MULTI-CHANNEL LINEAR PREDICTION-BASED SPEECH DEREVERBERATION WITH LOW-RANK POWER SPECTROGRAM APPROXIMATION

Ante Jukić[1], Nasser Mohammadiha[1], Toon van Waterschoot[2], Timo Gerkmann[1], Simon Doclo[1]

[1]University of Oldenburg, Department of Medical Physics and Acoustics
and the Cluster of Excellence Hearing4All, Oldenburg, Germany
[2]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium
ante.jukic@uni-oldenburg.de

ABSTRACT

In many acoustic conditions the recorded speech signals may be severely affected by reverberation, leading to a reduced speech quality and intelligibility. In this paper we focus on a blind speech dereverberation method based on multi-channel linear prediction (MCLP) in the short-time Fourier transform domain, which is typically performed in each frequency bin independently without taking into account the spectral structure of the speech signal. Since it is widely accepted that a speech spectrogram can be well approximated with a low-rank matrix, e.g., using a spectral dictionary, in this paper we propose to incorporate a low-rank matrix approximation of the speech spectrogram into the MCLP-based speech dereverberation. The low-rank approximation is obtained using nonnegative matrix factorization with Itakura-Saito divergence. Experimental results for several measured acoustic systems show that incorporating a low-rank approximation improves the dereverberation performance in terms of instrumental speech quality measures.

Index Terms— speech dereverberation, low-rank approximation, speech enhancement, multi-channel linear prediction, nonnegative matrix factorization

1. INTRODUCTION

The microphone recordings of a speech signal captured in a room are typically affected by reverberation, resulting from the reflections of the sound against the walls and objects within the room. In severe cases, speech intelligibility and automatic speech recognition performance may be significantly reduced due to reverberation [1, 2]. Therefore, effective reverberation suppression is required in several speech communication applications, such as hands-free telephony and voice-controlled systems. Recently, several dereverberation methods have been proposed in the literature [3–9].

A blind speech dereverberation method based on multi-channel linear prediction (MCLP) in the short-time Fourier transform (STFT) domain was proposed in [3, 5]. In each frequency bin it uses an autoregressive model of reverberation, assuming that the late reverberation can be predicted from the previous frames of the observed reverberant signals. The estimation of the prediction coefficients is based on maximum-likelihood estimation assuming a time-varying Gaussian (TVG) model for the desired speech signal, resulting in an iterative optimization scheme. The original MCLP-based method has been extended to multiple-input multiple-output and time-varying systems [10, 11], general sparse models for the desired signal [12], and has been applied in joint dereverberation and denoising [13–15].

The MCLP-based dereverberation method [3, 5] and most of its extensions operate in each frequency bin independently. While this leads to a highly efficient parallel implementation, a significant drawback is that the spectral structure of the speech signal is not exploited. Hence, including some spectral structure in the algorithm could possibly be beneficial. For example, in [8, 13] an all-pole model was used to exploit the smoothness of the speech power spectra, and pre-trained speech log-spectral priors have been used in [16]. Alternatively, it is also widely accepted that spectrograms of audio signals, including speech, can be well modeled using a low-rank approximation [17–19]. In this paper, we propose to incorporate a low-rank power spectrogram approximation into the MCLP-based dereverberation method. The low-rank approximation is obtained through nonnegative matrix factorization (NMF) using Itakura-Saito (IS) divergence [20], which is a common choice for audio signals, and also fits well in the MCLP framework [16]. The NMF problem is solved using an unsupervised or supervised variant of NMF, with the latter employing a pre-trained spectral dictionary. The obtained results show that the incorporating the low-rank approximation improves the dereverberation performance of the MCLP method.

