Sparse Linear Prediction-based Dereverberation for Signal Enhancement in Distant Speaker Verification

Marcin Witkowski, Magdalena Rybicka, and Konrad Kowalczyk
AGH University of Science and Technology, Institute of Electronics, Kraków, Poland
witkow@agh.edu.pl, mrybicka@agh.edu.pl, konrad.kowalczyk@agh.edu.pl

***Abstract*—In this paper we investigate several sparse dereverberation methods which provide signal enhancement in reverberant environments with the aim to apply them as a preprocessing step for the distant speaker verification (SV) task. First, we present multichannel linear prediction (LP) based techniques which promote sparsity of the dereverberated speech, whose performance has never been verified in the context of speaker recognition. In particular, we describe two existing sparse LP-based methods and present a novel LP-based method in which speech sparsity is enforced by adopting the so-called split Bregman approach. We then study the performance of both sparse and nonsparse dereverberation approaches for signal enhancement, and investigate the gain offered by these methods when applied as a preprocessing step for two different speaker verification systems based on DNN-based speaker embedding extraction. The results of performed experiments indicate that the proposed sparse approach and one of the existing methods consistently achieve significant improvements in distant speaker verification in reverberant environments, and that the SV results are well in line with signal enhancement achieved by the compared techniques.**

***Index Terms*—speaker recognition, dereverberation, linear prediction, sparse optimization, DNN-based speaker embedding**

I. Introduction

It is well known that the performance of speaker verification (SV) systems degrades when acoustic conditions are different to those in the training dataset. Recent solutions to tackle this problem require large amount of training data [1] or carefully prepared datasets that enable system adaptation by transfer learning [2], [3], enrollment adaptation [3] or backend modelling [4], [5]. To mitigate the detrimental effect of room reverberation, an alternative approach consists in speech enhancement as a front-end processing to the speaker recognition system. Recent VOiCES from a Distance Challenge [6] demonstrated the popularity of a linear prediction (LP) based dereverberation technique known as the Weighted Prediction Error (WPE) [7], as an effective method to cope with the reverberant signal in the speaker recognition task [6], [8]–[10]. Since nonreverberant speech exhibits a sparse structure in the time-frequency domain, compared with the representation of the reverberant speech, speech sparsity has been additionally exploited in the literature to achieve stronger dereverberation [11], [12]. Such methods either assume sparse signal distributions, e.g. a Complex Generalized Gaussian (CGG) [11] or a Laplacian [13] distribution, or incorporate ℓ_1-norm sparsity terms in the cost function such as in Alternating Direction Method of Multipliers (ADMM) [12].

This paper presents the performance gain offered by using sparse LP-based dereverberation as a pre-processing step for speaker verification in reverberant environments, which to our best knowledge has not yet been reported in the literature. Specifically, we present a novel sparse split Bregman (SSB) dereverberation method [14], and compare it against three existing techniques, namely the nonsparse WPE method [7] and two sparse methods known as CGG [11] and ADMM [12]. The SSB method can be considered a generalization of [7], [11] and [12]. In particular, its cost function follows from sparse distribution as in CGG [11], however, the sparsity term of ADMM [12] is also incorporated, with an additional term which ensures the LP-based solution [7]. This leads to a great improvement of the quality of the dereverberated signal over the ADMM method and better control of sparsity than in the CGG method. Stronger suppression of reverberation is shown to further enhance the speech signal, and also to increase the accuracy of distant speaker recognition. The SV results are presented for two SV systems, namely the state-of-the-art DNN-based speaker embedding extractor known as the Time Delay Neural Network (TDNN) [15] and a modified ResNet18 which has been proposed by the current authors in [16].

II. Sparse linear prediction based dereverberation

In this section, we present a signal model used in linear prediction based dereverberation, describe state-of-the-art nonsparse and sparse methods, and present a novel method which promotes speech sparsity using the split Bregman approach.

