INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Combating Reverberation in NTF-based Speech Separation Using a Sub-Source Weighted Multichannel Wiener Filter and Linear Prediction

Mieszko Fraś, Marcin Witkowski, Konrad Kowalczyk

AGH University of Science and Technology, Institute of Electronics, Krakow, Poland
{fras, witkow, konrad.kowalczyk}@agh.edu.pl

Abstract

Sound source separation (SS) from the microphone signals capturing speech in reverberant conditions is a formidable task. This paper addresses the problem of joint separation and dereverberation of speech using the multichannel Wiener filter (MWF) that is tailored to the sub-source modeling of each speech source with a full-rank mixing matrix. Specifically, the parameters of the proposed sub-source-weighted (SSW) spatial filter are estimated using the sub-source based expectation maximization (EM) algorithm with multiplicative updates (MU) and the localization prior distribution (LP) on the mixing matrix (SSEM-MU-LP). In addition, we strengthen dereverberation by incorporating a Generalized Weighted Prediction Error (GWPE) algorithm. The proposed method is evaluated using a large dataset of two-channel recordings of clean speech convolved with both real and synthesized impulse responses. The results of the experiments show the superior performance of the proposed method in reverberant conditions in comparison to using the standard NTF-based separation with the vanilla MWF in terms of signal-to-distortion ratio (improvement of $3-5.6$ dB) and other commonly used sound separation metrics.

Index Terms: source separation, nonnegative tensor factorization, multichannel Wiener filter, subsource modelling, dereverberation, weighted prediction error

1. Introduction

We consider a task of multichannel blind source separation (BSS) of speakers from convolutive mixtures recorded in a reverberant environment. Many approaches to tackle this issue have been proposed in the literature, including independent component analysis (ICA) [1, 2], independent vector analysis (IVA) [3], nonnegative matrix factorization (NMF) [4], nonnegative tensor factorization (NTF) [5] as well as deep learning methods [6]. In all these methods, the declining efficiency resulting from the presence of high reverberation remains one of the major challenges.

One possible solution to the problem of far-field source separation, adopted also in this work, is to first blindly dereverberate the signals, and next perform source separation. In such a task, the Multichannel Linear Prediction-based (MCLP) dereverberation is particularly attractive since it preserves the phase information, enabling to apply its output signals as input to the BSS algorithm. The state-of-the-art examples of blind MCLP dereverberation are the well-known weighted prediction error (WPE) methods, which incorporate variance weighting for a single [7] and multiple output signals [8]. In [9], a convolutional beamformer (CBF) is used for joint BSS and dereverberation. Other approaches include a combination of multichannel linear filtering-based dereveberation with a determined NMF algorithm [10] or integration of DNN with a weighted power minimization distortionless response (WPD) beamforming [11].

In this work, we focus on NTF-based approaches in which the parameters of the multichannel Wiener filter (MWF) are estimated using an expectation-maximization (EM) algorithm [12, 13]. Their advantage is that blind separation is performed without the need of having training data or any knowledge about the source signals or acoustic conditions. These methods are commonly based on a multichannel convolutive signal model, where the microphone mixture is subject to factorization into a mixing matrix that represents the spatial properties of the signal, and the power spectra of the sources. In [14] it has been shown that using full-rank mixing matrices (the so-called sub-sources) allows for a better representation of reverberant mixtures by modeling not only the spatial position but also the spatial width of each source. A combination of sub-source modeling with generalized EM algorithm and the Itakura-Saito (IS) divergence results in a sub-source based EM algorithm with multiplicative updates (SSEM-MU), which achieves fast convergence and improved separation performance [15, 16, 17].

Recently, a method to incorporate the statistical localization prior (LP), which models reverberation according to the statistical theory of room acoustics, into the SSEM-MU algorithm has been proposed by the current authors [18]. In contrast to the existing methods [19], the model adopted in [18] does not require any knowledge about the sources, the room and its characteristics. Note that the required localization information can be conveniently obtained using, e.g., the GCC method which is robust to reverberation [20]. Due to the independent modeling of the directional and diffuse sounds, the SSEM-MU-LP algorithm is capable of estimating the powers of individual sources in reverberant conditions.