2. PROBLEM FORMULATION

We consider a single speech source captured by M microphones in an acoustic enclosure. Let $s(k,n)$ denote the clean speech signal in the STFT domain, with frequency bin index $k \in \{1,\dots,K\}$, and time frame index $n \in \{1,\dots,N\}$. The STFT coefficients of the observed reverberant signal $x_m(k,n)$ at the m-th microphone can be modeled as

$$x_m(k,n) = \sum_{l=0}^{L_h-1} h_m(k,l)s(k,n-l) + v_m(k,n), \qquad (1)$$

with $h_m(k,l)$ modeling the acoustic transfer function of length L_h between the speech source and m-th microphone, and $v_m(k,n)$ representing the additive noise and modeling error. As in [5], by assuming $v_m(k,n) = 0$, the convolutive model in (1) can be simplified and the signal at the reference microphone (e.g., $m = 1$) can be written as

$$x_1(k,n) = d(k,n) + \sum_{m=1}^{M} \sum_{l=0}^{L_g-1} g_m(k,l)x_m(k,n-\tau-l), \quad (2)$$

This research was supported by the Marie Curie Initial Training Network DREAMS (Grant agreement no. ITN-GA-2012-316969), and in part by the Research Foundation Flanders (FWO-Vlaanderen) and the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG).

where L_g is the length of the prediction filter g_m for each channel. The first term $d(k,n) = \sum_{l=0}^{\tau-1} h_1(k,l)s(k,n-l)$ in (2) represents the desired speech signal at the reference microphone and consists of the direct path signal and early reflections (determined by the prediction delay τ). The second term in (2) models the late reverberation that is predicted from the previous observations on all M microphones using the prediction coefficients $g_m(k,l)$ [5]. The MCLP model in (2) can be written in vector form as

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \sum_{m=1}^{M} \tilde{\mathbf{X}}_{\tau,m}(k)\mathbf{g}_m(k), \tag{3}$$

with

$$\begin{aligned}\mathbf{d}(k) &= [d(k,1),\ldots,d(k,N)]^T,\\ \mathbf{x}_m(k) &= [x_m(k,1),\ldots,x_m(k,N)]^T,\\ \mathbf{g}_m(k) &= [g_m(k,0),\ldots,g_m(k,L_g-1)]^T,\end{aligned}$$

and $\tilde{\mathbf{X}}_{\tau,m}(k) \in \mathbb{C}^{N\times L_g}$ denoting a convolution matrix constructed using $\mathbf{x}_m(k)$ delayed for τ frames. Furthermore, by stacking the matrices $\tilde{\mathbf{X}}_{\tau,m}(k)$ and vectors $\mathbf{g}_m(k)$ as

$$\tilde{\mathbf{X}}_{\tau}(k) = \left[\tilde{\mathbf{X}}_{\tau,1}(k),\ldots,\tilde{\mathbf{X}}_{\tau,M}(k)\right], \tag{4}$$

$$\mathbf{g}(k) = \left[\mathbf{g}_1^T(k),\ldots,\mathbf{g}_M^T(k)\right]^T, \tag{5}$$

the MCLP model can be rewritten as

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \tilde{\mathbf{X}}_{\tau}(k)\mathbf{g}(k). \tag{6}$$

Speech dereverberation using the MCLP model in (2) can be achieved by estimating the prediction coefficients $\mathbf{g}(k)$ for each frequency bin k. Using the estimated prediction coefficients $\hat{\mathbf{g}}(k)$, the desired (dereverberated) speech signal is then estimated as

$$\boxed{\hat{\mathbf{d}}(k) = \mathbf{x}_1(k) - \tilde{\mathbf{X}}_{\tau}(k)\hat{\mathbf{g}}(k).} \tag{7}$$