The aim of linear prediction based dereverberation is to estimate the desired speech component (i.e., speech that propagates over the direct and early reflection paths) by subtracting the undesired reverberant speech component (i.e., speech convolved with late room reverberation) predicted using a linear filter from the microphone signals. The signal of the m-th microphone in the short-time Fourier transform (STFT) domain is given by

$$\mathbf{x}^m = \mathbf{d}^m + \mathbf{X}_D \ \mathbf{c}, \tag{1}$$

where $\mathbf{d}^m = [d^m(k,0), d^m(k,1), ..., d^m(k,N-1)]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ is the desired signal vector, the prediction filter coefficient vec-

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 European Union under the European Regional Development Fund, from the National Science Centre under grant number DEC-2017/25/B/ST7/01792, and was supported in part by PLGrid Infrastructure.

tor is denoted as $\mathbf{c} = [(\mathbf{c}^0)^{\mathrm{T}}, (\mathbf{c}^1)^{\mathrm{T}}, ..., (\mathbf{c}^{M-1})^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{ML_c \times 1}$ with $\mathbf{c}^m = [c(k,0), c(k,1), ..., c(k, L_c - 1)]^{\mathrm{T}} \in \mathbb{C}^{L_c \times 1}$, and the convolution matrix delayed by D time frames is given by $\mathbf{X}_D = [\bar{\mathbf{x}}_D(0), \bar{\mathbf{x}}_D(1), ..., \bar{\mathbf{x}}_D(N-1)]^{\mathrm{T}} \in \mathbb{C}^{N \times ML_c}$ where $\bar{\mathbf{x}}_D(n) = [(\mathbf{x}_D^0)^{\mathrm{T}}(n), (\mathbf{x}_D^1)^{\mathrm{T}}(n), ..., (\mathbf{x}_D^{M-1})^{\mathrm{T}}(n)]^{\mathrm{T}} \in \mathbb{C}^{ML_c \times 1}$ with $\mathbf{x}_D^m(n) = [x^m(k, n-D), x^m(k, n-D-1), ..., x^m(k, n-D-L_c+1)]^{\mathrm{T}} \in \mathbb{C}^{L_c \times 1}$. The microphone index is given by $m = 0, 1, ..., M-1$, the length of the prediction filter is denoted as L_c, whilst $n = 0, 1, ..., N-1$ and $k = 0, 1, ..., K-1$ denote the time and frequency indices. In the following, we assume that each subband frequency in the STFT domain is modelled independently and that we dereverberate the signal of the reference microphone. Thus hereafter we discard the microphone and frequency indices.

A. Nonsparse linear prediction dereverberation

The most popular nonsparse LP-based dereverberation technique is the Weighted Prediction Error (WPE) method [7], which is derived by modeling the desired speech in each frequency bin using a circular complex Gaussian distribution $\mathcal{N}_{\mathbb{C}}\big(d(n); 0, \lambda_d(n)\big)$ with zero mean and an unknown time-varying variance $\lambda_d(n)$. With a linear relation (1) between the desired signal $\mathbf{d}$ and the filter coefficients $\mathbf{c}$, the likelihood function is defined as

$$\mathcal{L}\big(\mathbf{c}, \boldsymbol{\lambda}_d\big) = \prod_{n=0}^{N-1} \frac{1}{\pi \lambda_d(n)} e^{-\frac{|d(n)|^2}{\lambda_d(n)}}, \tag{2}$$

which yields the negative log-likelihood to be minimized

$$\min_{\mathbf{c}, \boldsymbol{\lambda}_d} \sum_{n=0}^{N-1} \frac{|d(n)|^2}{\lambda_d(n)} + \log_e \lambda_d(n) \ . \tag{3}$$

In order to estimate $\mathbf{c}$ and $\boldsymbol{\lambda}_d = [\lambda_d(0), ..., \lambda_d(N-1)]^{\mathrm{T}} \in \mathbb{R}^{N \times 1}$, the joint optimization (3) is split into two sub-problems that are solved in an alternating fashion. The method that minimizes (3) will be referred to as the WPE method [7].