In this paper, we aim to improve the performance of speech separation from reverberant mixtures by proposing a modification to the standard MWF so that it exploits the sub-source model with the localization prior, in which the direct sound is contained predominantly in the first sub-source and late reverberation is concentrated in other sub-sources. The proposed sub-source weighted (SSW) MWF performs joint separation and dereverberation of speech using information estimated from the SSEM-MU-LP algorithm [18]. To further increase the robustness of SSEM-MU-LP towards late reverberation, we additionally incorporate the weighted prediction error (WPE) dereverberation algorithm [8]. The proposed method is presented in Sec. 2, and it is followed by the experimental evaluation, results and concluding discussion provided in Secs. 3 and 4.

2. Sub-Source Based Speech Separation in Reverberant Conditions

2.1. Sub-source signal model

Consider a signal model for an I-channel convolutive mixture of J sources in the short-time Fourier transform (STFT) do-

http://dx.doi.org/10.21437/Interspeech.2021-1230

main, which can be written as

$$\tilde{\mathbf{x}}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_f, \tag{1}$$

where $\tilde{\mathbf{x}}_{fn} = [\tilde{X}_{1fn}, \tilde{X}_{2fn}, \ldots, \tilde{X}_{Ifn}]^{\mathrm{T}} \in \mathbb{C}^I$ is the vector of the microphone signals (after or without an additional dereverberation which is described in Sec. 2.4), by $n = 1, \ldots, N$ and $f = 1, \ldots, F$ denote the time and frequency indices, respectively, $\mathbf{s}_{fn} = [S_{11fn}, \ldots, S_{1Ifn}, S_{21fn}, \ldots, S_{2Ifn}, \ldots, S_{J1fn}, \ldots, S_{JIfn}]^{\mathrm{T}} \in \mathbb{R}_+^{JI}$ is the vector of the so-called sub-sources that share the same spectral variance within each source and are mutually independent between the sources, $\mathbf{A}_f = [\mathbf{A}_{1f}, \mathbf{A}_{2f}, \ldots, \mathbf{A}_{Jf}]^{\mathrm{T}}$ is the matrix consisting of time-invariant mixing matrices $A_{jf} \in \mathbb{C}^{I \times I}$ for each source, and $\mathbf{b}_f = [B_{f1}, B_{f2}, \ldots, B_{fI}]^{\mathrm{T}} \in \mathbb{C}^I$ is the noise vector modeled as $\mathbf{b}_f \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{\Sigma}_{b,f})$, with the noise covariance matrix given by $\mathbf{\Sigma}_{b,f} = \boldsymbol{\sigma}_{b,f}^2 \boldsymbol{I}_{I \times I}$.

2.2. Sub-source weighted multichannel Wiener filter (SSW-MWF) for joint separation and reverberation reduction

In this section, we present the proposed sub-source weighted multichannel Wiener filter, which recovers the spatial image of the direct components of each source, while it reduces the reverberant signal component. In order to model the spectral and spatial cues of the sources, the signal model (1) can be first represented with a local Gaussian model (LGM) as [14]

$$\tilde{\mathbf{x}}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{jfn}, \tag{2}$$

where vector $\mathbf{y}_{jfn} = [Y_{j1fn}, Y_{j2fn}, \ldots, Y_{jIfn}]^{\mathrm{T}} \in \mathbb{C}^I$ contains the spatial images that jointly represent the j-th source as captured by the microphones. The microphone signal vector is modeled using a zero-mean circular complex Gaussian

$$\mathbf{y}_{jfn} \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{R}_{jf} V_{jfn}), \tag{3}$$

where $\mathbf{R}_{jf} \in \mathbb{C}^{I \times I}$ denotes the time-invariant, full-rank spatial covariance matrix and $\mathbf{V} \in \mathbb{R}_+^{J \times F \times N}$ denotes the matrix with non-negative spectral variances $\mathbf{V}_j \in \mathbb{R}_+^{F \times N}$ for the j-th source. Noting that $\mathbf{A}_{jf}$ is full-rank and that the spatial covariance matrix $\mathbf{R}_{jf}$ can be non-uniquely represented as

$$\mathbf{R}_{jf} = \mathbf{A}_{jf} \mathbf{A}_{jf}^H, \tag{4}$$

following the derivations in [18], we can model the mixing matrix for each source using the complex Gaussian distribution

$$\mathbf{A}_{jf} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{U}_{jf}, \mathbf{\Sigma}_{jf}), \tag{5}$$