3. MCLP-BASED SPEECH DEREVERBERATION

In [3, 5] speech dereverberation based on MCLP has been formulated assuming a TVG model for the desired speech signal. In the TVG model the desired signal in each time-frequency bin is modeled as a zero-mean complex-valued Gaussian variable, with time- and frequency-dependent variance. The probability density function for the desired signal $d(k,n)$ can then be written as

$$p\left(d(k,n)\right) = \frac{1}{\pi\lambda(k,n)} e^{-\frac{|d(k,n)|^2}{\lambda(k,n)}}, \tag{8}$$

where the variance $\lambda(k,n)$ is an unknown parameter that needs to be estimated. Note that the TVG model does not include any dependency across frequency or time, i.e., it is assumed that the STFT coefficients $d(k,n)$ in different time-frequency bins are independent random variables. The likelihood function for the k-th frequency bin can be written as

$$\mathcal{L}\left(\mathbf{g}(k),\boldsymbol{\lambda}(k)\right) = p\left(\mathbf{d}(k)\right) = \prod_{n=1}^{N} p\left(d(k,n)\right), \tag{9}$$

with $\boldsymbol{\lambda}(k) = [\lambda(k,1),\ldots,\lambda(k,N)]^T$. The unknown prediction coefficients $\mathbf{g}(k)$ and variances $\boldsymbol{\lambda}(k)$ can be estimated by maximizing the likelihood function in (9), or equivalently by minimizing the negative log-likelihood as

$$\min_{\mathbf{g}(k),\boldsymbol{\lambda}(k)} \mathbf{d}^H(k)\mathcal{D}_{\boldsymbol{\lambda}(k)}^{-1}\mathbf{d}(k) + \sum_{n=1}^{N}\log\lambda(k,n), \tag{10}$$

where $\mathcal{D}_{\boldsymbol{\lambda}(k)} = \mathrm{diag}\left(\boldsymbol{\lambda}(k)\right)$ is a diagonal matrix constructed using the vector $\boldsymbol{\lambda}(k)$, and constant terms in the negative log-likelihood have been omitted. In [5] it has been proposed to solve the optimization problem in (10) using an alternating optimization procedure. In the first step, the cost function in (10) is minimized with respect to the prediction coefficients $\mathbf{g}(k)$, assuming that an estimate for the variances $\hat{\boldsymbol{\lambda}}(k)$ is available. In this case, the following quadratic optimization problem is obtained

$$\hat{\mathbf{g}}(k) = \arg\min_{\mathbf{g}(k)} \mathbf{d}^H(k)\mathcal{D}_{\hat{\boldsymbol{\lambda}}(k)}^{-1}\mathbf{d}(k). \tag{11}$$

By combining (6) and (11), the prediction coefficients can be estimated as

$$\boxed{\hat{\mathbf{g}}(k) = \left(\tilde{\mathbf{X}}_{\tau}^H(k)\mathcal{D}_{\hat{\boldsymbol{\lambda}}(k)}^{-1}\tilde{\mathbf{X}}_{\tau}(k)\right)^{-1}\tilde{\mathbf{X}}_{\tau}^H(k)\mathcal{D}_{\hat{\boldsymbol{\lambda}}(k)}^{-1}\mathbf{x}_1(k).} \tag{12}$$

In the second step, the cost function in (10) is minimized with respect to the variances $\boldsymbol{\lambda}(k)$, assuming that an estimate for the prediction coefficients $\hat{\mathbf{g}}(k)$ is available. In this case, an estimate of the desired speech signal $\hat{\mathbf{d}}(k)$ is first calculated using (7), and the variance $\lambda(k,n)$ is estimated as

$$\hat{\lambda}(k,n) = \arg\min_{\lambda(k,n)>0} \frac{|\hat{d}(k,n)|^2}{\lambda(k,n)} + \log\lambda(k,n), \tag{13}$$

with the solution given as $\hat{\lambda}(k,n) = |\hat{d}(k,n)|^2$, or in short

$$\boxed{\hat{\boldsymbol{\lambda}}(k) = |\hat{\mathbf{d}}(k)|^2,} \tag{14}$$

where the absolute value and the power are applied element-wise. The alternating procedure consisting of (12) and (14) is iterated for a fixed number of iterations or until convergence. At the beginning of the alternating procedure, the variance estimates are initialized as $\hat{\boldsymbol{\lambda}}(k) = |\mathbf{x}_1(k)|^2$ [5].

