INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint Blind Source Separation and Dereverberation for Automatic Speech Recognition using Delayed-Subsource MNMF with Localization Prior

Mieszko Fraś, Marcin Witkowski, Konrad Kowalczyk

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

Abstract

Overlapping speech and high room reverberation deteriorate the accuracy of automatic speech recognition (ASR). This paper proposes a method for jointly optimum source separation and dereverberation using delayed subsource multichannel nonnegative matrix factorization (MNMF). We formulate a subsource-based signal model that accounts for late room reverberation using time-delayed microphone signals from several past time frames. We then propose a maximum a posteriori (MaP) estimator based on MNMF with localization prior on the mixing matrix suitable for direct-path and reverberant signal components estimation. Finally, two algorithms are derived, namely the original and simplified delayed subsource MNMF, which are shown to outperform many state-of-the-art approaches. The results of experimental evaluations, performed using real and simulated data, indicate superior performance of the proposed processing in terms of the word error rate (WER) as well as signal-to-distortion ratio (SDR).

Index Terms: source separation, dereverberation, automatic speech recognition, nonnegative matrix factorization

1. Introduction

Automatic speech recognition (ASR) can achieve tremendous accuracy on low-reverberant recordings of a single speaker taken with a close-talk microphone [1, 2]. On the other hand, speech recorded in real-life conditions is often contaminated by interfering sounds and reverberation. For robust ASR, it is advantageous to reduce the detrimental effects of highly overlapping speech and strong room reverberation, for which dereverberation and source separation can sequentially be performed.

For multichannel dereverberation, a popular approach is the weighted prediction error (WPE) method [3, 4], while for source separation (SS), it is common to formulate a probabilistic generative model such as multichannel nonnegative matrix factorization (MNMF) [5, 6, 7]. The latter proved to be highly effective in convolutive source separation performed blindly in unseen acoustic conditions, for which preparation of training data would be cumbersome. Such a sequential approach, however, is not optimal, due to mutually-dependent relationships between the dereverberation and source separation processes.

Recently, some attempts have been made to join WPE-based multichannel linear prediction with constrained variants of MNMF [8, 9] in which the spatial model is constrained to be first-rank [10] or jointly diagonalizable [11], with the aim to reduce high degrees of freedom of a full-rank spatial covariance matrix (SCM) and lower the method's sensitivity to parameter initialization. One interesting approach to further increase the reliability of MNMF is to incorporate prior information, e.g. about the relative source positions, into the probabilistic framework, which yields the so-called maximum a posteriori (MaP) estimator [12, 13]. To enhance separation of reverberant speech, in [14] WPE was followed by MNMF with a localization prior. However, apart from suboptimal sequential processing, the method would retrieve the entire reverberant source image and could only cover part of reverberation that falls within the short-time Fourier transform (STFT) analysis window.

In this paper, we aim to estimate direct-path speech signals (i.e. separated non-reverberant speech) from multichannel microphone mixtures with high speech overlap and strong reverberation, for improving the effectiveness of back-end ASR. To this end, we propose a novel delayed subsource MNMF (DS-MNMF) method which models source spectra using nonnegative tensor factorization (NTF) [6] and models joint late reverberation components as subsources based on spectral information inferred from the 'time-lagged' microphone signals. We propose a MaP estimator, designed for direct signal estimation, obtained by carefully selecting the mean and covariance of the localization prior on direct-path and late reverberation subsources. Furthermore, we present the derived update equations for the resulting expectation-maximization (EM) algorithm, and propose a less computationally complex variant of the proposed DS-MNMF based on a simplified model. Finally, an extensive experimental evaluation of the proposed algorithms is performed against state-of-the-art, in simulated and real rooms, showing significant improvements in terms of word error rate (WER) and signal-to-distortion ratio (SDR) for various reverberation conditions.