with the mean and covariance matrices respectively given by $\boldsymbol{U}_{jf} \in \mathbb{C}^{I \times I}$ and $\mathbf{\Sigma}_{jf} \in \mathbb{C}^{I \times I}$. Since we aim to separate the direct-path source signals, we construct the mixing matrix $\mathbf{A}_{jf}$ so that the signal that propagates over the direct path is concentrated only in the mean of the prior for the first sub-source (i.e., for $i = 1$), which can be written as

$$\boldsymbol{U}_{jf} = \left[\mathbf{d}_{jf}, \mathbf{O}_{I \times (I-1)}\right], \tag{6}$$

where $\mathbf{O}_{I \times (I-1)}$ is the zero matrix and $\mathbf{d}_{jf} = \left[1, e^{-\jmath\omega\,\tau_2}, \ldots, e^{-\jmath\omega\,\tau_I}\right]^{\mathrm{T}}$ denotes the steering vector for the direct-path signal, where $\omega = 2\pi f$, $\jmath = \sqrt{-1}$ denotes the angular frequency, and τ_i denotes the delay between the direct signal that propagated from the source to the i-th microphone relative to the delay for the reference (in our case first) microphone. On the other hand, room reverberation is modelled using the so-called spatial coherence matrix $\mathbf{\Omega}_f \in \mathbb{C}^{I \times I}$ [21] whose elements are given by $[\mathbf{\Omega}_f]_{ii'} = \mathrm{sinc}(\frac{\omega ||\mathbf{p}_i - \mathbf{p}_{i'}||_2}{\nu})$, where $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ are the respective microphone positions and ν denotes the wave velocity. Since the diffuse late reverberation cancels out on average, it is taken into account only in the covariance matrix of the prior, which can be expressed as

$$\mathbf{\Sigma}_{jf} = \mathbf{\Omega}_f. \tag{7}$$

Based on the aforementioned definitions, we propose to retrieve the spatial image of the direct-path source signal using the sub-source weighted multichannel Wiener filter (MWF) tailored to preserve primarily the signal of the first sub-source, which can be expressed as

$$\widehat{\mathbf{Y}}_{jfn} = \mathbf{A}_{jf}\, \mathbf{\Lambda}\, \mathbf{A}_{jf}^H V_{jfn} \left[\sum_{j=1}^{J} \mathbf{A}_{jf} \mathbf{A}_{jf}^H V_{jfn}\right]^{-1} \tilde{\mathbf{X}}_{fn}. \tag{8}$$

where $\mathbf{A}_{jf,i}$ is the mixing matrix for the i-th sub-source of the j-th source, and $\mathbf{\Lambda} = \mathrm{diag}([\alpha_1, \alpha_2, \ldots, \alpha_I]^{\mathrm{T}})$ denotes the diagonal matrix with real elements $\alpha_i \in [0, 1]$ on the main diagonal. For instance, for the suppression of nondirect-path signals, we can set $\alpha_1 = 1$ and all the remaining weights as $\alpha_i \ll 1$.

2.3. EM algorithm for NTF-based sub-source separation

Given the aforementioned signal model, the maximum a posteriori (MAP) estimator of the parameters $\mathbf{\Theta} = \{\mathbf{A}, \mathbf{V}\}$ of the probabilistic model, with the statistical localization prior distribution over the mixing matrix, can be formulated as

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\mathrm{argmax}}\ P(\mathbf{\Theta}|\tilde{\boldsymbol{X}}, \boldsymbol{S}) = \underset{\mathbf{\Theta}}{\mathrm{argmax}}\, P(\tilde{\boldsymbol{X}}, \boldsymbol{S}|\mathbf{\Theta}) P(\mathbf{\Theta}), \tag{9}$$

where $\tilde{\boldsymbol{X}}$ is the observed microphone mixture and the latent data consists of the sub-sources $\boldsymbol{S}$. The resulting negative log-likelihood (up to irrelevant constant values) is given by [18]