4. LOW-RANK POWER SPECTROGRAM APPROXIMATION

In several speech enhancement methods a low-rank model for the speech signal has been successfully applied. For example, in NMF-based speech denoising [19, 21] and source separation methods [18] the magnitude or power spectrograms have been modeled as a low-rank matrix, typically using a dictionary with a small number of spectral vectors. In the previously presented MCLP-based dereverberation procedure, the estimated prediction coefficients $\hat{\mathbf{g}}(k)$ in each frequency bin depend on the estimated variances $\hat{\boldsymbol{\lambda}}(k)$. Instead of estimating these variances independently in each time-frequency bin as the power spectrogram of the estimated desired speech signal $\hat{\mathbf{d}}(k)$, cf. (14), in this section we propose to estimate the variances using a low-rank approximation of the power spectrogram.

We define the matrix $\hat{\mathbf{D}} = \{\hat{d}(k,n)\} \in \mathbb{C}^{K\times N}$ containing all STFT coefficients of the estimated desired speech signal, i.e., the k-th row of $\hat{\mathbf{D}}$ is equal to $\hat{\mathbf{d}}^T(k)$. Then the power spectrogram of

the desired speech signal is given as $|\hat{\mathbf{D}}|^2$, where the absolute value and the power are applied element-wise. Similarly, we define the nonnegative matrix $\mathbf{\Lambda} \in \mathbb{R}_{0+}^{K \times N}$, with its k-th row equal to $\boldsymbol{\lambda}^T(k)$. Assuming that the power spectrogram $|\hat{\mathbf{D}}|^2$ can be modeled as a rank-R matrix with $R < \min\{K, N\}$, we employ NMF to find its nonnegative low-rank approximation in the form $\hat{\mathbf{\Lambda}} = \hat{\mathbf{W}}\hat{\mathbf{H}}$, where $\hat{\mathbf{W}} \in \mathbb{R}_{0+}^{K \times R}$ and $\hat{\mathbf{H}} \in \mathbb{R}_{0+}^{R \times N}$ are nonnegative matrices. The matrix $\hat{\mathbf{W}}$ can be interpreted as a spectral dictionary containing R spectral profiles, while matrix $\hat{\mathbf{H}}$ contains activation coefficients for the dictionary elements across the time frames. The NMF problem is typically formulated as

$$\min_{\mathbf{W},\mathbf{H}} J\left(|\hat{\mathbf{D}}|^2, \mathbf{WH}\right), \text{ s.t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{15}$$

where the cost function is a divergence J that quantifies the discrepancy between two matrices. Selection of the divergence J in general depends on the considered application. Most commonly used divergences for NMF are Euclidean distance, Kullback-Leibler (KL) divergence, and Itakura-Saito (IS) divergence, which are all special cases of the β-divergence [20, 22]. Here we employ the IS divergence given as

$$J_{IS}\left(|\hat{\mathbf{D}}|^2, \mathbf{\Lambda}\right) = \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{|\hat{d}(k,n)|^2}{\lambda(k,n)} - \log\frac{|\hat{d}(k,n)|^2}{\lambda(k,n)} - 1. \tag{16}$$

The motivation for selecting the IS divergence becomes clear by comparing a single term in the sum in (16) with the cost function in (13), and observing that they only differ up to a constant independent of $\lambda(k,n)$ [16]. Also, as opposed to Euclidean and KL divergence, the IS divergence is scale-invariant and thus better suited for modeling data with a large dynamic range, such as audio signals [17]. From a probabilistic perspective, NMF with the IS divergence applied on the power spectrogram corresponds to a maximum-likelihood estimation in a Gaussian composite model of the STFT [17].

