DOES DEREVERBERATION HELP MULTICHANNEL SOURCE SEPARATION? A CASE STUDY

Nicolás López[1,2], Mounira Maazaoui[1], Yves Grenier[1], Gaël Richard[1] and Ivan Bourmeyster[2]

[1]Institut Mines-Télécom - Télécom ParisTech - CNRS/LTCI - 37/39 rue Dareau, 75014 Paris, France
[2]Arkamys - 31 rue Pouchet, 75017 Paris, France

ABSTRACT

Multichannel blind source separation performances rapidly degrade when the mixtures are highly reverberated. In fact, blind source separation algorithms usually focus on the separation task without dealing with the dereverberation problem. Some recent studies attempted to reduce the reverberation by introducing a dereverberation module before or after the blind source separation but only limited success was obtained in improving the separation performance in highly reverberant rooms. In this article, we conduct a number of experiments combining state of the art spectral enhancement-based dereverberation and source separation algorithms showing that, in this particular case, speech enhancement does not improve the performance of blind source separation.

Index Terms— Blind source separation, speech dereverberation, spectral subtraction, microphone array.

1. INTRODUCTION

In a multichannel acoustic scene analysis context, one important task is to separate different audio sources that are active simultaneously. This is the case for example in robot audition where the robot equipped with a microphone array must separate the speech signal from several competing talkers, so that it can recognize a given sentence. In this context, blind source separation (BSS) techniques use the multichannel information received at the sensors to recover in separate channels the acoustic events related to a given source. A common approach for BSS is to assume an instantaneous mixture of independent and equally distributed sources. When these conditions are actually met, which is equivalent to consider that sources are propagated in an anechoic environment, methods like independent component analysis give satisfying separation results.

In practice, instantaneous blind source separation techniques are known to fail in reverberant rooms [1], where the mixtures become convolutive. State of the art methods deal with this limitation by working in the frequency domain so that the convolutive mixture can be approximated with an instantaneous one. Methods based on independent component analysis, Non Negative Matrix Factorization and sparse optimization have shown satisfying separation results when the reverberation is low or moderate. However, when the room is highly reverberant the separation performances degrade dramatically. This is because longer Room Impulse Responses (RIR) require longer analysis windows to span all the convolution effects in the frequency domain. But by using longer windows the assumption of independence between the sources does not hold anymore. The separation performances are then bounded by the trade-off between the independence of the sources and the length of the convolutive filter. In a recent work, Maazaoui *et al.* used beamforming methods as a preprocessing step for BSS [2]. By focalizing the directivity of the sensor array towards the sources, the reverberation from the jammer direction is attenuated and as a consequence the separation performances are improved.

Speech dereverberation (SD) techniques have been largely studied in recent years, leading to better reverberation reduction than beamforming techniques [3]. One should then expect to improve the separation performance by previously applying some SD processing to the mixture. In [4], Yoshioka *et. al* proposed to use a multichannel SD algorithm based on linear prediction as a preprocessing step for BSS. The SD and BSS filters were jointly optimized leading to significantly better separation results in rooms with reverberation times of 0.3 and 0.5 seconds. Similar results were observed with the multichannel approach proposed in [5].

In this paper we investigate the influence of single channel spectral enhancement-based dereverberation as a preprocessing step for multichannel BSS for a large range of reverberation times. We considerer a simple framework: single channel SD is applied to every channel and the dereverberated mixtures are separated by multichannel BSS. By using a single channel approach for SD we propose a system that is independent of the geometry of the microphone array. It also allows to parallelize the SD task, allowing for faster processing in real-time applications. We use state-of-the-art methods for this study. SD is performed with the method proposed by Habets *et al.* in [6] and BSS with method by Maazaoui *et al.* in [2]. We show that, in this particular configuration, reducing the reverberation does not lead to an improvement on the separation performances.

This paper is organized as follows: in Section 2 we briefly

introduce the methods used for SD and BSS. In section 3 we present two variants of the sequential algorithm under investigation. Experimental results are given in Section 4 before drawing some conclusions in Section 5.