2. Proposed Delayed Subsource MNMF

2.1. Signal model

Let us consider an I-channel microphone mixture of J reverberant sources, which can be represented within a single time-frequency bin of the STFT as $\boldsymbol{x}_{fn} = \sum_{j=1}^{J} \boldsymbol{y}_{fn}^{(j)}$, where $\boldsymbol{x}_{fn} = [X_{1,fn}, X_{2,fn}, \ldots, X_{I,fn}]^{\mathrm{T}} \in \mathbb{C}^I$ is the microphone mixture, $\boldsymbol{y}_{fn}^{(j)} = [Y_{1,fn}^{(j)}, Y_{2,fn}^{(j)}, \ldots, Y_{I,fn}^{(j)}]^{\mathrm{T}} \in \mathbb{C}^I$ represents the spatial image of the j-th source (i.e., source signal as captured by $i = 1, \ldots, I$ microphones), $j = 1, \ldots, J$ is the source index, while $f = 1, \ldots, F$ and $n = 1, \ldots, N$ denote the frequency bin and time frame indices, respectively.

The complex spectrum of the j-th source $S_{fn}^{(j)} \in \mathbb{C}$ can be modelled as $S_{fn}^{(j)} \sim \mathcal{N}_c(0, V_{fn}^{(j)})$, where $V_{fn}^{(j)} \in \mathbb{R}_+$ denotes the non-negative spectral variance, which can be structured with the joint NTF model [15], i.e. as

$$V_{fn}^{(j)} = \sum_{k=1}^{K} Q_k^{(j)} W_{fk} H_{kn}, \tag{1}$$

10.21437/Interspeech.2023-2520

where $Q_k^{(j)} = [\boldsymbol{Q}]_k^{(j)}$, $W_{fk} = [\boldsymbol{W}]_{fk}$, $H_{kn} = [\boldsymbol{Q}]_{kn}$ are the elements of the respective NTF matrices, and K denotes the number of NTF components. The columns of matrix $\boldsymbol{W}$ represent the frequency profiles, the rows of matrix $\boldsymbol{H}$ represent time activations, while matrix $\boldsymbol{Q}$ maps component k to the j-th source.

Assuming that the duration of an early part of the room impulse response (RIR) is shorter than the length of the time frame of the STFT [16], in this paper we propose to model the reverberant microphone mixture as a sum of early signal components $\boldsymbol{d}_{fn}^{(j)}$ for J sound sources and a joint late reverberation component $\boldsymbol{r}_{fn}$ for all sources, which contains a sum of delayed late reverberation components from the past time frames. The proposed microphone mixture model is thus given by

$$\boldsymbol{x}_{fn} = \sum_{j=1}^{J} \underbrace{\boldsymbol{a}_f^{(j)} S_{fn}^{(j)}}_{\boldsymbol{d}_{fn}^{(j)}} + \underbrace{\sum_{\tau=\delta}^{L_a} \boldsymbol{c}_{f\tau} X_{\mathrm{ref},f,n-\tau}}_{\boldsymbol{r}_{fn}} + \boldsymbol{b}_f, \qquad (2)$$

where $\boldsymbol{a}_f^{(j)} = [A_{1,f}^{(j)}, A_{2,f}^{(j)}, \ldots, A_{I,f}^{(j)}]^{\mathrm{T}} \in \mathbb{C}^I$ is the time-invariant transfer function between the j-th source and I microphones, $X_{\mathrm{ref},f,n-\tau}$ denotes the mixture signal at the reference microphone, $\boldsymbol{c}_{f\tau} \in \mathbb{C}^I$ is a time-invariant convolutional transfer function for $\tau = \delta, \delta+1, \ldots, L_a$, δ is the delay between the early and late components, whilst L_a is the length of the convolutional transfer functions. The noise vector $\boldsymbol{b}_f = [B_{1,f}, B_{2,f}, \ldots, B_{I,f}]^{\mathrm{T}} \in \mathbb{C}^I$, which follows complex Gaussian distribution $\boldsymbol{b}_f \sim \mathcal{N}_c(0, \boldsymbol{\Sigma}_{b,f})$, is used in this work for the so-called simulated annealing (for details about the annealing procedure please view [5]).