$$\begin{aligned} \log P(\tilde{\boldsymbol{X}}|\boldsymbol{S}, \mathbf{\Theta}) + \log P(\boldsymbol{S}|\mathbf{\Theta}) + \log P(\boldsymbol{A}) \stackrel{\mathrm{c}}{=} \\ \sum_{f,n} \mathrm{Tr}\Big\{ \mathbf{\Sigma}_{b,fn}^{-1} \big(\mathbf{R}_{\tilde{x}\tilde{x},fn} - \mathbf{A}_f \mathbf{R}_{\tilde{x}s,fn}^H - \mathbf{R}_{\tilde{x}s,fn} \mathbf{A}_f^H \\ + \mathbf{A}_f \mathbf{R}_{ss,fn} \mathbf{A}_f^H \big) \Big\} + \sum_{f,n} \log |\mathbf{\Sigma}_{b,f}| + I \sum_{j,f,n} \mathrm{d}_{IS}\left(\xi_{jfn} | V_{jfn}\right) \\ - \gamma \log \mathcal{N}_{\mathbb{C}}\left(\mathbf{A}_f | \boldsymbol{U}_f, \mathbf{\Sigma}_f\right), \end{aligned} \tag{10}$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace operator, $\mathrm{d}_{IS}(\xi_{jfn}|V_{jfn})$ denotes the Itakura-Saito divergence [16], and the sufficient statistics are given by $\mathbf{R}_{\tilde{x}\tilde{x},fn} = \tilde{\mathbf{X}}_{fn} \tilde{\mathbf{X}}_{fn}^H$, $\mathbf{R}_{\tilde{x}s,fn} = \tilde{\mathbf{X}}_{fn} \mathbf{S}_{fn}^H$, $\mathbf{R}_{ss,fn} = \mathbf{S}_{fn} \mathbf{S}_{fn}^H$ and $\xi_{jfn} = \frac{1}{I} \sum_i |S_{ji,fn}|^2$. In order to estimate the model parameters $\mathbf{\Theta}$, we apply the recently proposed SSEM-MU-LP algorithm [18]. In the E step, the conditional expectation of sufficient statistics is calculated, similarly to [17]. In the M step, the model parameters are updated using standard MU rules and the closed-form solution for the mixing matrix $\mathbf{A}$ is used [18]. For the detailed description of the update equations, the reader is referred to [18].

2.4. WPE-based multichannel dereverberation

As additional means of reducing late reverberation, we propose to preprocess the microphone signals using the well-known multichannel Generalized Weighted Prediction Error (GWPE)

[8] algorithm, which preserves the phase information of the microphone signals and does not impose a condition on the number of sources. Specifically, the GWPE signal model in the STFT domain for the f-th subband reads

$$\mathbf{X}_f = \tilde{\mathbf{X}}_f + \mathbf{X}_{Df}\mathbf{C}_f\,, \tag{11}$$

where $\tilde{\mathbf{X}}_f = [\tilde{\mathbf{x}}_{f1}^{\mathrm{T}}, \tilde{\mathbf{x}}_{f2}^{\mathrm{T}}, \ldots, \tilde{\mathbf{x}}_{fN}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{N\times I}$ denotes the matrix of the estimated dereverberated signal (which corresponds to the linear prediction error) together with the noise signal, $\mathbf{X}_f \in \mathbb{C}^{N\times I}$ denotes the microphone signal matrix that is defined similarly, $\mathbf{C}_f \in \mathbb{C}^{IL_c\times I}$ is a matrix with the prediction filters, each of length L_c for the MIMO system, and $\mathbf{X}_{Df} = [\mathbf{x}_{D,f1}, \mathbf{x}_{D,f2}, \ldots, \mathbf{x}_{D,fN}]^{\mathrm{T}} \in \mathbb{C}^{N\times IL_c}$ is a related convolution matrix built from the microphone signals. Each vector $\mathbf{x}_{D,fn} = [\mathbf{x}_{D1fn}^{\mathrm{T}}, \mathbf{x}_{D2fn}^{\mathrm{T}}, \ldots, \mathbf{x}_{DIfn}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{IL_c}$ contains I stacked vectors $\mathbf{x}_{Difn} = [X_{if(n-D)}, X_{if(n-D-1)}, \ldots, X_{if(n-D-L_c+1)}]^{\mathrm{T}} \in \mathbb{C}^{L_c}$ with the reversed signal parts of length L_c, delayed by a prediction delay D for the respective i-th channel. For each channel, the signal might be separately modelled using a circular complex Gaussian distribution with a zero-mean and time-varying variance. When the microphone array is small, the covariance matrix might be spatially averaged across the estimated signals [8] and the cost function becomes [22]

$$\underset{\mathbf{C}_f, \lambda_{fn}}{\operatorname{argmin}}\ \mathcal{N}_{\mathbb{C}}(\tilde{\mathbf{x}}_{fn}\,|\,\mathbf{0}, \lambda_{fn}\mathbf{I}_{IxI}). \tag{12}$$

