SPEECH DEREVERBERATION WITH MULTI-CHANNEL LINEAR PREDICTION AND SPARSE PRIORS FOR THE DESIRED SIGNAL

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

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

ABSTRACT

The quality of recorded speech signals can be substantially affected by room reverberation. In this paper we focus on a blind method for speech dereverberation based on the multi-channel linear prediction model in the short-time Fourier domain, where the parameters of the model are estimated using a maximum-likelihood procedure. Contrary to the conventional approach, we propose to model the desired speech signal using a general sparse prior that can be represented as a maximization over scaled complex Gaussians. Experimental evaluation, employing a parametric complex generalized Gaussian prior for the desired speech signal, shows that instrumentally predicted speech quality can be improved compared to the conventional approach.

***Index Terms*—** Dereverberation, speech enhancement, model-based signal processing, sparse priors

1. INTRODUCTION

Speech signals captured in an enclosed space with microphones placed at a distance are typically corrupted by reverberation, caused by reflections against the walls and objects within the enclosure. Although moderate reverberation can have beneficial effects, in severe cases it can lead to a significant decrease of speech intelligibility and automatic speech recognition performance [1]. Several speech communication applications, such as hands-free telephony, teleconferencing and voice-controlled systems, require effective solutions for reverberation suppression. Various dereverberation techniques have been proposed in the literature. One class of methods is based on blind identification of the room impulse responses (RIRs) between the source and the microphone array, followed by multichannel equalization [2]. This strategy could, in theory, result in a perfect dereverberation. While several robust equalization techniques were proposed [3], their performance is still affected by RIR estimation errors, and accurate blind channel identification remains an issue. More robust speech dereverberation techniques are based on spectral enhancement, but with a trade-off between speech distortion and reverberation suppression [4]. Recently, several blind speech dereverberation techniques were proposed that do not require any knowledge about the acoustical properties of the enclosure [1, 5, 6].

A blind dereverberation method based on multi-channel linear prediction (MCLP) was proposed in [5], with an efficient implementation in the short-time Fourier domain (STFT). The method is based on an autoregressive model of the reverberation process, assuming that the reverberant components can be predicted from the previous samples. An additional delay is introduced in the MCLP model to preserve the short-time correlation of the desired signal and suppress only late reverberation [5]. In this conventional approach, the complex-valued STFT coefficients of the desired speech signal are modeled using a time-varying Gaussian (TVG) model that assumes that the coefficients can be modeled locally (i.e., at each time-frequency bin) with a complex Gaussian distribution with unknown variance. The unknown parameters of the MCLP model and the variances of the TVG model are then estimated by an iterative maximum-likelihood scheme.

In this paper we propose to use a general circular sparse prior for the STFT coefficients of the desired speech signal. The prior is represented as a maximization over scaled Gaussians [7] that can be interpreted as a TVG model with a hyperprior on the unknown variance. The proposed algorithm is derived in the general case, for a wide range of possible sparse priors. However, in the experiments we use a parametric family of complex generalized Gaussian priors [8]. The results show that the proposed approach outperforms the conventional approach.

The paper is organized as follows. In Section 2 we introduce the notation and formulate the problem of speech dereverberation using the MCLP model. Section 3 contains a brief overview of the conventional approach, while the proposed approach is presented in Section 4 with experimental results in Section 5.

2. PROBLEM FORMULATION

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

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

where $h_m(l,k)$ models the acoustic transfer function of length L_h between the speech source and the m-th microphone, and $(.)^*$ denotes the complex conjugate operator. The additive term $e_m(l,k)$ jointly represents modeling errors and the additive noise signal. As in [5], by assuming $e_m(n,k) \equiv 0$ the convolutive model in (1) can

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).