In order to more conveniently formulate (2), we treat J early source components and $L = L_a - \delta$ delayed late reverberation components as $M = J + L$ subsources, and stack them together in a single subsource vector $\boldsymbol{s}_{fn} = [S_{fn}^{(1)}, \ldots, S_{fn}^{(J)}, (X_{\mathrm{ref},f,n-\delta}), \ldots, (X_{\mathrm{ref},f,n-L_a})]^T \in \mathbb{C}^M$. We can then rewrite (2) in vector notation as

$$\boldsymbol{x}_{fn} = \boldsymbol{A}_f \boldsymbol{s}_{fn} + \boldsymbol{b}_f \qquad (3)$$

where $\boldsymbol{A}_f = [\boldsymbol{a}_f^{(1)}, \ldots, \boldsymbol{a}_f^{(J)}, \boldsymbol{c}_f^{(J+1)}, \ldots, \boldsymbol{c}_f^{(J+L)}] \in \mathbb{C}^{I \times M}$, in which for convenience we use the following notation $\boldsymbol{c}_f^{(J+1)} = \boldsymbol{c}_{f\delta}$ and $\boldsymbol{c}_f^{(J+L)} = \boldsymbol{c}_{fL_a}$. In sections to follow, we will refer to the transfer function for the m-th subsource as $[\boldsymbol{A}_f]^{(m)}$, i.e. $[\boldsymbol{A}_f]^{(m)} = \boldsymbol{a}_f^{(m)}$ for $m \leq J$ and $[\boldsymbol{A}_f]^{(m)} = \boldsymbol{c}_f^{(m)}$ for $m > J$.

2.2. Maximum a posteriori estimator

In order to estimate the parameters of the probabilistic model $\boldsymbol{\Theta} = \{\boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Sigma}_b\}$ with a prior distribution over the mixing matrix $\boldsymbol{A}$, we formulate the following posterior in which $\boldsymbol{X}$ is the observed microphone mixture and the latent data consisting of subsources $\boldsymbol{S}$. The log-posterior of complete data $\{\boldsymbol{X}, \boldsymbol{S}\}$ is given by

$$\log P(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{S}) = \log P(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{\Theta}) + \log P(\boldsymbol{S}|\boldsymbol{\Theta}) + \log P(\boldsymbol{A}). \qquad (4)$$

In order to design a suitable prior over the mixing matrix, we assume that transfer function for m-th subsource can be modeled using a complex Gaussian distribution

$$[\mathbf{A}_f]^{(m)} \sim N_c(\boldsymbol{u}_f^{(m)}, \boldsymbol{\Sigma}_f^{(m)}), \qquad (5)$$

with mean vector $\boldsymbol{u}_f^{(m)} \in \mathbb{C}^I$ and covariance matrix $\boldsymbol{\Sigma}_f^{(m)} \in \mathbb{C}^{I \times I}$. In contrast to state-of-the-art source separation which adopt the location priors to retrieve reverberant source signals [12, 13], in this work, our goal is to restore the non-reverberant source signals. Early source signal components propagate over the direct path and several early reflection paths between the source and the microphones, and hence they are primarily concentrated in the mean of the prior. On the other hand, late reverberation is diffuse in nature and hence it cancels out on average, whilst it should be taken into account in the covariance of the 'lagged' subsources, which can be expressed by

$$\boldsymbol{\Sigma}_f^{(m)} = \begin{cases} \frac{1}{\gamma} \boldsymbol{I}_{I \times I} & \text{, for } m \leq J, \\ \boldsymbol{\Omega}_f & \text{, for } m > J, \end{cases} \qquad (6)$$

where $\gamma \in \mathbb{R}_+$ is a hyperparameter that controls the strength of the prior (in this work always set to 1), $\boldsymbol{I}_{I \times I}$ denotes an identity matrix, and $\boldsymbol{\Omega}_f \in \mathbb{C}^{I \times I}$ denotes the spatial coherence matrix [17] whose elements are given by $[\boldsymbol{\Omega}_f]_{ii'} = \mathrm{sinc}(\kappa ||\mathbf{p}_i - \mathbf{p}_{i'}||_2)$, where $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ are the respective microphone positions and κ denotes the wave number. The mean of the prior is defined as

$$\boldsymbol{u}_f^{(m)} = \begin{cases} \boldsymbol{\mu}_f^{(j)} & \text{, for } m \leq J, \\ \mathbf{0}_I & \text{, for } m > J, \end{cases} \qquad (7)$$

where $\mathbf{O}_I$ is the vector of zeros and $\boldsymbol{\mu}_f^{(j)} \in \mathbb{C}^I$ is the vector with relative transfer functions of the early parts of the RIRs between the j-th source and the microphones. In practice, it is often assumed that the direct propagation path contributes the most, in which case the steering vector is set based on information from the localization algorithm, e.g. [18] is used in this work.