2. MODELS AND METHODS

2.1. Single channnel dereverberation

For this study we use the single channel dereverberation method described by Habets in [6]. This state-of-the art method is based on a short term prediction of the late reverberant energy and spectral filtering. The signal at the mth microphone is written as $x_m(n) = x_m^e(n) + x_m^r(n)$ where $x_m^e(n)$ is the early signal that we want to recover and $x_m^r(n)$ is the late reverberation signal. $x_m^e(n)$ represents the direct path signal followed by n_e seconds of early reflections. Assuming that the early and late reverberation parts of the RIR are uncorrelated and that the observed signal is stationary it has been shown in [6] that the power spectral density (*psd*) of the reverberant signal at frequency f, frame k and microphone m, denoted $\lambda_m^x(f,k)$, can be written as:

$$\lambda_m^x(f,k) = e^{-2\delta n_e}\lambda_m^x(f,k-k_e) + \lambda_m^e(f,k),$$

where $\lambda_m^e(f,k)$ is the *psd* of the signal affected by the early reflections of the RIR and $\lambda_m^x(f,k) = |X_m(f,k)|^2$. The estimator for the *psd* of late reverberation is given by:

$$\hat{\lambda}_m^r(f,k) = \mathrm{e}^{-2\delta n_e}\lambda_m^x(f,k-k_e). \quad (1)$$

Here, δ is the decay rate of the room and k_e is a frame delay given by $k_e = \frac{n_e f_s}{R}$, where R is the frame rate of the STFT and f_s is the sampling frequency. The decay rate is related to the reverberation time T_{60} by:

$$\delta = \frac{3\ln(10)}{T_{60}}. \quad (2)$$

The framework adopted for single channel speech enhancement based dereverberation is illustrated in Figure 1. The time domain signal $x(n)$ is analyzed with a Short Time Fourier Transform (STFT) filterbank. We first estimate the late reverberant energy using Eq. (1) using the reverberation time measured directly from the RIR. Then we design the time-frequency dereverberation filter $G(f,k)$ using the Optimally Modified Log Spectral Amplitude (OM-LSA) estimator as described in [7]. The OM-LSA framework is an extension of the state-of-the-art Log Spectral Amplitude estimator for speech signals introduced in [8] which takes into account the speech presence uncertainty. This leads to a non linear filter that is a function g of the STFT of the observed signal $X_m(f,k)$ and the late reverberation *psd* $\hat{\lambda}_m^r(f,k)$:

$$G_m(f,k) = g\left(X_m(f,k)\ ,\ \hat{\lambda}_m^r(f,k)\right). \quad (3)$$

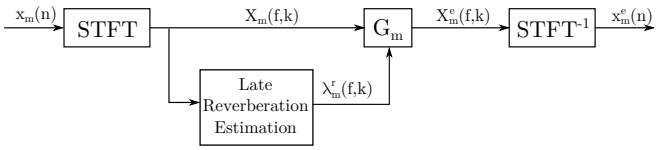


Fig. 1. Single channel dereverberation framework as proposed in [6].

Finally the dereverberated signals are recovered in the matrix $\mathbf{X^e}(f,k)$ where the mth line $X_m^e(f,k)$ is obtained by spectral filtering as:

$$X_m^e(f,k) = G_m(f,k)\, X_m(f,k) \quad (4)$$

2.2. Multichannel blind source separation

We apply a blind source separation algorithm to the M outputs of the dereverberation stage $\mathbf{X^e}(f,k)$ in the time-frequency domain. We use a sparsity separation criterion based on the ℓ_1 norm minimization to estimate the separation matrix $\mathbf{W}(f)$ as originally proposed in [2]. The optimization technique used to update the separation matrix $\mathbf{W}(f)$ is the natural gradient proposed by Amari *et al.* in 1996 [9], the update equation is written as:

$$\mathbf{W}_{t+1}(f) = \mathbf{W}_t(f) - \mu\nabla\psi\left(\mathbf{W}_t(f)\right)\mathbf{W}_t^H(f)\,\mathbf{W}_t(f) \quad (5)$$

$\psi\left(\mathbf{W}(f)\right)$ is our loss function, μ is an adaptation step and t refers to the iteration. In order to induce the sparsity of the separated sources we use the loss function defined by:

$$\psi\left(\mathbf{W}(f)\right) = \sum_{i=1}^{N}\sum_{k=1}^{T}|Y_i(f,k))| \quad (6)$$

where N is the number of sources and T the total number of signal frames. The output signal is $\mathbf{Y}(f,k) = \mathbf{W}(f)\,\mathbf{X^e}(f,k)$ where each line of $\mathbf{Y}$ contains one separated source. For the complete derivation of the natural gradient in Eq. (5) the interested reader should refer to [2].

3. SEQUENTIAL DEREVERBERATION AND BSS ALGORITHMS

In this section we present the two strategies chosen to estimate the multichannel late reverberation starting from the single channel estimator presented in 2.1.