be expressed in the multi-channel linear prediction form as

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

with $d(n,k) = \sum_{l=0}^{D-1} h_1^*(l,k) s(n-l,k)$ being the desired speech signal (at the reference microphone $m = 1$) consisting of the anechoic speech signal and early reflections determined by the prediction delay D, and $g_m(l,k)$ denoting the regression coefficients [5]. The MCLP model in (2) can be written in a more compact form as

$$\boxed{x_1(n,k) = d(n,k) + \bar{\mathbf{g}}(k)^H \bar{\mathbf{x}}(n-D,k),} \quad (3)$$

where $\bar{\mathbf{g}}(k) \in \mathbb{C}^{ML_g}$ is a multi-channel regression vector, and $\bar{\mathbf{x}}(n,k)$ consists of previous samples in each of the M channels, i.e.,

$$\bar{\mathbf{x}}(n,k) = \left[\bar{\mathbf{x}}_1(n,k)^T, \dots, \bar{\mathbf{x}}_M(n,k)^T\right]^T, \quad (4)$$

with $\bar{\mathbf{x}}_m(n,k) = [x_m(n,k), \dots, x_m(n-L_g+1,k)]^T$.

The problem of speech dereverberation can now be formulated as a blind estimation of the desired speech signal $d(n,k)$ from the reverberant observations $x_m(n,k), \forall m,n,k$. Using (3), the desired signal can be estimated as

$$\hat{d}(n,k) = x_1(n,k) - \hat{\bar{\mathbf{g}}}(k)^H \bar{\mathbf{x}}(n-D,k), \quad (5)$$

with $\hat{(.)}$ denoting an estimated value. Therefore, dereverberation can be performed by estimating the regression vectors $\hat{\bar{\mathbf{g}}}(k)$, and applying (5). Note that in the following we will work in each frequency bin independently, so the index k will often be omitted for notational convenience.

3. CONVENTIONAL APPROACH

Several speech dereverberation methods were proposed using the time-varying Gaussian model for the desired signal [5, 6]. More specifically, the desired signal in each time-frequency bin is modeled as a zero-mean random variable by a circular complex Gaussian distribution with unknown and time-varying variance. The probability density function for the desired signal $d(n,k)$ can then be written as

$$\boxed{\mathcal{N}_{\mathbb{C}}\big(d(n,k); 0, \lambda(n,k)\big) = \frac{1}{\pi\lambda(n,k)} \mathrm{e}^{-\frac{|d(n,k)|^2}{\lambda(n,k)}},} \quad (6)$$

where $\lambda(n,k)$ is considered to be an unknown parameter that needs to be estimated. Since the TVG model does not include any dependency across frequencies, the index k can be omitted and the likelihood function for a single frequency bin can be written as [5]

$$\mathcal{L}(\bar{\mathbf{g}}, \boldsymbol{\lambda}) = \prod_{n=1}^{N} \mathcal{N}_{\mathbb{C}}\big(d(n); 0, \lambda(n)\big), \quad (7)$$

with $\boldsymbol{\lambda} = [\lambda(1), \dots, \lambda(N)]^T$. The regression vector $\bar{\mathbf{g}}$ is estimated by maximizing the likelihood with respect to the unknown parameters (variances and the regression vector), i.e., by solving the following optimization problem

$$\boxed{\min_{\boldsymbol{\lambda}>0, \bar{\mathbf{g}}} \sum_{n=1}^{N} \frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n).} \quad (8)$$

Since the joint minimization of (8) with respect to $\bar{\mathbf{g}}$ and $\boldsymbol{\lambda}$ can not be performed analytically, it was proposed in [5] to use an alternating optimization procedure.

Estimation of $\bar{\mathbf{g}}$: In the first step, the cost function in (8) is minimized with respect to $\bar{\mathbf{g}}$. Assuming that the variances $\boldsymbol{\lambda}$ are fixed, a least squares problem is obtained, i.e.,

$$\min_{\bar{\mathbf{g}}} \sum_{n=1}^{N} \frac{|x_1(n) - \bar{\mathbf{g}}^H \bar{\mathbf{x}}(n-D)|^2}{\lambda(n)}, \quad (9)$$

with the optimal regression vector $\bar{\mathbf{g}}$ given as

$$\bar{\mathbf{g}} = \mathbf{A}^{-1}\mathbf{b}, \quad (10)$$

where

$$\mathbf{A} = \sum_{n=1}^{N} \frac{\bar{\mathbf{x}}(n-D)\bar{\mathbf{x}}^H(n-D)}{\lambda(n)}, \ \mathbf{b} = \sum_{n=1}^{N} \frac{\bar{\mathbf{x}}(n-D) x_1^*(n)}{\lambda(n)} \quad (11)$$

Estimation of $\boldsymbol{\lambda}$: In the second step, the cost function in (8) is minimized with respect to $\boldsymbol{\lambda}$. Assuming now that the regression vector $\bar{\mathbf{g}}$ is fixed the optimal variances can be calculated as

$$\boxed{\lambda(n) = |d(n)|^2.} \quad (12)$$

This alternating procedure is repeated until a convergence criterion is satisfied or a maximum number of iterations is exceeded. Additionally, to prevent division by zero a small positive constant ε is included as a lower bound for the estimated variance. The presented approach is often referred to as the weighted prediction error (WPE) method [5].