Having proposed a suitable prior over the mixing matrix, we introduce the conditional expectation operator $\mathbb{E}_{X|S,\Theta^l}[\cdot]$ to the negative log-posterior of (4), which yields an auxiliary cost function $Q_p(\Theta, \Theta^l)$ to be minimized, which reads

$$\begin{aligned} Q_p(\Theta, \Theta^l)_{fn} = \sum_{f,n} \mathrm{Tr}\Big\{ \boldsymbol{\Sigma}_{b,fn}^{-1} \Big(\widehat{\mathbf{R}}_{xx,fn} - \mathbf{A}_f \widehat{\mathbf{R}}_{xs,fn}^H - \\ - \widehat{\mathbf{R}}_{xs,fn} \mathbf{A}_f^H + \mathbf{A}_f \widehat{\mathbf{R}}_{ss,fn} \mathbf{A}_f^H \Big) \Big\} + I \sum_{j,f,n} \mathrm{d}_{IS}\left(\widehat{\xi}_{fn}^{(j)} | V_{fn}^{(j)} \right) + \\ + \sum_{f,n} \log |\boldsymbol{\Sigma}_{b,f}| - \sum_{m,f} \log N_c\left([\mathbf{A}_f]^{(m)} | \boldsymbol{u}_f^{(m)}, \boldsymbol{\Sigma}_f^{(m)} \right), \end{aligned} \qquad (8)$$

where $\mathrm{Tr}\{\cdot\}$ denotes the trace operator, $\mathrm{d}_{IS}(\xi_{fn}^{(j)} | V_{fn}^{(j)})$ denotes the Itakura-Saito divergence [19], and the expectations of sufficient statistics are given by $\widehat{\mathbf{R}}_{xx,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H]$, $\widehat{\mathbf{R}}_{xs,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{x}_{fn}\mathbf{s}_{fn}^H]$, $\widehat{\mathbf{R}}_{ss,fn} = \mathbb{E}_{X|S,\Theta^l}[\mathbf{s}_{fn}\mathbf{s}_{fn}^H]$, and $\widehat{\xi}_{fn}^{(j)} = \mathbb{E}_{X|S,\Theta^l}[|S_{fn}^{(j)}|^2]$.

2.3. The proposed DS-MNMF algorithm

The proposed optimization criterion (8) is minimized using an expectation-maximization algorithm that iterates between the expectation (E) step, in which the conditional expectation of sufficient statistics are calculated, and the maximization (M) step, in which the parameters $\boldsymbol{\Theta}$ are updated. Note that derivations in this work are partially similar to these presented in [6], [12], [13], therefore, for brevity, only final equations are provided when possible. Step E consists of the following updates:

$$\widehat{\mathbf{R}}_{xx,fn} = \mathbf{A}_f \mathbf{R}_{ss,fn} \mathbf{A}_f^H + \boldsymbol{\Sigma}_{b,f}, \qquad (9)$$

$$\widehat{\mathbf{R}}_{xs,fn} = \mathbf{R}_{xx,fn}\mathbf{G}_{s,fn}^{H}, \tag{10}$$

$$\widehat{\mathbf{R}}_{ss,fn} = \mathbf{G}_{s,fn}\mathbf{R}_{xx,fn}\mathbf{G}_{s,fn}^{H} + (\mathbf{I}_{JI} - \mathbf{G}_{s,fn}\mathbf{A}_f)\mathbf{R}_{ss,fn}, \tag{11}$$

$$\widehat{\xi}_{fn}^{(j)} = \widehat{\mathbf{R}}_{ss,fn}(j,j), \tag{12}$$

$$\mathbf{G}_{s,fn} = \mathbf{R}_{ss,fn}\mathbf{A}_f^{H}\widehat{\mathbf{R}}_{xx,fn}^{-1}. \tag{13}$$

