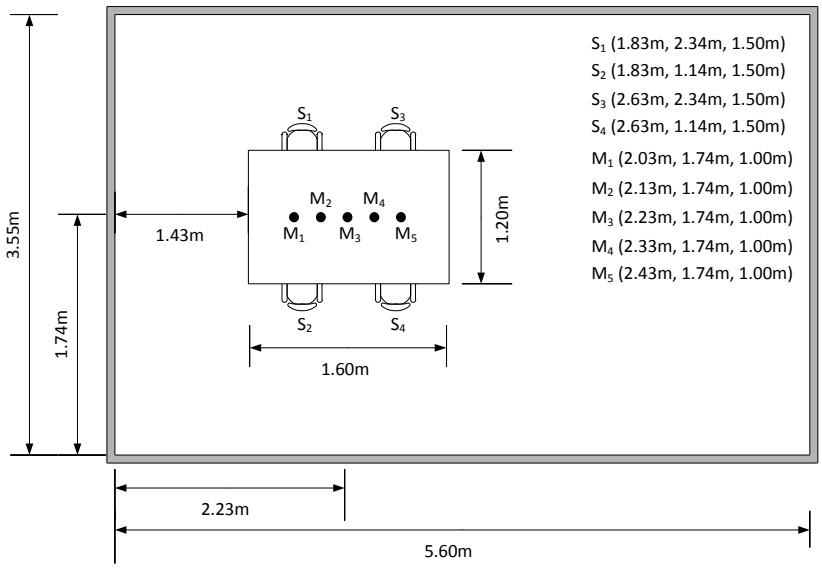

Fig. 2. Room setup

3 Computer Simulations

The overall framework depicted in Fig. 1 has been developed on a freeware software platform, namely NU-Tech [10], suitable for real-time audio processing. 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.

The acoustic scenario under study is made 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. 2. A similar setup is used in the AMI [6] sub-corpus addressed in simulations described later on. Such a corpus contains the 'IS' meetings, well suited for evaluation of algorithms working in multi-party conversational speech scenarios: Indeed they have been used in [9] to test the performances 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 [11], thus synthetically generating the microphone signals. No background noise has been added. Three different reverberation conditions have been taken into account corresponding to $T_{60} = 120, 240, 360$ ms respectively, with IRs 1024 taps long. 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.

Two quality indexes have been used to evaluate the algorithm performances. First the Normalized Segmental Signal-to-Reverberation Ratio (NSegSRR) has been used, which is defined as follows [1]:

$$\text{NSegSRR} = 10 \log_{10} \left(\frac{\|\mathbf{s}_m\|_2}{\|(1/\alpha)\hat{\mathbf{s}}_m - \mathbf{s}_m\|_2} \right), \quad m = 1, \ldots, M \tag{3}$$

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.

Finally, to evaluate the BCI algorithm performances, the Normalized Projection Misalignment (NPM) has been used:

$$\mathrm{NPM}\,(k) = 20\log_{10}\left(\|\epsilon\,(k)\|/\|\mathbf{h}\|\right), \tag{4}$$

where $\epsilon(k) = \mathbf{h} - \frac{\mathbf{h}^T\mathbf{h}_t(k)}{\mathbf{h}_t^T(k)\mathbf{h}_t(k)}\mathbf{h}_t(k)$ is the projection misalignment, $\mathbf{h}$ is the real IR vector whereas $\mathbf{h}_t(k)$ is the estimated one at the k-th iteration (i.e., the frame index). In this case, the lower the NPM value, the better it is.

3.1 Experimental Results

Computer simulations discussed in this section are related to the meeting *IS1009b* of the corpus [6]. It has a total length of $33'15''$ and all the four participants are female speakers. The amount of speaking time for each speaker, including overlap, is $7'47''$, $5'10''$, $7'20''$, $9'00''$ for speaker s_1, s_2, s_3 and s_4 respectively, whereas the total overlap is $3'05''$.

As stated in previous section, three distinct acoustic scenarios have been addressed, corresponding to the aforementioned T_{60} values: For each of them the non-processed and processed cases have been evaluated. Moreover two 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 Section 2).

Experimental results presented in Table 3.1, clearly show that consistent NPM and NSegSRR improvements are registered in processed audio files due to the use of the proposed algorithmic framework. The reported values have been calculated assuming that all algorithms have reached convergence, i.e. considering the last 2 seconds of each speaker. NPM values have to be referred to an initial value of about 0 dB, obtained initializing the overall channel IRs vector to satisfy the unit-norm constraint [7] while NSegSRR values for the non-processed audio files are reported in Table 1.