4. PROPOSED APPROACH

It is widely accepted that the STFT coefficients of speech signals can be well modeled using sparse priors, both locally [9–11] as well as globally [12]. Although the real and imaginary parts of the complex-valued coefficients are often assumed to be independent to simplify computations, it was observed that the distribution of the complex-valued speech coefficients is actually approximately circular [13,14]. In the proposed approach we therefore use the MCLP model (3) to model the reverberation process, with a sparse circular prior for the desired speech signal. The proposed prior can be interpreted as a TVG model with an additional hyperprior for the variance. Similar modification could be used with different local models (e.g., locally Laplacian model in [15]).

4.1. Representation of sparse priors

Intuitively, a prior is considered to be sparse when it is super-Gaussian, i.e., it exhibits a higher peak at the origin and heavier tails than the corresponding Gaussian prior. Here we consider a circular sparse prior for a complex-valued random variable Z that can be represented as

$$p(z) = \mathrm{e}^{-f(|z|)}. \quad (13)$$

In general, $p(z)$ can represent a proper sparse prior (i.e., a probability density), or an *improper* (non-integrable) sparse prior. Formally, it can be shown that when $f'(t)/t$ is decreasing on $(0,\infty)$, the prior will be super-Gaussian, i.e., sparse [7]. In this case, $p(z)$ can be conveniently represented as a maximization over scaled Gaussians with different variances, i.e.,

$$p(z) = \max_{\lambda>0} \mathcal{N}_{\mathbb{C}}(z; 0, \lambda)\psi(\lambda), \quad (14)$$

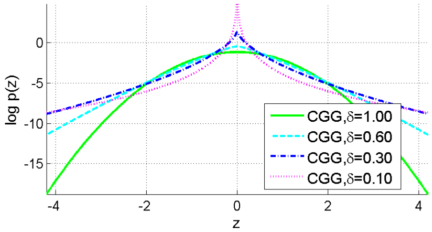

Fig. 1. Logarithm of the CGG prior (15) for different values of the shape parameter δ and variance fixed to 1.

where ψ is a scaling function that can be interpreted as a hyperprior on the variance λ [7]. This representation is often referred to as the convex type due to its roots in convex analysis, as opposed to the integral type of representations, such as Gaussian scale mixtures [7]. The scaling function ψ in (14) is related to f in (13), but the scaling function is often not required explicitly in practical algorithms [7].

An example of a parametric circular super-Gaussian prior is the complex generalized Gaussian (CGG) given as [8]

$$p(z) = \frac{\delta}{\pi\gamma\Gamma(1/\delta)} \mathrm{e}^{-\left(\frac{|z|^2}{\gamma}\right)^{\delta}}, \tag{15}$$

with the scale parameter $\gamma > 0$, the shape parameter $\delta \in (0, 1)$, and Γ denoting the Gamma function. The circular Gaussian distribution is obtained by setting $\delta = 1$, while smaller values of the shape parameter result in more sparse priors, i.e., a stronger peak at zero and heavier tails. This can also be seen from the plot of $\log p(z)$ in Figure 1. Since the CGG prior can be written in the form (13) with f given as

$$f(t) = \left(\frac{t^2}{\gamma}\right)^{\delta} - \log \frac{\delta}{\pi\gamma\Gamma(1/\delta)}, \tag{16}$$

it can be represented using a convex representation in the form (14).