3.1. Single channel reverberation estimation

In a first configuration, each input to the BSS system is dereverberated independently. The late reverberation *psd* of the mth channel is estimated according to Eq. (1) and an individual filter is derived for each channel. This results in M

dereverberated signals. In this configuration, the channels do not share any information about the late reverberation, a channel dependent perturbation signal is used for the design of the M dereverberation filters. We refer to this method as Single Channel Dereverberation (SCD).

3.2. Global reverberation estimation

Our second configuration tries to emphasize the spatial diversity at the microphone level to get a better knowledge of the reverberant field. We assume that the power of late reverberation is approximately equal for each microphone. This allow us to use a global estimate of the late reverberation *psd* defined by:

$$\hat{\lambda}^r(f,t) = \frac{1}{M}\sum_{m=1}^{M}\hat{\lambda}^r_m(f,t) \tag{7}$$

This global estimation of the late reverberation was firstly proposed in [6] where the dereverberation algorithm was used as a postprocessing step to a spatial processor (e.g. beamformer). We estimate $\hat{\lambda}^r_m(f,k)$ for every input channel and align properly the estimated signals before computing $\hat{\lambda}^r$. Finally we design the channel dependent dereverberation filters $G_m(f,t)$ that are now steered by the same perturbation signal $\hat{\lambda}^r(f,t)$. The approach is referred to as Multiple Channel Dereverberation (MCD).

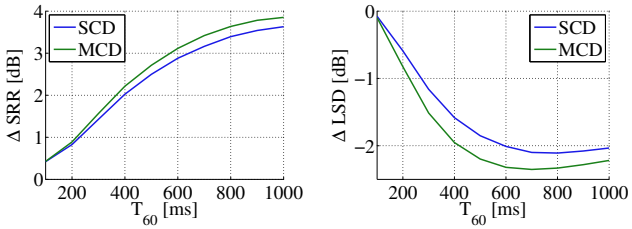
4. EXPERIMENTS AND RESULTS

4.1. Experimental settings

For our experiment we simulate the acoustics of a room with fixed dimensions of $[6 \times 4 \times 5]$m. We consider 42 mixtures with $N = 2$ speech sources at different locations captured by a linear microphone array with $M = 10$ microphones spaced by 5 cm. For each source position and each microphone we compute 10 Room Impulse Responses with reverberation times ranging from 100 to 1000 ms using the Fast Image-Source Method described in [10]. The reverberant mixtures are obtained by filtering each source with the corresponding RIRs. The reverberation time of the room is estimated directly from the RIR using Schroeder's backwards integration method [11]. For each microphone, we estimate the late reverberation *psd* using Eq. (1) and we compute the dereverberation filter using both strategies described in the previous section. The time domain dereverberated signals $\mathbf{x}^e_m(n)$ are fed to the source separation stage with the same settings of [2] and we recover an estimate of the two separated sources denoted $y_1(n)$ and $y_2(n)$.

4.2. Dereverberation Performance

First we evaluate the dereverberation stage using the segmental Signal to Reverberant Ratio (SRR) and the Log Spectral Distortion (LSD) as described in [6]. For each microphone, we use the early signal $x^e_m(n)$ as a reference. We compute each measure for the dereverberated signal and study the improvement related to the unprocessed one for each microphone. Figure 2(a) shows the average improvement of the SRR (ΔSRR) as a function of the reverberation time for the single channel and multiple channel dereverberation methods (SCD and MCD respectively) described in Section 3. In both cases, the reverberation level of the mixture is reduced for every reverberation time. In our application we have a mixture of two sources affected by two different RIRs while the model for dereverberation presented in Section 2.1 assumes a single source affected by a single RIR. However the late reverberation is diffuse so we can consider that it is constant in the room. Under this assumption, we approximate the late reverberation of each individual source by an estimate of $\hat{\lambda}^r_m(f,t)$ for the mixture. If we compare the single channel and the multiple channel approaches we see that the spatial averaging of the estimated reverberant field slightly improves the dereverberation capability of the algorithm, showing that the use of spatial diversity benefits SD.