Contrary to state-of-the-art approaches, we propose to compute matrix $\mathbf{R}_{ss,fn}$ using the following update rule

$$\mathbf{R}_{ss,fn} = \mathbf{\Gamma}_f \, \mathbf{\Sigma}_{s,fn}, \tag{14}$$

where $\mathbf{\Gamma}_f = \mathrm{diag}([\Gamma_f^{(1)}, \ldots, \Gamma_f^{(M)}]) \in \mathbb{R}_{+}^{M\times M}$ is a diagonal matrix which contains the attenuation weights for the successive 'lagged' subsources corresponding to the delayed late reverberation components, while a diagonal matrix composed of the variances of all subsources is given by

$$\mathbf{\Sigma}_{s,fn} = \mathrm{diag}([V_{fn}^{(1)}, \ldots, V_{fn}^{(J)}, |X_{\mathrm{ref},f,n-\delta}|^2, \ldots, |X_{\mathrm{ref},f,n-L_a}|^2]). \tag{15}$$

Note that (15) follows directly from the definition of $\boldsymbol{s}_{fn}$ in (3) and that the attenuation weights can be conveniently computed during the normalization of matrix $\mathbf{A}_f$, as described below.

In the M-step, we minimize the cost function (8) over $\boldsymbol{A}_f$, which yields the following update rule for the mixing matrix

$$\mathbf{A}_f = \left[\mathbf{\Sigma}_f^{-1} + \left(\overline{\mathbf{R}}_{ss,f} \otimes \mathbf{I}_I\right)^{T}\right]^{-1} \left[\mathbf{\Sigma}_f^{-1}\boldsymbol{U}_f + \overline{\mathbf{R}}_{xs,f}\right], \tag{16}$$

where the Kronecker product operator is denoted by $\otimes$, while the mean and covariance matrices $\boldsymbol{U}_f$ and $\mathbf{\Sigma}_f$ are given by

$$\boldsymbol{U}_f = [\boldsymbol{u}_f^{(1)}, \boldsymbol{u}_f^{(2)}, \ldots, \boldsymbol{u}_f^{(M)}]^{\mathrm{T}}, \tag{17}$$

$$\mathbf{\Sigma}_f = \begin{bmatrix} \mathbf{\Sigma}_f^{(1)} & & 0 \\ & \ddots & \\ 0 & & \mathbf{\Sigma}_f^{(M)} \end{bmatrix}, \tag{18}$$

$\overline{\mathbf{R}}_{xs,f} = \left[\overline{\mathbf{R}}_{xs,f}^{(1)}, \overline{\mathbf{R}}_{xs,f}^{(2)}, \ldots, \overline{\mathbf{R}}_{xs,f}^{(M)}\right]^{\mathrm{T}}$, $\overline{\mathbf{R}}_{ss,f}$, and $\overline{\mathbf{R}}_{xs,f}^{(m)}$ are power normalized matrices given by

$$\overline{\mathbf{R}}_{ss,f} = \frac{M\sum_n \widehat{\mathbf{R}}_{ss,fn}}{\mathrm{Tr}\{\sum_n \widehat{\mathbf{R}}_{ss,fn}\}}, \tag{19}$$

$$\overline{\mathbf{R}}_{xs,f}^{(m)} = \frac{M\sum_n \widehat{\mathbf{R}}_{xs,fn}^{(m)}}{\mathrm{Tr}\{\sum_n \widehat{\mathbf{R}}_{ss,fn}\}}. \tag{20}$$

Note that the presented power normalization allows to estimate the spatio-spectral properties without any prior knowledge about the room or reverberation.

Importantly, normalization of $\mathbf{A}_f$ involves, firstly, computation of the attenuation weights according to

$$\Gamma_f^{(m)} = \frac{1}{I}\sum_{i=1}^{I} |A_{i,f}^{(m)}|^2, \tag{21}$$

and secondly, normalizing each column of $\mathbf{A}_f$ by its value for the reference microphone, i.e. as $[\overline{\boldsymbol{A}}_f]^{(m)} = \frac{1}{A_{\mathrm{ref},f}^{(m)}}[\boldsymbol{A}_f]^{(m)}$.