B. Promoting sparsity using a sparse prior distribution

Speech sparsity in the STFT domain can be modeled using an appropriate sparse distribution. In [11] Jukic et al. propose to use the so-called Complex Generalized Gaussian (CGG) as a general circular sparse prior, for which the likelihood function can be represented as

$$\mathcal{L}\big(\mathbf{c}, \boldsymbol{\lambda}_d\big) = \prod_{n=0}^{N-1} \max_{\lambda_d(n) \geq 0} \ \frac{1}{\pi \lambda_d(n)} e^{-\frac{|d(n)|^2}{\lambda_d(n)}} \ \psi\big(\lambda_d(n)\big), \tag{4}$$

where $\psi\big(\lambda_d(n)\big)$ is a scaling function which makes the distribution sparse. As shown in [11], the resulting negative log-likelihood cost function is then given by

$$\min_{\mathbf{c}, \boldsymbol{\lambda}_d} \sum_{n=0}^{N-1} \frac{|d(n)|^2}{\lambda_d(n)} + \log_e \lambda_d(n) - \log_e \psi\big(\lambda_d(n)\big). \tag{5}$$

In the following, the method that minimizes (5) will be referred to as the CGG method [11].

C. Promoting sparsity through the sparsity term

Another approach to enforcing sparsity of the solution is to incorporate an additional sparsity term in the cost function, typically formulated as the first-order or zero-order norm on the desired speech. In [12] the cost function consists of only the sparsity term with a constraint which follows the LP-based signal model (1), which reads

$$\min_{\mathbf{d}, \mathbf{c}} \|\mathbf{d}\|_{\mathbf{w},1} \quad \text{subject to} \quad \mathbf{d} + \mathbf{X}_D\, \mathbf{c} = \mathbf{x}, \tag{6}$$

where $\|\mathbf{d}\|_{\mathbf{w},1} = \sum_{n=0}^{N-1} w(n)|d(n)|$ denotes weighted ℓ_1-norm of vector $\mathbf{d}$ with nonnegative weights $\mathbf{w} = [w(0), ..., w(N-1)]^{\mathrm{T}} \in \mathbb{R}_{>0}^{N \times 1}$. The constrained problem (6) can then be solved using the so-called Alternating Direction Method of Multipliers (ADMM) [12]. Thus the method that minimizes (6) will be referred to as ADMM.

D. Sparse split Bregman method

In this section, we present a novel sparse dereverberation method in which dereverberated speech is modelled using sparse prior distribution. However, in addition, the sparsity term is integrated into the second cost function in order to ensure that obtained linear prediction based solution is indeed sparse. The proposed optimization can be formulated by the following two cost functions which are solved in an alternating fashion

$$\operatorname*{argmin}_{\boldsymbol{\lambda_d}} \mathbf{d}^{\mathrm{H}} \mathcal{D}_{\boldsymbol{\lambda}_d}^{-1} \mathbf{d} + \log_e \det\{\mathcal{D}_{\boldsymbol{\lambda_d}}\} - \log_e \det\{\mathcal{D}_{\psi(\boldsymbol{\lambda_d})}\} \tag{7a}$$

$$\operatorname*{argmin}_{\mathbf{d}} \mathbf{d}^{\mathrm{H}} \mathcal{D}_{\boldsymbol{\lambda}_d}^{-1} \mathbf{d} + \|\mathbf{d}\|_{\mathbf{w},1} \ \text{ subject to } \ \mathbf{d} + \mathbf{X}_D\, \mathbf{c} = \mathbf{x}, \tag{7b}$$

where $\mathcal{D}_{\boldsymbol{\lambda}_d} = \operatorname{diag}\{\boldsymbol{\lambda}_d\}$ denotes the diagonal matrix with elements of vector $\boldsymbol{\lambda}_d$ on the main diagonal, and diagonal matrix $\mathcal{D}_{\psi(\boldsymbol{\lambda}_d)} = \operatorname{diag}\{\psi(\boldsymbol{\lambda}_d)\}$ is defined analogously. Optimization (7a) follows directly from (5) and thus it has a closed-form solution that is equivalent to the solution of the CGG method [11]. On the other hand, optimization (7b) is convex but nondifferentiable, and thus its closed-form solution does not exist. To address this problem we use the so-called split Bregman method [17] which enables to find the solutions to two constrained sub-problems: (i) differentiable problem corresponding to the first term in (7b) and (ii) the nondifferentiable problem corresponding to the second term in (7b), which are computed separately in an alternate fashion. The optimum solution of the latter sub-problem can be found using the so-called shrinkage operator [17] given by (8c). The final update equations for the sparse dereverberation method derived using the split Bregman approach are given by [14]