Problem (12) can be solved using the maximum likelihood approach by the alternative estimation of both the spatial covariance matrix and the filter coefficients. The update rules read

$$\lambda_{fn}^{(\tau)} = \frac{1}{I(2\delta+1)}\sum_{i=1}^{I}\sum_{\eta=n-\delta}^{n+\delta}|\tilde{X}_{if\eta}^{(\tau-1)}|^2\,, \tag{13a}$$

$$\mathbf{C}_f^{(\tau)} = \left(\mathbf{X}_{Df}^H\mathcal{D}_{\boldsymbol{\lambda}_f^{(\tau)}}^{-1}\mathbf{X}_{Df}\right)^{-1}\mathbf{X}_{Df}^H\mathcal{D}_{\boldsymbol{\lambda}_f^{(\tau)}}^{-1}\mathbf{X}_f\,, \tag{13b}$$

$$\tilde{\mathbf{X}}_{fn}^{(\tau)} = \mathbf{X}_f - \mathbf{X}_{Df}\mathbf{C}_f^{(\tau)}\,, \tag{13c}$$

where $\mathcal{D}_{\boldsymbol{\lambda}_f^{(\tau)}} = \mathrm{diag}(\boldsymbol{\lambda}_f^{(\tau)})$ where $\boldsymbol{\lambda}_f^{(\tau)} = [\lambda_{f1}^{(\tau)}, \ldots, \lambda_{fN}^{(\tau)}]^{\mathrm{T}}$ is a spatially averaged variance vector, δ is a time frame context for the variance smoothing and τ denotes the iteration index. The GWPE algorithm loop, given by (13a)-(13c), finishes when the convergence condition is met or the maximum number of iterations T is reached. Hereafter we will refer to the described dereverberation method as multichannel WPE.

3. Experimental Evaluation and Results

3.1. Description of performed experiments

The evaluation of the proposed modifications to the NTF-based speech separation is performed using two datasets of reverberant microphone mixtures synthesized by convolving speech recordings with room impulse responses (RIRs). In the first dataset, the impulse responses were generated using the image-source method [23]. We simulated a room of size $10 \times 10 \times 4$ m with absorption coefficients selected so that a set of 5 different reverberation times (RT60) ranging from 0.3 s to 0.9 s was obtained. For each RT60 value, 5 random setups of sources and arrays were drawn, in which the position of a two-element microphone array with inter-microphone spacing of 0.05 m was selected randomly around the room center, while the two speech sources were located randomly at a distance of around 2 m from the array. In the second dataset, the room impulse responses were taken from the MIRD database [24], from which we randomly selected the two-channel RIRs from 13 measured source positions at a constant distance of 2 m from the microphone array, which consisted of 2 microphones with a spacing of 0.06 m. We randomly selected 5 pairs of RIRs measured in rooms with RT60 of 0.36 s and 0.61 s, respectively. Each pair of dry, non-reverberant recordings consisted of samples from the TIMIT database [25] with approximately matched lengths of both signals and mismatched utterances spoken by different speakers. The length of each sample was adjusted to the shorter recording of each pair. In order to use all TIMIT speakers, the entire set of 315 clean sample pairs consisted of 192 female-male and 123 male-male pairs. The microphone mixtures were obtained by convolving nonreverberant speech with the simulated or measured RIRs. The signals sampled at 16 kHz were processed with a 2048-point STFT with 50% overlap.

The NTF-related parameters of the SSEM-MU-LP algorithm (the mapping matrix **Q**, the matrix with frequency profiles **W**, and the matrix with time activations **H**) were initialized with random values drawn from the uniform distribution. The random values in matrices **W** and **H** were additionally scaled with the power of the observed mixture averaged over time and frequency, respectively. The mixing matrix **A** was initialized randomly, and the localization prior was based on the relative position information about the speakers. The maximum number of iterations in the EM algorithm was set to 5.

In case of performing additional dereverberantion, we used the *Nara-wpe* implementation [22] of the GWPE as a separate preprocessing block, where both STFT and ISTFT operated on frames of length 512 samples with a shift of 160 samples (which corresponded to 10 ms). As in [26], the prediction delay D was set to 1. The filter length was set empirically using the adjusted formula $L_c = 50 \cdot \mathrm{RT60}$, and the context range and maximum number of iterations were set to $\delta = 2$ and $T = 3$, respectively.