We will consider two variants of the IS divergence-based low-rank approximation. The first variant uses unsupervised NMF, by solving the following optimization problem

$$\{\hat{\mathbf{W}}, \hat{\mathbf{H}}\} = \arg\min_{\mathbf{W},\mathbf{H}} J_{IS}\left(|\hat{\mathbf{D}}|^2, \mathbf{WH}\right), \text{ s.t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0. \tag{17}$$

In this case both the spectral dictionary and the activation coefficients are estimated by minimizing the IS divergence between the known matrix $|\hat{\mathbf{D}}|^2$ and its low-rank approximation. The variances are then estimated as $\hat{\mathbf{\Lambda}} = \hat{\mathbf{W}}\hat{\mathbf{H}}$. The second variant uses supervised NMF, by solving the following optimization problem

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} J_{IS}\left(|\hat{\mathbf{D}}|^2, \mathbf{W}_{\text{trained}}\mathbf{H}\right), \text{ s.t. } \mathbf{H} \geq 0. \tag{18}$$

In this case the matrix $\mathbf{W}_{\text{trained}}$ is fixed and equal to a given pre-trained spectral dictionary and only the matrix $\mathbf{H}$ is estimated. The dictionary is learned on the power spectrograms of training samples, for example using a database of clean speech signals [19]. The variances are then estimated as $\hat{\mathbf{\Lambda}} = \mathbf{W}_{\text{trained}}\hat{\mathbf{H}}$.

The complete dereverberation scheme is summarized in Algorithm 1. The method starts with initializing the variances using a low-rank approximation of the power spectrogram of the observed signal at the reference microphone $|\mathbf{X}_1|^2$. The low-rank approximation step corresponds to the supervised or unsupervised NMF problem (18) or (17), depending on whether the pre-trained dictionary is provided or not. In each iteration, the prediction coefficients for the MCLP model are updated, followed by a NMF-based low-rank approximation for the variance update. Note that the original MCLP method is obtained by omitting the low-rank approximation and using $\hat{\mathbf{\Lambda}} \leftarrow |\hat{\mathbf{D}}|^2$. The presented iterative scheme is repeated for a predetermined number of iterations $i_{\max}$. Note that we are not enforcing the estimated desired speech signal to have a low-rank spectrogram, since it is still calculated using MCLP as in (7). In terms of computational complexity, the most expensive step of the algorithm is computation of NMF. In the experimental section, for both supervised and unsupervised NMF, we minimize the IS divergence using an iterative procedure with multiplicative updates [20]. This is a standard approach to NMF with the IS divergence [22], while faster [23] and online [24] algorithms have been proposed recently.

Algorithm 1 Outline of the proposed algorithm combining MCLP and low-rank power spectrogram approximation.

parameters: L_g and τ in (6), rank R in (17)/(18)
input: $\mathbf{X}_m = \{x_m(k,n)\}, \forall m$
initialization: $\hat{\mathbf{\Lambda}} = \text{low\_rank\_approx}\left(|\mathbf{X}_1|^2\right)$
for $i = 1, \ldots, i_{\max}$ **do**
  for $k = 1, \ldots, K$ **do**
    $\hat{\mathbf{g}}(k) \leftarrow$ calculate using (12)
    $\hat{\mathbf{d}}(k) \leftarrow \mathbf{x}_1(k) - \tilde{\mathbf{X}}_\tau(k)\hat{\mathbf{g}}(k)$
  end for
  $\hat{\mathbf{\Lambda}} \leftarrow \text{low\_rank\_approx}\left(|\hat{\mathbf{D}}|^2\right)$
end for

5. EXPERIMENTS

In this section the dereverberation performance of the original MCLP method [5] and the proposed MCLP methods using a low-rank power spectrogram approximation (both the unsupervised variant, labeled MCLP+NMF, and the supervised variant with a pre-trained dictionary, labeled MCLP+NMF+dict) will be evaluated.

We have considered two different acoustic scenarios with a single speech source. The first acoustic system (AC_1) consists of $M = 2$ measured RIRs from the MARDY database [25] in a room with a reverberation time of $RT_{60} \approx 450$ ms. The distance between the source and the microphones is approximately 3 m, and the direct-to-reverberant ratio (DRR) is 3.3 dB at the reference microphone. The second acoustic system (AC_2) consists of $M = 4$ measured RIRs in a room with a reverberation time of $RT_{60} \approx 750$ ms. The distance between the source and the microphones is approximately 2.3 m, and the DRR is -3.6 dB at the reference microphone.