$$\boldsymbol{\lambda}_d^{(i)} = \max\big\{|\mathbf{d}^{(i-1)}|^{2-p}, \varepsilon_\lambda\big\}, \tag{8a}$$

$$\begin{aligned} \mathbf{c}^{(i)} &= \big(\mathbf{X}_D^{\mathrm{H}} \mathcal{D}_{\boldsymbol{\lambda}_d}^{-1} \mathbf{X}_D + \frac{\alpha}{2} \mathbf{X}_D^{\mathrm{H}} \mathbf{X}_D\big)^{-1} \\ &\quad \Big(\mathbf{X}_D^{\mathrm{H}} \mathcal{D}_{\boldsymbol{\lambda}_d}^{-1} \mathbf{x} + \frac{\alpha}{2} \mathbf{X}_D^{\mathrm{H}} \big(\mathbf{x} - \mathbf{d}^{(i-1)} + \boldsymbol{\mu}^{(i-1)}\big)\Big), \end{aligned} \tag{8b}$$

$$\mathbf{d}^{(i)} = \mathcal{D}_{\mathbf{h}} \max\big\{1 - \mathcal{D}_{\mathbf{w}}(\alpha|\mathbf{h}|)^{-1}, G_{\min}\big\}, \tag{8c}$$

$$\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}^{(i-1)} - \big(\mathbf{d}^{(i)} - \mathbf{x} + \mathbf{X}_D\, \mathbf{c}^{(i)}\big), \tag{8d}$$

where vector $\mathbf{h}$ and matrices $\mathcal{D}_{\mathbf{h}}$ and $\mathcal{D}_{\mathbf{w}}$ in (8c) are given by

$$\mathbf{h} = \mathbf{x} - \mathbf{X}_D\mathbf{c}^{(i)} + \boldsymbol{\mu}^{(i-1)}, \tag{9}$$

$$\mathcal{D}_{\mathbf{h}} = \operatorname{diag}\{\mathbf{h}\}, \tag{10}$$

$$\mathcal{D}_{\mathbf{w}} = \operatorname{diag}\{(|\mathbf{h}| + \varepsilon_\lambda)^{-1}\}. \tag{11}$$

In the presented update equations i denotes the iteration index, α is the penalty parameter for the constraint in (7b), $\boldsymbol{\mu}$ denotes the so-called auxiliary Bregman variable, ε_λ and $G_{\min}$ denote small real constants for algorithm robustness, and function $\max\{[\cdot], \cdot\}$ returns a vector with maximum values.

The presented sparse split Bregman (SSB) method can be considered a generalization of the nonsparse WPE [7], sparse CGG [11], and sparse ADMM [12] methods. In fact, the update equations for WPE and CGG are given by (8a), (8b), and update on $\mathbf{d}$ is obtained by reordering (1) with parameters $\alpha = 0$ and $p = 0$ for WPE and $\alpha = 0$ for CGG. The ADMM is obtained by removing the first term from (7b) and dropping the cost function (7a), which yields (8b) with the first terms with $\mathcal{D}^{-1}_{\boldsymbol{\lambda}_d}$ in brackets removed, (8c) and (8d). Simplification of (8b) for the ADMM allows for the computation of $\left(\mathbf{X}_D^{\mathrm{H}}\mathbf{X}_D\right)^{-1}\mathbf{X}_D^{\mathrm{H}}$ outside the loop and leads to a significant computational complexity reduction compared with the other three methods.