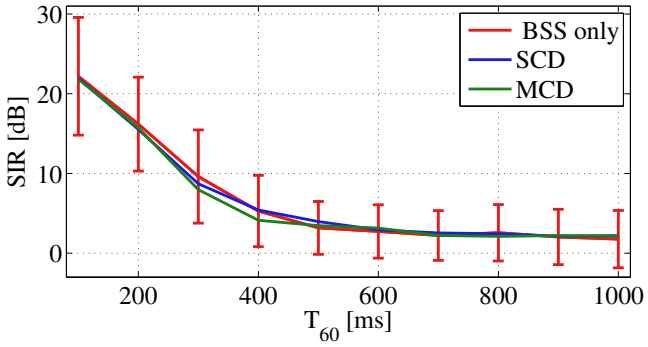
(a) SRR improvement (b) LSD improvement

Fig. 2. SRR and LSD improvement after the dereverberation step using Single Channel Dereverberation (SCD) and Multiple Channel Dereverberation (MCD).

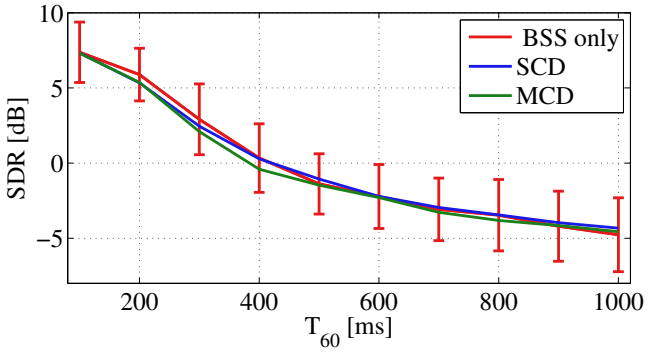
This improvement is obtained at the cost of higher spectral distortion as illustrated in Figure 2(b). The distortion introduced by the dereverberation stage remains moderate even for high reverberation times. In the MCD case, the averaged estimated reverberation uses information about the other channels for the design of the dereverberation filter, this results in spectral distortions due to the incoherences between those channels. The analysis of the dereverberation part shows that at the input of the BSS block we have a signal with less reverberation than the microphone signal but with slightly more spectral distortion. In the following Section we study how this affects the separation performance of the separation step.

4.3. Source Separation Performance

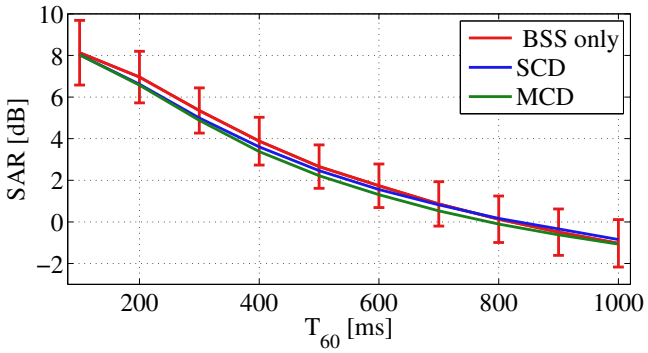
For the evaluation of the separation task we use the BSS-eval Toolbox[12]. We study the Signal to Interference Ratio (SIR), the Signal to Distortion Ratio (SDR) and the Signal to Artifact Ratio (SAR) measures. We compute the separation scores for each separated source in each reverberant condition. The



(a) Average SIR



(b) Average SDR



(c) Average SAR

Fig. 3. Average BSS-eval scores for the separation: without dereverberation (red), using SCD (blue) and using MCD (green). The error bar shows the standard deviation when no dereverberation is performed.

anechoic signals are used as source images. We compare the performance in three configurations: BSS without SD, BSS with SCD and BSS with MCD.

Figure 3 shows the BSS-eval scores as a function of the reverberation time for the three cases. Figure 3(a) gives the SIR separation performance of the algorithms. It is clear that BSS succeeds for reverberation times shorter than 300 ms where the SIR is higher than 10 dB. Then the performance rapidly

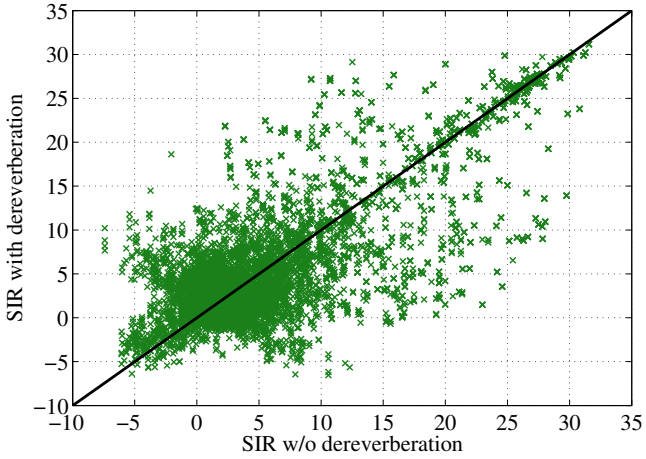


Fig. 4. Scatter plot of the SIR score in the MCD case. Each point represents a separated source in one of the considered reverberant conditions.