With regards to Table 3, the SDiar system has shown a Diarization Error Rate (DER) [9] equal to: 6.36% ($T_{60} = 120$ ms), 6.61% ($T_{60} = 240$ ms) and 7.16% ($T_{60} = 360$ ms). The speech enhancement framework performances decrease when the real SDiar system is employed: this is mainly due to the occurrence of speaker errors (i.e. the confusion of one speaker identity with another

Table 1. NSegSRR values for non-processed audio files of meeting *IS1009b*

	NSegSRR (dB)			
$\boldsymbol{T_{60}}$	s_1	s_2	s_3	s_4
120 ms	-4.98	-4.77	-6.78	-4.18
240 ms	-6.11	-6.45	-20.06	-9.59
360 ms	-6.61	-7.56	-27.55	-11.76

Table 2. *'Oracle' Speaker Diarization case study:* NPM and NSegSRR values for dereverberated audio files of meeting *IS1009b*

	NPM (dB)				**NSegSRR** (dB)			
T_{60}	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
120 ms	-13.23	-3.09	-6.16	-9.02	6.65	5.83	5.11	6.67
240 ms	-10.96	-1.70	-6.74	-10.19	7.00	1.29	5.68	6.69
360 ms	-11.52	-1.90	-7.83	-12.69	6.87	1.07	5.25	5.54

Table 3. *'Real' Speaker Diarization case study:* NPM and NSegSRR values for dereverberated audio files of meeting *IS1009b*

	NPM (dB)				**NSegSRR** (dB)			
T_{60}	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
120 ms	-12.27	-0.69	-1.99	-7.80	3.52	2.97	2.21	8.11
240 ms	-6.47	-0.20	-1.08	-4.05	-0.17	-1.13	-0.84	0.25
360 ms	-4.48	-0.11	-0.67	-2.75	-3.06	-4.23	-5.04	-2.90

one) which makes the BCI algorithm convergence problematic, thus reducing the dereverberation capabilities of the SDer procedure. Nevertheless still significant improvements are obtained w.r.t. the results attained in the non-processed case study (see Table 1). Moreover, it must also be underlined that IRs could be estimated during the SDiar training phase (performed using 60 s of speech for each speaker), thus accelerating the overall system convergence fulfilment in the real testing phase. However in this way the authors want to stress the fact that the IRs can be estimated continuously even if some changes, such as speaker movements, occur in the room. Similar results have been obtained with other meeting data and thus they have not been reported for the sake of conciseness.

4 Conclusions

In this paper, an advanced multi-channel algorithmic framework to enhance the speech quality in multi-party 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 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 operations in order to optimize the performances. All the algorithms work in real-time and a PC-based implementation of them has been discussed in this contribution. Performed simulations, 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-machine interaction scenarios. However, as aforementioned, some refinements to make the BCI algorithm more robust to errors in speakers' activity detection are currently under test. As future works, the impact of noise will be considered and suitable procedures will be developed to reduce its impact. Moreover, the application of

the proposed framework in keyword spotting [12], dominance estimation [13], emotion recognition [14] tasks or similar will be analysed.

References

1. Naylor, P., Gaubitch, N.: Speech Dereverberation. Signals and Communication Technology. Springer, Heidelberg (2010)
2. Rotili, R., De Simone, C., Perelli, A., Cifani, S., Squartini, S.: Joint multichannel blind speech separation and dereverberation: A real-time algorithmic implementation. In: Huang, D.-S., McGinnity, M., Heutte, L., Zhang, X.-P. (eds.) ICIC 2010. CCIS, vol. 93, pp. 85–93. Springer, Heidelberg (2010)
3. Rotili, R., Principi, E., Squartini, S., Schuller, B.: Real-time speech recognition in a multi-talker reverberated acoustic scenario. In: Proc. of ICIC, August 11-14 (to appear, 2011)
4. Rotili, R., Principi, E., Squartini, S., Piazza, F.: Real-time joint blind speech separation and dereverberation in presence of overlapping speakers. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part II. LNCS, vol. 6676, pp. 437–446. Springer, Heidelberg (2011)
5. Araki, S., Hori, T., Fujimoto, M., Watanabe, S., Yoshioka, T., Nakatani, T., Nakamura, A.: Online meeting recognizer with multichannel speaker diarization. In: Proc. of Conf. on Signals, Systems and Computers, pp. 1697–1701 (November 2010)
6. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al.: The AMI meeting corpus: A pre-announcement. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 28–39. Springer, Heidelberg (2006)
7. Huang, Y., Benesty, J.: A class of frequency-domain adaptive approaches to blind multichannel identification. IEEE Trans. on Speech and Audio Process. 51(1), 11–24 (2003)
8. Rotili, R., Cifani, S., Principi, E., Squartini, S., Piazza, F.: A robust iterative inverse filtering approach for speech dereverberation in presence of disturbances. In: Proc. of IEEE APCCAS, pp. 434–437 (December 2008)
9. Vinyals, O., Friedland, G.: Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In: Proc. of IEEE International Conference on Semantic Computing, pp. 426–431 (August 2008)
10. Squartini, S., Ciavattini, E., Lattanzi, A., Zallocco, D., Bettarelli, F., Piazza, F.: NU-Tech: implementing DSP algorithms in a plug-in based software platform for real time audio applications. In: Proc. of 118th Conv. of the AES (2005)
11. Habets, E.: Room impulse response (RIR) generator (May 2008), `http://home.tiscali.nl/ehabets/rirgenerator.html`
12. Wöllmer, M., Marchi, E., Squartini, S., Schuller, B.: Robust multi-stream keyword and non-linguistic vocalization detection for computationally intelligent virtual agents. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part II. LNCS, vol. 6676, pp. 496–505. Springer, Heidelberg (2011)
13. Hung, H., Huang, Y., Friedland, G., Gatica-Perez, D.: Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans. on Audio, Speech, and Lang. Process. 19(4), 847–860 (2011)
14. 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, 1062–1087 (February 2011)