III. DNN-based speaker verification

In this section, we provide an overview of two network architectures for speaker embedding extraction which are used in speaker verification experiments, namely, the well-known Time Delay Neural Network (TDNN) [15] and the modified ResNet18 (mR18) proposed by the current authors in [16]. In both systems, the extracted embeddings undergo cosine distance scoring (CDS) in order to avoid the impact of backend adaptation on the overall systems' performance.

A. TDNN-based speaker embedding

X-vectors are speaker embeddings extracted from a TDNN-based architecture, which consists of 5 time-delay (TD) layers that span over 15 time frames context of network input, followed by a pooling layer that computes mean and standard deviation [15]. Next the calculated statistics are propagated through a fully connected layer followed by the output softmax layer with the number of outputs equal to the number of speakers in a training set. The x-vector embedding is extracted from the output of the fully connected layer.

B. Modified ResNet18-based speaker embedding

The first layer of the modified ResNet18 [16] is the 2D convolutional layer with filter size of 7x7 and downsampling stride of 2x2. The next part of the network is composed of 4 main segments, each consisting of 2 ResNet blocks, where each of the ResNet blocks is built of 2 convolutional layers with identity shortcut connections. The convolutional layers in the ResNet block have a filter size of 3x3, have an identical output size and do not involve downsampling, except for the first layer of each segment which has a stride of 1x2. The output dimensions for each of 4 segments are {64, 128, 256, and 512}. The output from the residual part is forwarded to the statistics pooling layer which computes mean and standard deviation of the feature vector obtained after ResNet segments in the time dimension. The output of statistics pooling is fed to the fully connected layer with output size of 512, followed by a softmax output layer with dimension equal to the number of speakers in the training set. Speaker embedding is extracted as the output of the fully connected layer.

IV. Performed experiments

In experiment 1, we evaluate the dereverberation performance of sparse and nonsparse LP-based methods presented in Sec. II. The microphone signals are obtained as a convolution of 2620 nonreverberant speech files from the Librispeech *test-clean* subset [18] with Room Impulse Responses (RIRs) generated using the image-source method [19]. To simulate rooms with reverberation times (RT60) ranging from 0.4 to 1 s, we randomly select room dimensions 5 - 10 m for the width and length, 2 - 4 m for the room height, and appropriately adjust the wall absorption coefficients from the range 0.1 - 0.5. In each room, we generate RIRs for 5 random positions of the source and a 2-element and 4-element circular microphone array with random source-array distances of 1 - 2 m and inter-microphone spacing of 0.2 - 0.3 m, respectively. Each evaluation result is obtained by averaging over the results computed for all 2620 speech files, each convolved with a randomly selected RIR simulated for a given RT60. As evaluation metrics we use the Perceptual Evaluation of Speech Quality (PESQ) [20], the Frequency Weighted Segmental Signal-to-Noise Ratio (WSNR) [21], and the Cepstral Distance (CD) [21]. The presented improvement values of ΔPESQ and ΔWSNR are calculated as the differences between the output dereverberated signal and the reverberant input signal at the reference microphone, whilst the CD improvement (ΔCD) is computed as a subtraction of CDs for the input and the output signals.