Finally, the parameters of the NTF model are updated using multiplicative update rules given by

$$Q_k^{(j)} \longleftarrow Q_k^{(j)} \frac{\sum_{fn} W_{fk}H_{kn}\hat{\xi}_{fn}^{(j)}(V_{fn}^{(j)})^{-2}}{\sum_{fn} W_{fk}H_{kn}(V_{fn}^{(j)})^{-1}}, \tag{22}$$

$$W_{fk} \longleftarrow W_{fk} \frac{\sum_{jn} H_{kn}Q_k^{(j)}\hat{\xi}_{fn}^{(j)}(V_{fn}^{(j)})^{-2}}{\sum_{jn} H_{kn}Q_k^{(j)}(V_{fn}^{(j)})^{-1}}, \tag{23}$$

$$H_{kn} \longleftarrow H_{kn} \frac{\sum_{jf} W_{fk}Q_k^{(j)}\hat{\xi}_{fn}^{(j)}(V_{fn}^{(j)})^{-2}}{\sum_{jf} W_{fk}Q_k^{(j)}(V_{fn}^{(j)})^{-1}}, \tag{24}$$

which ought to be computed interchangeably several times per iteration. In order to avoid scale, phase and permutation indeterminacy, matrices $\boldsymbol{Q}, \boldsymbol{W}$, and $\boldsymbol{H}$ are normalized following the procedure described in detail in [5]. As a result, all time-dependent amplitude information is relegated to matrix $\boldsymbol{H}$.

The noise matrix $\mathbf{\Sigma}_b$ is updated using simulated annealing with the aim to accelerate algorithm convergence and reduce the likelihood of getting stuck at local minima (see [5] for details).

Finally, based on the estimated spatial and spectro-temporal information, extraction of the separated non-reverberant source signals is achieved using the convolutional weighted parametric multichannel Wiener filter recently presented in [20].

2.4. SDS-MNMF algorithm with a simplified signal model

In order to reduce the computational complexity of DS-MNMF that rises exponentially with an increasing source number J and late reverberation delay L, we propose a simplified version of the proposed algorithm, referred hereafter as SDS-MNMF, in which we simplify the signal model (2) to obtain

$$\boldsymbol{x}_{fn} = \sum_{j=1}^{J} \boldsymbol{a}_f^{(j)} S_{fn}^{(j)} + \boldsymbol{c}_f \sum_{\tau=\delta}^{L_a} \Gamma_{f,\tau} X_{\mathrm{ref},f,n-\tau} + \boldsymbol{b}_f. \tag{25}$$

Note that in (25), there is only a single 'resultant' transfer function $\boldsymbol{c}_f$, while the attenuation weights for subsequent time lags $\Gamma_{f,\tau}$ are included in the model. Such a simplification results in lowering the overall number of subsources from $M = J + L$ to $M = J + 1$. The mixing matrix then becomes $\boldsymbol{A}_f = [(\boldsymbol{a}_f^{(1)})^T, \ldots, (\boldsymbol{a}_f^{(J)})^T, \boldsymbol{c}_f^T]^T \in \mathbb{C}^{I\times(J+1)}$ and (14) changes to

$$\mathbf{R}_{ss,fn} = \mathrm{diag}([V_{fn}^{(1)}, \ldots, V_{fn}^{(j)}, \sum_{\tau=\delta}^{L_a} \Gamma_{f,\tau}|X_{\mathrm{ref},f,n-\tau}|^2]), \tag{26}$$

while other update equations from Sec. 2.3 remain unchanged, apart from (21), and hence this algorithm requires prior estimation of the attenuation matrix $\mathbf{\Gamma}_f$. In experimental evaluation presented in this paper, in order to obtain $\mathbf{\Gamma}_f$ without additional prior knowledge on room acoustics, we run 10 iterations of the basic DS-MNMF algorithm, and then run the SDS-MNMF algorithm with fixed $\mathbf{\Gamma}_f$ values.