The efficacy of the compared processing was assessed using the standard separation evaluation measures described in [27], namely, the signal-to-distortion ratio (SDR), image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR). As the reference speech signals, we consistently used the direct-path speech signals as captured by the microphones. For each reverberation time, the evaluation measure values were computed as an average over the results obtained for 315 recordings of different speakers in random source array setups. The recently presented SSEM-MU-LP algorithm with standard MWF (equivalent to the proposed MWF with $\alpha_1 = \alpha_2 = 1$), which has been shown in [18] to outperform state-of-the-art NTF-based source separation algorithms such as SSEM [13], SSEM-MU [17], is used across all experiments as a reference algorithm. All separation results presented in this paper show an improvement (denoted with Δ) achieved by the studied method over the aforementioned reference algorithm.

We compare the BSS with the proposed sub-source weighted MWF that extracts the first sub-source only ($\alpha_1 = 1$ and $\alpha_2 = 0$), the BSS with the proposed sub-source weighted MWF which extracts the first and partially the second sub-source (empirically set $\alpha_2 = 0.1$), the BSS with standard MWF operating on the microphone signals dereverberated using the WPE algorithm (denoted as WPE&stdMWF), and the BSS with the first two aforementioned filters which operate on the microphone signals dereverberated using the WPE algorithm [denoted respectively as WPE&MWF(0) and WPE&MWF(0.1)].

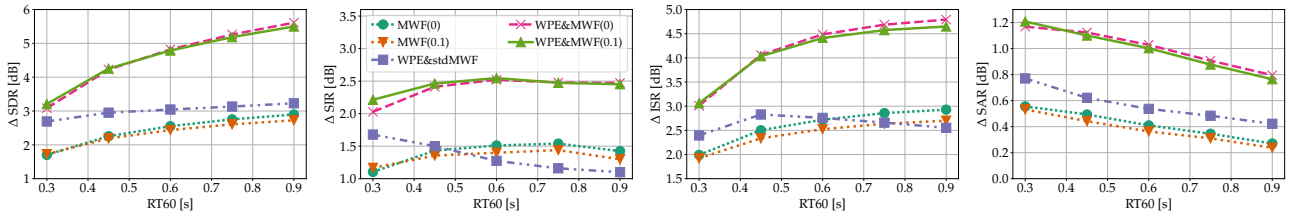

Figure 1: *Improvements (denoted with Δ) in SDR, ISR, SIR and SAR measures obtained by the proposed processing over the vanilla BSS with standard MWF without WPE-based dereverberation for synthesized RIRs with RT60 ranging from 0.3 s to 0.9 s. The BSS with the proposed sub-source weighted MWF for $\alpha_2 = 0$ and $\alpha_2 = 0.1$ [denoted as MWF(0) and MWF(0.1)], the BSS with standard MWF after dereverberation with the WPE algorithm (denoted as WPE&stdMWF), and the full proposed processing, i.e. the BSS with the sub-source weighted MWF after dereverberation with the WPE algorithm [denoted as WPE&MWF(0) and WPE&MWF(0.1)].*

3.2. Results and discussion

The results of the experiments performed on the dataset created using the simulated impulse responses are presented in Fig. 1. As can be clearly observed, all of the presented approaches result in a notable gain in separation performance for all reverberation times and all evaluation measures. In particular, the proposed sub-source weighted multichannel Wiener filter which extracts the first sub-source only and the version which in addition extracts part of the second sub-source [denoted as MWF(0) and MWF(0.1)], perform very similarly. The SDR results are rising with an increasing level of reverberation, reaching nearly 3 dB at RT60 = 0.9 s. This indicates that the reverberant signal component is primarily estimated as the second sub-source by the SSEM-MU-LP algorithm. Thus, the speech extraction of primarily the first sub-source signal using the proposed sub-source weighted MWF is a suitable method of achieving an additional dereverberation of the separated speech signals. The WPE dereverberation of the microphone signals applied before the source separation (denoted as WPE&stdMWF), allows for obtaining a better estimation of the MWF parameters, with an improvement over the lack of dereverberation oscillating around 3 dB in SDR values. In this case, the late reverberation is largely reduced so that the BSS algorithm can operate on much less reverberant data, thereby improving its performance. Although SDR results are higher in this case, the ISR drops quite significantly for high reverberation times, which may indicate that part of the remaining reverberation in the separated signals may still contain the reverberant components of both sources. The drop in SIR for high RT60 values shows that the WPE struggles to achieve full dereverberation in presence of two simultaneously active sources in highly reverberant conditions, and that further reverberation reduction is still desired. The best performance is achieved when both WPE-based dereverberation and sub-source based BSS with the proposed filtering are jointly applied [WPE&MWF(0) and WPE&MWF(0.1)]. The proposed method enables to reduce the late reverberation so that the spectral components of speech are better estimated by the source separation algorithm. The reduction of the reverberant component present in the second sub-source further improves the separation results, reaching up to nearly 5.6 dB of improvement in SDR over the same BSS method which does not incorporate any reverberation reduction. Note that none of the presented variants introduces distortions to the desired extracted direct-path signals. It could be even argued that it can lead to a slight improvement of the quality of the recovered signal in comparison with the standard MWF. Finally, the ISR results indicate that the application of the proposed processing allows for a more precise capturing of the spatial characteristics of the speech sources, which may be advantageous if further spatial signal processing is of interest.