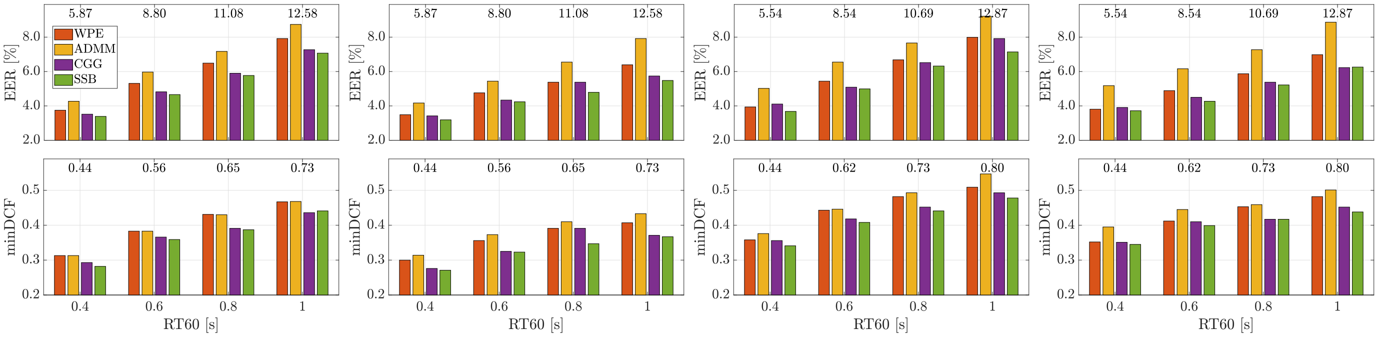
In experiment 2 we investigate the influence of applying a sparse dereverberation method on the accuracy of two SV systems described in Sec. III. Both embedding extractors are trained similarly as in the VoxCeleb Kaldi recipe, using the entire VoxCeleb2 [22] and VoxCeleb1 [23] training part, augmented only with environmental noise and music from the MUSAN database [24]. In order to relate the SV results to the results of experiment 1, the reverberant dataset is the same in both experiments. Text-independent trials are generated so that the first 4 speech samples from each chapter are used for enrollment and reverberant samples for scoring. All in all, the trial set contains 3068 target and 33024 nontarget trials. Note that there are no cross-gender nontarget trials and no semi-chapter target trials. Finally, note that the output signal, i.e. after dereverberation, is computed using (1) with the filter coefficients obtained in the last iteration of the algorithm.

For both SV systems, as input features to the DNN we use 64-dimensional log-Mel filter bank coefficients obtained from 25 ms frames with 10 ms overlap. Feature extraction is followed by mean-normalization over the 3 s window and energy-based Kaldi's Voice Activity Detector [25]. We used two

TABLE I

Dereverberation performance of the WPE, ADMM, CGG, and the SSB methods in terms of CD, PESQ, and WSNR improvements between the output and input values for simulated data with 2- and 4-channel setup (higher values indicate better performance). The reference values (for unprocessed signal of the closest microphone) for PESQ, CD and WSNR are {1.47, 1.28, 1.20 1.16}, {3.46, 4.25, 4.73, 5.08} and {7.62, 4.84, 3.25, 2.24}, respectively, for RT60 equal to {0.4, 0.6, 0.8 and 1 s}.

Measure	Δ PESQ								Δ CD								Δ WSNR							
RT60 [s]	0.4		0.6		0.8		1		0.4		0.6		0.8		1		0.4		0.6		0.8		1	
Num. ch.	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4
WPE	0.59	0.65	0.51	0.58	0.40	0.51	0.34	0.48	0.98	1.08	1.21	1.32	1.24	1.42	1.23	1.56	4.35	5.02	5.04	5.78	4.86	5.91	4.70	6.27
ADMM	0.55	0.55	0.47	0.50	0.38	0.43	0.30	0.38	0.80	0.77	0.99	1.08	1.07	1.19	1.09	1.31	3.44	3.22	3.85	4.03	3.94	4.28	3.74	4.30
CGG	0.68	0.70	0.63	0.69	0.53	0.62	0.44	0.59	1.15	1.21	1.41	1.50	1.44	1.60	1.43	1.75	4.51	5.23	5.29	6.33	5.16	6.59	4.96	6.95
SSB	**0.81**	**0.84**	**0.73**	**0.79**	**0.61**	**0.72**	**0.52**	**0.69**	**1.18**	**1.26**	**1.46**	**1.55**	**1.51**	**1.67**	**1.52**	**1.84**	**4.95**	**5.59**	**5.69**	**6.55**	**5.50**	**6.71**	**5.22**	**6.95**



(a) mR18, 2-mic (b) mR18, 4-mic (c) TDNN, 2-mic (d) TDNN, 4-mic

Fig. 1. SV performance for the mR18 (a,b) and the TDNN (c,d) systems with a 2-channel (a,c) and 4-channel (b,d) dereverberation as front-end processing for RT60s equal to 0.4, 0.6, 0.8, and 1 s (lower values indicate better performance). Unprocessed reference values are reported at the top of the bar plots.

evaluation metrics, namely the Equal Error Rate (EER) and the minimum value of the Detection Cost Function (minDCF), which is computed for parameters $C_{miss} = C_{fa} = 1$ and $P_{tar} = 0.01$.