3. Experimental Evaluation

The evaluation dataset consisted of two-channel reverberant mixtures with highly overlapping speech, obtained by convoloving clean speech signals of $J = 2$ with the respective room impulse responses (RIRs) between the sources and $I = 2$ microphones. As clean signals, 2620 speech utterances from the Librispeech *test-clean* [22] part were used, such that each of 1310 mixtures contained the signals of two different speakers of a similar length. The RIRs were either taken from the REVERB challenge [23] with rooms denoted as *small*, *medium* and *large* or simulated via the image-source method [24] with 60 dB reverberation times (RT60) of 300, 600, and 900 ms. Signals were sampled at 16 kHz and processed with 512 point STFT with 50% overlap. As evaluation metrics, the word error

Table 1: *Word error rate (WER) [%] and signal-to-distortion ratio (SDR) results for various reverberation conditions in simulated and real rooms obtained for the two proposed and a number of state-of-the-art source separation algorithms, also with WPE preprocessing.*

RIR type	Simulated with Image Source Method						Real from REVERB Challenge					
RT60 [ms] / Room	300		600		900		*small*		*medium*		*large*	
Measure	WER	SDR	WER	SDR	WER	SDR	WER	SDR	WER	SDR	WER	SDR
microphone mixture	105.10	-0.5	106.36	-2.0	106.92	-3.4	103.01	-0.2	102.63	-1.4	103.86	-2.3
single source with reverberation	2.78	11.8	4.45	5.4	11.56	2.2	2.48	15.6	2.94	7.5	9.81	4.7
GEM-MU [15]	86.11	2.4	90.62	-0.1	94.13	-2.3	86.9	3.2	94.83	1.8	96.21	1.3
Fast-MNMF [21]	87.09	2.4	102.52	-0.5	107.99	-2.3	66.65	2.8	85.87	0.3	92.11	-0.9
WPE [4] + GEM-MU [15]	53.94	6.7	63.06	4.7	79.50	2.7	83.71	3.3	91.50	2.0	92.33	1.5
WPE [4] + Fast-MNMF [21]	77.44	2.2	93.11	0.6	103.33	-0.9	61.45	2.6	74.71	1.7	82.20	1.2
DS-MNMF with $L = 0$ (no reverb)	13.21	9.9	32.81	6.2	50.41	3.2	20.24	7.7	33.22	5.0	37.56	3.9
WPE + DS-MNMF with $L = 0$	10.67	10.2	28.02	7.0	46.65	4.0	19.71	7.3	32.31	5.6	37.79	4.5
Proposed DS-MNMF	9.40	**11.1**	**25.07**	**7.7**	**43.02**	**5.2**	19.09	**7.9**	31.27	**6.2**	35.72	**5.2**
Proposed SDS-MNMF	**9.34**	**11.1**	25.44	7.6	44.09	5.1	**18.65**	7.8	**30.40**	6.0	**35.61**	5.0

Table 2: *Word error rate (WER) [%] results for a different number of delayed late reverberation subsources, for the proposed original (DS-MNMF) and simplified (SDS-MNMF) algorithms.*

RT60 [ms]		300	600	900
DS-MNMF	$L = 0$	13.21	32.81	50.41
	$L = 2$	10.20	27.48	45.47
	$L = 4$	9.40	**25.07**	**43.02**
	$L = 6$	**9.23**	25.28	43.66
SDS-MNMF	$L = 0$	13.21	32.81	50.41
	$L = 2$	10.20	27.29	45.47
	$L = 4$	9.34	**25.44**	**44.09**
	$L = 6$	**9.33**	25.95	44.49

rate (WER) was used to assess the performance of the ASR on the separated signals and the Signal-to-Distortion Ratio (SDR) to assess the quality of source separation and dereverberation, respectively. In the ASR task, we used the pretrained *asr-transformer-transformerlm* model [25] from the SpeechBrain toolkit [26] with the beam size of 10 and CTC weight set as 0.52 during tests.