In all experiments, the sampling frequency was $f_s = 16$ kHz, and the STFT was calculated using a Hann window with a frame length of 64 ms and a frame shift of 16 ms. We set length of the prediction filters for each channel to $L_g = 14$ for $M = 2$, and $L_g = 8$ for $M = 4$. The prediction delay is set to $\tau = 2$ frames, and the maximum number of iterations is set to $i_{\max} = 10$. For both unsupervised and supervised NMF variants we use the multiplicative updates as in [20], and set the tolerance on the relative change of the IS divergence to 10^{-4} (typically around hundred iterations were performed for NMF). The dictionary $\mathbf{W}_{\text{trained}}$ with R columns for the supervised NMF variant was learned on a subset of the TIMIT database [26] with the training matrix composed of the power spectrograms of 190 utterances from 38 speakers. The dictionary was obtained by applying unsupervised NMF with rank R on the train-

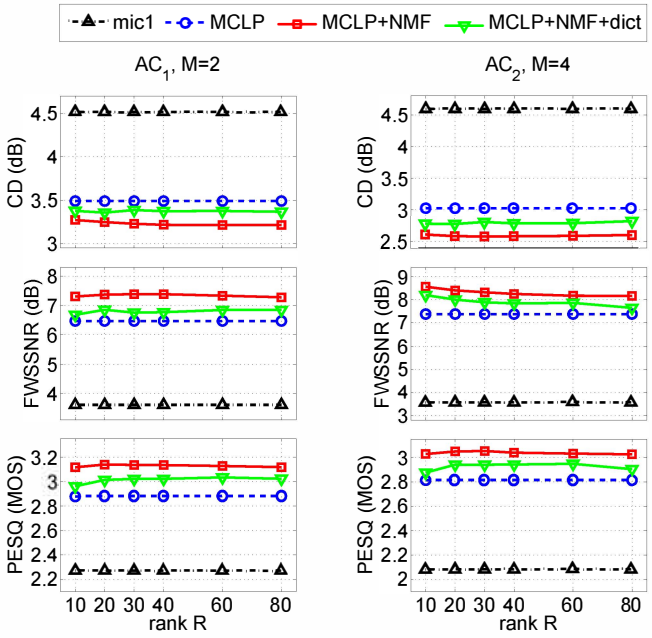
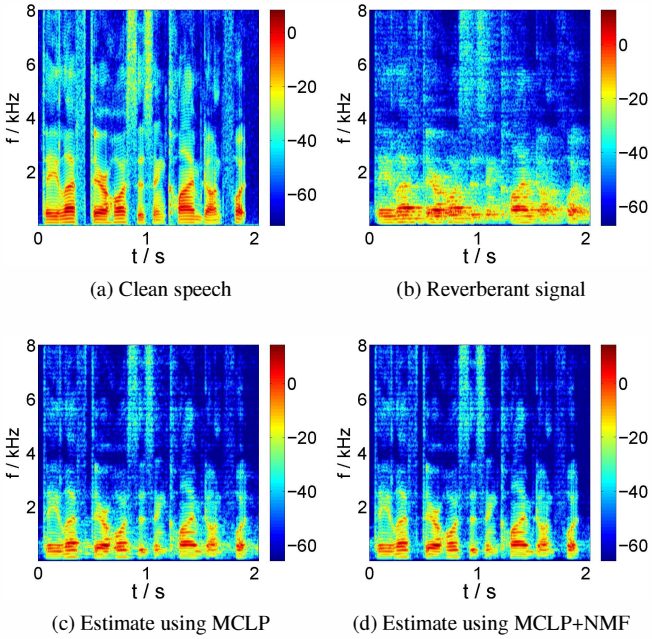


Fig. 1: Evaluated speech quality measures for AC_1 (left) and AC_2 (right) vs. rank R of the power spectrogram approximation.