4.2. Speech dereverberation using general sparse priors

We now propose to model the coefficients of the desired speech signal using a circular sparse prior $p\big(d(n)\big) = \mathrm{e}^{-f(|d(n)|)}$ with a convex representation given as

$$\boxed{p\big(d(n)\big) = \max_{\lambda(n)>0} \mathcal{N}_{\mathbb{C}}\big(d(n); 0, \lambda(n)\big)\psi\big(\lambda(n)\big).} \tag{17}$$

This can be interpreted as a generalization of the TVG model, with an additional hyperprior for the variance $\lambda(n)$. Similarly as in the conventional approach, the regression vector $\bar{\mathbf{g}}$ can be estimated by maximizing the likelihood, i.e., by solving the following optimization problem

$$\boxed{\min_{\boldsymbol{\lambda}>0,\bar{\mathbf{g}}} \sum_{n=1}^{N} \frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n) - \log \psi\big(\lambda(n)\big),} \tag{18}$$

with $d(n)$ depending on $\bar{\mathbf{g}}$ through (5). The likelihood can again be maximized using an alternating optimization procedure.

Estimation of $\bar{\mathbf{g}}$: Assuming that the variances $\boldsymbol{\lambda}$ are fixed, the same least-squares problem is obtained as in the conventional approach, with the solution given by (10).

Estimation of $\boldsymbol{\lambda}$: Assuming that the regression vector $\bar{\mathbf{g}}$ is fixed, each $\lambda(n)$ can be obtained by solving the following problem

$$\min_{\lambda(n)>0} \frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n) - \log \psi\big(\lambda(n)\big). \tag{19}$$

By exploiting the relation between the scaling function ψ and f (for details we refer to a similar derivation in [7]) the optimal value of $\lambda(n)$ can be expressed as

$$\lambda(n) = \frac{2|d(n)|}{f'\big(|d(n)|\big)}, \tag{20}$$

for a general sparse prior in (13). Note that although the optimization problem in (19) includes ψ, the optimal $\lambda(n)$ for this subproblem depends only on f, so the scaling function ψ does not need to be given explicitly.

In the case of a CGG prior for the desired signal, the optimal value of $\lambda(n)$ can be written using (16) and (20) as

$$\lambda(n) = \frac{\gamma^{\delta}}{\delta} |d(n)|^{2(1-\delta)}. \tag{21}$$

This expression depends on the shape and the scaling parameters of the CGG prior in (15). However, since the estimation of $\bar{\mathbf{g}}$ using (10) and (11) is invariant to a scaling of $\boldsymbol{\lambda}$, the update in (21) can be simplified to

$$\boxed{\lambda(n) \leftarrow |d(n)|^{2(1-\delta)},} \tag{22}$$

that depends only on the shape parameter $\delta \in (0, 1)$ of the CGG prior. An outline of the proposed approach with a CGG prior is given in Table 1. Note that the variance update (12) in the conventional approach corresponds to setting $\delta = 0$ in the obtained update (22).

By comparing the optimization problem (8) with the proposed approach (18) it can be seen that the conventional approach is obtained by setting $\psi(\lambda) = \text{const}$ in the proposed approach. In this case the prior for the desired signal, as interpreted in the proposed framework with (14), is expressed as

$$p\big(d(n)\big) \propto \max_{\lambda(n)>0} \mathcal{N}_{\mathbb{C}}\big(d(n); 0, \lambda(n)\big) = \frac{1}{\pi \mathrm{e} |d(n)|^2}. \tag{23}$$

Note that this is an *improper* prior on the desired signal since it is not integrable. Also, it strongly favors values of the desired signal that are close to the origin, i.e., it is a strong sparse prior for the desired signal. This interpretation also highlights the role of sparsity in the conventional approach although it was derived using a TVG model.

```
parameters: $L_g$ and $D$ in (2), $\delta$ in (22)
input: $x_m(n,k), \forall n, m, k$
initialization: $\hat{\lambda}(n,k) \leftarrow |x_1(n,k)|^{2(1-\delta)}$
for all $k$ do
    repeat
        $\mathbf{A}(k), \mathbf{b}(k) \leftarrow$ calculate using (11)
        $\hat{\bar{\mathbf{g}}}(k) \leftarrow \mathbf{A}(k)^{-1}\mathbf{b}(k)$
        $\hat{d}(n,k) \leftarrow x_1(n,k) - \hat{\bar{\mathbf{g}}}(k)^H \bar{\mathbf{x}}(n-D,k)$
        $\hat{\lambda}(n,k) \leftarrow \max\{|\hat{d}(n,k)|^{2(1-\delta)}, \varepsilon(k)\}$
    until condition satisfied
end for
```

Table 1. Outline of the proposed approach with a CGG prior. The conventional approach [5] is obtained by setting $\delta = 0$.