Table 1: *The SDR, SIR, ISR and SAR improvements obtained by the studied methods for genuine RIRs from the MIRD database [24] measured using a 2-element array in a room with RT60s of 0.36 s and 0.61 s.*

Metric	Δ SDR	Δ SIR	Δ ISR	Δ SAR
RT60 [10 ms]	36 | 61	36 | 61	36 | 61	36 | 61
MWF (0)	1.3 | 1.7	0.9 | 0.2	0.0 | 0.9	0.0 | 0.0
MWF (0.1)	1.6 | 1.6	1.3 | 0.2	0.2 | 0.9	0.0 | 0.0
WPE&stdMWF	2.0 | 2.5	0.8 | 1.2	1.1 | 1.1	2.3 | 2.8
WPE&MWF (0)	2.3 | 3.4	2.2 | 1.5	0.3 | 1.4	0.9 | 2.6
WPE&MWF (0.1)	2.6 | 3.4	2.8 | 1.6	0.6 | 1.4	1.1 | 2.7

The results of experiments performed on the dataset created using real, measured RIRs are presented in Table 1. In general, these results confirm the main conclusions drawn from the aforementioned results. In particular, the sole application of WPE is shown to provide better improvement than the sub-source filter alone. This can be explained by the relatively low level of early reflections in the measured RIRs, so that the late reverberation suppression achieved by the WPE already facilitates better separation. Observing the results of the final proposed processing, the usage of the sub-source weighted Wiener filter with $\alpha_2 = 0.1$ seems to be the most robust version, which provides both best separation (SIR results) and output signal quality (SAR results).

4. Conclusions

In this paper, we have presented a method to perform robust separation of speech sources in various reverberant acoustic conditions. The proposed sub-source weighted multichannel Wiener filter, which exploits the NTF-based sub-source spatial model, supplemented by WPE-based dereverberation, is shown to achieve significant improvements over the analogous source separation model that does not reduce room reverberation.

5. Acknowledgements

This research received financial support from the Foundation for Polish Science under grant number First TEAM/2017-3/23 (POIR.04.04.00-00-3FC4/17-00) which is co-financed by the EU and was supported in part by PLGrid Infrastructure.

6. References

[1] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 569270, 2003.

[3] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 165–172.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[5] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[6] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[8] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[9] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation," in *Proc. Interspeech*, 2020.

[10] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 31–35.

[11] T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, and S. Araki, "DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6399–6403.

[12] L. Benaroya, L. M. Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP).*, vol. 6. IEEE, 2003, pp. VI–613.

[13] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1118–1133, 2011.

[14] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[15] A. Ozerov, C. Févotte, R. Blouet, and J. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 257–260.

[16] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[17] A. Ozerov, C. Févotte, and E. Vincent, "An introduction to multichannel NMF for audio source separation," in *Audio Source Separation*. Springer, 2018, pp. 73–94.

[18] M. Fraś and K. Kowalczyk, "Maximum a posteriori estimator for convolutive sound source separation with sub-source based NTF model and the localization probabilistic prior on the mixing matrix," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 526–530.

[19] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 1–11, 2013.

[20] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone arrays*. Springer, 2001, pp. 157–180.

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

[22] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IEEE Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[25] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993. [Online]. Available: https://hdl.handle.net/11272.1/AB2/SWVENO

[26] M. Witkowski and K. Kowalczyk, "Split Bregman approach to linear prediction based dereverberation with enforced speech sparsity," *IEEE Signal Processing Letters*, vol. 28, pp. 942–946, 2021.

[27] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.