3.1. Results of experiments and discussion

In the first experiment, we aim to verify the influence of the number of delayed subsources L on the performance of the two proposed algorithms, namely of the original DS-MNMF and the simplified SDS-MNMF, in various reverberation levels. The WER results presented in Table 2 for the simulated RIRs indicate that the performance of both algorithms improves for an increasing number of 'time-lags' that model late reverberation, with a significant improvement observed for $L > 2$, and reaching an optimum value of $L = 4$ for the considered reverberation levels and the STFT frame size. Note that the simplified algorithm achieves nearly similar gains as the original algorithm, while it provides faster performance due to the smaller sizes of the mixing and subsource covariance matrices.

In the second experiment, we compare the proposed algorithms against popular state-of-the-art separation methods on datasets with real and simulated RIRs. As the main reference (baseline) algorithm, we used the generalized EM algorithm with multiplicative updates (GEM-MU) [15], which is similar in structure to the proposed algorithm, but without the localization prior and additional, 'lagged' subsources that model late reverberation. As a second reference algorithm, a popular Fast-MNMF [21] was used. Both methods were also tested in sequence with the generalized WPE [4] method, a popular approach to perform multichannel dereverberation prior to other speech enhancement and separation tasks. Dereverberation followed by separation is then denoted as WPE + GEM-MU and WPE + Fast-MNMF, respectively, and for the WPE preprocessing parameter L was selected according to the formula presented in [27] for optimum dereverberation, which results in $L = 8, 15, 30$ for RT60 $= 300, 600$ and 900 ms, respectively. Apart from the two proposed algorithms, in which we set $L = 4$ as concluded from the first experiment, we also present the results of the DS-MNMF with the localization prior but without any 'time-lags' for modelling late reverberation, which is performed on signals dereverberated with WPE preprocessing. For reference, we also present the results obtained for the reverberant mixture (without any separation or dereverberation performed) as well as for the reverberant speech of a single speaker.

The results of the second experiment in terms of WER and SDR values are presented in Table 1. As can be observed from the first two rows, performing ASR on microphone mixtures turns out to be a formidable task, while reverberation alone has a negative yet minor effect on the overall metric values. Importantly, both reference algorithms struggle to converge in most cases, which results in high WER and low SDR results. Although WPE preprocessing can help to mitigate the negative effect of reverberation, a non-optimum sequential processing seems insufficient for the joint separation and dereverberation task. In contrast, the proposed algorithm is jointly optimum, and hence it yields the best results by a high margin. Interestingly, incorporation of the localization prior seems to make a significant performance boost, while the simplified algorithm performs almost equally well as the original DS-MNMF.

4. Conclusions

This paper presents a novel MNMF algorithm for joint separation and dereverberation in which late reverberation components are modelled with delayed subsources using past time frames. Experimental evaluations show that both proposed algorithms outperform state-of-the-art methods in the joint task.

5. Acknowledgements

This research was funded in part by the National Science Centre, Poland DEC-2021/42/E/ST7/00452, by program "Excellence initiative – research university" for AGH-UST, by the Foundation for Polish Science First TEAM/2017-3/23 (POIR.04.04.00-00-3FC4/17-00), and supported by PLGrid infrastructure. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

6. References

[1] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2019, pp. 449–456.

[2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2018, pp. 4774–4778.

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

[4] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[6] 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.

[7] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[8] 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.

[9] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Autoregressive fast multichannel nonnegative matrix factorization for joint blind source separation and dereverberation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 511–515.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[11] N. Ito and T. Nakatani, "Fastmnmf: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 371–375.

[12] 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, p. 149, 2013.

[13] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 526–530.

[14] M. Fras, M. Witkowski, and K. Kowalczyk, "Combating reverberation in ntf-based speech separation using a sub-source weighted multichannel wiener filter and linear prediction." in *Interspeech*, 2021, pp. 3895–3899.

[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] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[17] P. Naylor and N. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer-Verlag, 2010.

[18] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," *Microphone arrays: signal processing techniques and applications*, pp. 157–180, 2001.

[19] 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.

[20] M. Fraś and K. Kowalczyk, "Convolutional weighted parametric multichannel wiener filter for reverberant source separation," *IEEE Signal Processing Letters*, vol. 29, pp. 1928–1932, 2022.

[21] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, 2016.

[24] 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.

[25] "ASR model based on transformers trained using Librispeech," https://huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech, accessed: 19.01.2023.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

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