(a) Clean speech (b) Reverberant signal

(c) Estimate using MCLP (d) Estimate using MCLP+NMF

Fig. 2: The spectrograms of the clean speech, reverberant observation, MCLP estimate, and MCLP+NMF estimate (rank $R = 40$) for AC_2.

ing matrix, and each column of the obtained dictionary is further normalized so that it sums to one.

The considered speech dereverberation algorithms were tested on a set of 10 utterances from the TIMIT database (speakers not contained in the training set), with an average length of approximately 4.3 s. The reverberant observations were obtained by convolving the utterances with the respective RIRs. The dereverberation performance is evaluated in terms of several instrumental speech quality measures, i.e., cepstral distance (CD), perceptual evaluation of speech quality (PESQ), and frequency-weighted segmental signal-to-noise ratio (SNR) [27]. The measures were evaluated with the clean speech signal as the reference and then averaged over all utterances. The results obtained for AC_1 and AC_2 with different ranks R of the approximation are presented in Figure 1. It can be observed that for all experiments the dereverberation performance is improved by including the low-rank power spectrogram approximation. Furthermore, the unsupervised NMF variant results in the best overall performance, with the performance gain relatively stable across a wide range of ranks R. The supervised NMF variant also results in an improved performance when compared to the original MCLP method but is outperformed by unsupervised NMF variant. This can be explained by the fact that unsupervised NMF results in a better low-rank approximation (in terms of divergence) than supervised NMF, due to the inclusion of a trained speaker-independent dictionary in the latter. However, it is possible that with a speaker-dependent (or less general) dictionary the performance of the supervised NMF variant could be improved. Another possible reason is that the dictionary is trained on clean speech, while the estimated desired speech also includes early reflections. This mismatch could possibly be avoided by training the dictionary on clean speech samples convolved with the early part of the RIRs, but this would make the dictionary learning dependent on the RIRs in the training set.

Figure 2 depicts spectrograms of the clean speech, the reverberant observation, and the estimated desired speech signals obtained using the original MCLP method and the MCLP method combined with unsupervised NMF. It can be observed that MCLP combined with unsupervised NMF suppresses more reverberation than the original MCLP method, with the improved performance especially visible in the low-frequencies and during the speech pauses.

6. CONCLUSIONS

In this paper we have presented a method to incorporate a low-rank structure of the speech power spectrogram into MCLP-based dereverberation. The proposed method uses NMF based on Itakura-Saito divergence for the low-rank approximation, either in an unsupervised or supervised way. The experimental results demonstrate that the proposed method improves the dereverberation performance when compared to the original MCLP method.

7. REFERENCES

[1] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 75–95, Aug. 1998.

[2] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, July 2006.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, May 2008, pp. 85–88.

[4] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, June 2009.

[5] 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, Sept. 2010.

[6] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sept. 2013.

[7] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, Aug 2014.

[8] B. Schwartz, S. Gannot, and E. Habets, "LPC-based speech dereverberation using Kalman-EM algorithm," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Antibes - Juan les Pins, France, Sept. 2014.

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

[10] 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, Dec 2012.

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

[12] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 23–26.

[13] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb 2009.

[14] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 7447–7451.

[15] M. Delcroix et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[16] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on Itakura-Saito distance minimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, May 2012, pp. 245–248.

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

[18] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 3, pp. 550–563, March 2010.

[19] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct 2013.

[20] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation (ICA)*, Berlin, 2006, pp. 32–39.

[21] K. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *Proc. INTERSPEECH*, pp. 411–414.

[22] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sept. 2011.

[23] D. L. Sun and C. Févotte, "Alternating direction method of multipliers for non-negative matrix fatorization with the beta-divergence," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 6201–6205.

[24] A. Lefevre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence," in *Proc. IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2011, pp. 313–316.

[25] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Paris, France, Sept. 2008.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, *TIMIT acoustic-phonetic continuous speech corpus*, Philadelphia: Linguistic Data Consortium, 1993.

[27] K. Kinoshita, M. Delcroix, T. Yoshioka, E. Habets, R. Haeb-Umbach, V. Leutnat, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, USA, Oct. 2013.