5. EXPERIMENTS

To evaluate the performance of the proposed approach with CGG priors, we performed an experiment using sound samples of 10 different speakers (5 male and 5 female), where the average length of the sound samples was 9.5 s and the sampling frequency was $f_s = 16$ kHz. The reverberant observations were generated by convolving each utterance with a set of measured room impulse responses. We used a setup with $M = 2$ microphones in a room with $RT_{60} \approx 750$ ms. In the experiment the STFT was calculated using a 64 ms Hamming window with 75% overlap. The parameters for the algorithm outlined in Table 1 were set as follows: the order of the regression vectors $L_g = 23$, the prediction delay $D = 3$, and $\varepsilon(k) = 10^{-8}$. The dereverberation performance was evaluated in terms of cepstral distance (CD), perceptual evaluation of speech quality (PESQ), frequency-weighted segmental signal-to-noise ratio (FWSSNR), and speech-to-reverberation modulation energy ratio (SRMR) [16]. The measures were evaluated with the anechoic speech as reference and averaged over all utterances.

In Figure 2 the conventional approach (labeled as WPE-CONV) is compared with the proposed approach based on a CGG prior for the desired signal (labeled as CGG). It can be seen that by selecting an appropriate value of the shape parameter δ the proposed approach outperforms the conventional WPE for all considered performance measures. Note that the conventional WPE method is often employed using only a single iteration [5], and the presented results indicate that in this case the proposed approach could be used for obtaining a better performance.

6. CONCLUSIONS

In this paper we have presented a blind method for speech dereverberation based on MCLP, where the desired signal is modeled using a general sparse prior. The presented speech model is a generalization of the time-varying Gaussian model, and provides a possibility for modeling global properties of the desired signal. Experimental results demonstrate that the proposed framework can be used to improve instrumentally predicted speech enhancement performance, by introducing a small modification to the conventional approach.

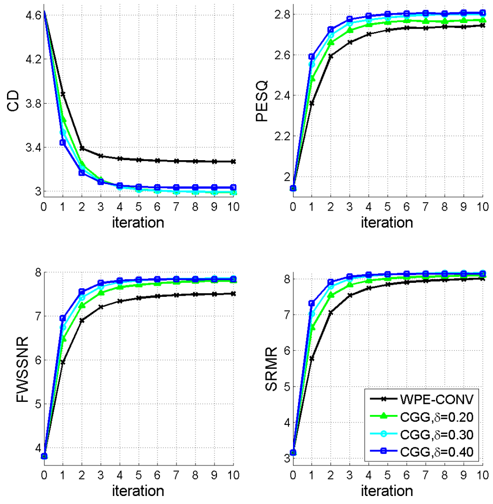

Fig. 2. Performance evaluation for the conventional (WPE-CONV) and the proposed approach (CGG). Iteration index zero denotes the value of a measure for the observed signal at the first microphone.

7. REFERENCES

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

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

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

[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. on Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1717–1731, Sept. 2010.

[6] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and Kalman smoother," in *Proc. EUSIPCO*, Marrakech, Morocco, Sept. 2013.

[7] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *Advances in Neural Information Processing Systems 18*, 2006.

[8] M. Novey, T. Adali, and A. Roy, "A Complex Generalized Gaussian Distribution - Characterization, Generation, and Estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1427–1433, Mar. 2010.

[9] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. ICASSP*, San Diego, California, USA, Mar. 1984, vol. 9, pp. 53–56.

[10] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. ICASSP*, Orlando, Florida, USA, May 2002, pp. I–253.

[11] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. IWAENC*, Tel Aviv, Israel, Sept. 2010.

[12] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. IWAENC*, Tel Aviv, Israel, Sept. 2010.

[13] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Aug. 2005.

[14] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, 2005.

[15] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," accepted for ICASSP, 2014.

[16] 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. WASPAA*, New Paltz, USA, Oct. 2013.