degrades until the floor value of 2 dB for highly reverberant enclosures, a drop of 20 dB compared to a reverberation time of 100 ms. There is also a degradation of the SDR and SAR where we lose up to 12 and 9 dB respectively. Regarding the standard deviations of the measured scores, we observed almost identical deviations in the three cases. For the sake of readability, in Figure 3 we only show the error bars in the case where no SD is applied. It is clear that the dereverberation preprocessing does not significantly improve BSS. This means that the BSS stage is not sensitive to the reverberation reduction that we assessed in Section 4.2. The SDR and SAR scores in Figures 3(b) and (c) confirm that, in this particular case, single channel dereverberation does not improve BSS.

We take a further look to the scores with the SIR scatter plot in Figure 4 representing the individual scores of each separated signal for all the reverberant conditions. We compare the scores of BSS without SD and those of the MCD approach. If we focus around 10 dB in the horizontal axis we observe cases where the SIR is improved by 15 dB with the preprocessing. But we also have cases where the separation fails with a degradation of the same amount while in average we do not observe any improvement.

4.4. Discussion

Knowing that reverberation degrades the separation performance as shown in Figure 3(a) we applied a dereverberation algorithm previous to the separation stage. The mixtures processed with SD have reduced reverberation when compared with the unprocessed ones as shown in Figure 2(a). However we could not observe any significant improvement of the separation by using a SD preprocessor. Here we draw some interpretations of this behavior.

BSS and Spectral Subtraction. The OM-LSA estimator involves spectral subtraction that can lead to negative components if the power of the disturbance is over estimated. This is avoided by thresholding the spectral amplitude to a low non-negative value. In addition, spectral subtractive techniques cannot recover the phase component of the target signal. As a consequence, the dereverberated signal is affected by non linear distortions that contradict the linear mixture assumption used for multichannel BSS. For the BSS we also assumed the mixture matrix and the separation matrix $\mathbf{W}(f)$ to be time invariant. However in a spectral enhancement framework the input is processed by different filters from a frame to another. All these elements suggest that spectral filtering techniques are not suitable as a preprocessing for systems relying on linearity assumptions. Dereverberation based on LTI filtering is more suitable for this framework [4].

Early Reflections. The estimator of the *psd* of the reverberation predicts the late reverberation using a parametric model of the room impulse response. This algorithm is robust against RIR fluctuations but it is not intended to perform perfect deconvolution and because of this, the processed signals are still affected by convolutive distortions. As reducing the tail of the reverberation does not affect the separation performances, we should study more deeply the role of the early part of the reverberation in the mixtures.

5. CONCLUSION

We presented in this paper a set of experiments to study the influence of state-of-the-art dereverberation algorithms in a blind source separation context. We used a non linear dereverberation method based on spectral subtraction in conjunction to a linear sparse optimization source separation algorithm. We presented two different strategies to attenuate the reverberation by using the spatial information available in the microphone array. We showed that the reverberation of the mixture signal was well reduced by the chosen technique. Even if the reverberation of the mixtures was efficiently reduced, we did not observe any improvement in the separation performance due to the dereverberation stage. BSS performs poorly for high reverberation times independently of the preprocessing we used. We suggest that spectral subtraction approaches for dereverberation are not suitable for applications involving a system based on linearity assumptions. We believe that further studies have to be carried to understand the effect of early reflections on BSS and to make BSS benefit from the efficiency of spectral subtraction-based approaches for dereverberation.

6. REFERENCES

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 109–116, 2003.

[2] M. Maazaoui, K. Abed-meraim, and Y. Grenier, "Blind source separation for robot audition using fixed HRTF beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 58, 2012.

[3] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer, 2010.

[4] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno, "Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 69–84, 2011.

[5] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1369–1380, 2013.

[6] E.A.P Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[7] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *Signal Processing Letters, IEEE*, vol. 9, no. 4, pp. 113–116, 2002.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, 1985.

[9] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, pp. 757–763, 1996.

[10] E.A. Lehmann and A.M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1429–1439, 2010.

[11] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.