Dereverberation is performed in the STFT domain with 20 ms frame length and 10 ms hop multiplied by a Hamming window, whilst half-rectangular Tuckey window is used in the ISTFT synthesis. As reference microphone (at which input metrics are computed), we select the microphone that is the closest to the source. The filter length is set according to the empirically found formula $L_c = 50 \cdot \mathrm{RT60}$ (halved for 4-element array) and prediction delay is set to $D = 1$. The parameters for ADMM and the proposed SSB method are empirically adjusted using a small subset of simulation data. The parameters of the proposed SSB method are set as $\alpha = 0.01$, $G_{\min} = 0.06$, and $p = 0.5$. Referring to notation in equations (15), (19) and (22) in [12] we set the parameters for ADMM method as $G_{\min} = 0.04$, $\rho = 1$, $\varepsilon = 10^{-10}$ and $\gamma = 1$, while the weights are estimated using formula (19) from [12]. For the CGG and SSB methods, parameter p is set to 0.5, as in [11]. All methods are initialized with $\mathbf{c} = \mathbf{0}_{ML_c \times 1}$, $\boldsymbol{\mu} = \mathbf{0}_{N \times 1}$ and $\mathbf{d} = \mathbf{x}$. Alternating update loop is finished either when the maximum number of 20 iterations is reached or the convergence condition $||\hat{\mathbf{d}}^{(i)} - \hat{\mathbf{d}}^{(i-1)}||_2 < 10^{-3}$ is met.

V. Results and discussion

Table I presents the results of experiment 1, in which we compare the dereverberation performance of three existing methods, namely of the nonsparse WPE, sparse CGG, sparse ADMM, and the sparse SSB method in terms of signal enhancement for different reverberation times in simulated rooms. The results show the differences between values obtained for the dereverberated signal and the unprocessed signal at the microphone closest to the source.

As can be observed, all four methods successfully reduce reverberation. Among the compared approaches, the SSB method introduces the strongest reverberation suppression for all examined rooms and evaluation measures. Specifically, the improvements achieved by SSB are 37%-53%, 20%-24% and 11%-13% over the improvements obtained with the nonsparse WPE method for ΔPESQ, ΔCD and ΔWSNR, respectively. In relation to the second-best performing CGG, the SSB method achieves significant 14%-20% improvement for ΔPESQ and slight gain of up to 6% and 10% for ΔCD and ΔWSNR, respectively. In contrast, the third sparse technique referred to as ADMM does not succeed in preserving the quality of speech in the output signal, which can be seen in all evaluation measures, which may be attributed to the lack of the first term in cost function (7b).

Increasing the number of channels brings about a significant performance gain regardless of the dereverberation technique. Note that for the best performing techniques, namely the SSB and CGG, an improvement is less pronounced when increasing the number of microphones in comparison with the nonspare WPE method, which achieves even 41% and 27% of relative improvement at RT60 = 1 s in CD and PESQ score. These results clearly show the superiority of the sparse approaches which provide very good dereverberation even for a low number of microphones.

Figure 1 reports the performance of two speaker verification systems, namely the mR18 and the TDNN, evaluated using the unprocessed signals and the signals dereverberated with the nonsparse WPE and three sparse methods for the 2- and 4-channel scenarios. Note that for clean (nonreverberant) data and the used trial set, the mR18 and the TDNN systems obtain the vanilla EER / minDCF of 1.27% / 0.188 and 1.60% / 0.239, respectively. Thus in general the modified ResNet-based system provides slightly lower errors than the TDNN-based system. As can be observed in Fig. 1, there is a clear, high gain in speaker recognition performance in both minDCF and EER when any dereverberation method is applied. Among the dereverberation techniques, the SV accuracy is always better for the SSB algorithm than for the nonsparse WPE and the sparse ADMM algorithms, while it achieves slightly better results than CGG for the vast majority of investigated scenarios. A general conclusion can be made that two sparse methods always outperform the nonsparse WPE, while the ADMM algorithm, with a slight speech signal degradation, is not a good choice among the LP-based dereverberation techniques for distant speaker verification.

For both SV systems, an improvement in EER and minDCF is more apparent for higher reverberation times, which is similar to the trend of the CD and the WSNR improvements presented in Table I. The highest relative improvement of 57% and 56% in terms of EER compared with the unprocessed signals is observed for the mR18 system with dereverberation using the sparse SSB method at RT60 equal to 0.8 s and 1 s, respectively. Similarly, the gain offered by 4-channel over the 2-channel pre-processing is more clearly exhibited at high RT60 values. For instance, a relative improvement of 23% can be observed for the 4-channel dereverberation over the 2-channel case for mR18 system with SSB preprocessing and RT60 of 1 s. Finally note that the benefits of sparse processing as well as using more signals in multichannel LP-based dereverberation are more significant at high reverberation times.

VI. Conclusions

In this paper, we have evaluated the application of several sparse and nonsparse LP-based dereverberation methods for signal enhancement in reverberant environments with the aim to preprocess the microphone signals for subsequent speaker verification. The results of experiments have shown that sparse methods offer significant gain over the typically used nonsparse counterpart in both studied tasks. In particular, we have shown that the novel SSB method performs particularly well.

References

[1] M. Mclaren, D. Castán, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," in *Proc. of Speaker Odyssey*, 2018, pp. 327–334.

[2] I. Szoke, M. Skacel, L. Mosner, J. Paliesek, and J. Cernocky, "Building and Evaluation of a Real Room Impulse Response Dataset," *IEEE J. Sel. Topic Signal Process.*, 2019.

[3] X. Qin, D. Cai, and M. Li, "Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interspeech*, 2019, pp. 4045–4049.

[4] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 4253–4256.

[5] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 4257–4260.

[6] M. K. Nandwana, J. van Hout, C. Richey, M. McLaren, M. A. Barrios, and A. Lawson, "The VOiCES from a Distance Challenge 2019," in *Proc. Interspeech 2019*, 2019, pp. 2438–2442.

[7] 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.

[8] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, G. Lavrentyeva, V. Volokhov, and A. Kozlov, "STC Speaker Recognition Systems for the VOiCES from a Distance Challenge," in *Proc. Interspeech*, 2019.

[9] P. Matějka, O. Plchot, H. Zeinali, L. Mošner, A. Silnova, L. Burget, O. Novotný, and O. Glembek, "Analysis of BUT Submission in Far-Field Scenarios of VOiCES 2019 Challenge," in *Proc. Interspeech*, 2019, pp. 2448–2452.

[10] T. Y. Chong, K. M. Tan, K. K. Teh, C. H. You, H. Sun, and H. D. Tran, "The I2R's ASR System for the VOiCES from a Distance Challenge 2019," in *Proc. Interspeech*, 2019, pp. 2458–2462.

[11] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[12] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "A general framework for incorporating time–frequency domain sparsity in multichannel speech dereverberation," *J. of the Audio Eng. Society*, vol. 65, no. 1/2, pp. 17–30, 2017.

[13] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2014, pp. 5172–5176.

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

[15] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 5329–5333.

[16] M. Rybicka and K. Kowalczyk, "On Parameter Adaptation in Softmax-based Cross-Entropy Loss for Improved Convergence Speed and Accuracy in DNN-based Speaker Recognition," in *Proc. Interspeech*, 2020.

[17] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. On Imaging Sciences*, vol. 2, no. 2, pp. 323–343, 2009.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[20] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2007.

[22] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," *Proc. Interspeech*, pp. 2616–2620, 2017.

[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE Automatic Speech Rec. and Unders. Workshop (ASRU)*, 2011.