SPEECH DEREVERBERATION WITH CONVOLUTIVE TRANSFER FUNCTION APPROXIMATION USING MAP AND VARIATIONAL DECONVOLUTION APPROACHES

Ante Jukić[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

Recordings of a speech signal in an enclosed space are typically corrupted with reverberation. In combination with background noise, these effects may severely degrade the speech quality. In this paper we aim to blindly recover the speech signal from the reverberant and possibly noisy observations, where the signals are represented using the convolutive transfer function model in the STFT domain. The problem of blind speech dereverberation is decomposed into a set of independent blind deconvolution problems that we propose to solve using a maximum a posteriori approach and a variational approach, exploiting the sparsity of the speech signal in the STFT domain. The corresponding optimization problems can be solved using an alternating optimization procedure. Experimental results show that the proposed approach based on variational estimation results in consistent improvements of the instrumentally predicted measures of speech enhancement and dereverberation.

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

1. INTRODUCTION

When recording speech signals in an enclosure, they are typically corrupted by reverberation, due to reflections of the sound against the walls and objects in the room. While moderate reverberation can be perceptually beneficial, in severe cases reverberation can lead to significantly decreased speech quality and intelligibility [1–3]. Therefore, effective reverberation suppression is crucial for applications such as hands-free telephony or other voice-based systems. Because of its practical significance, speech dereverberation has been an active research field in recent years [4]. Several dereverberation techniques have been proposed, which are based on, e.g., multichannel equalization [5, 6], spectral enhancement [7], or probabilistic modeling of speech signals [8–12].

In this paper, we consider speech dereverberation in the short-time Fourier transform (STFT) domain, where the reverberant recordings are modeled using a convolutive transfer function (CTF) model: the time-domain convolution between the speech signal and the room impulse responses (RIRs) is approximated by a convolution between the speech signal STFT coefficients and a CTF in each frequency bin independently [13]. Independent modeling of reverberation in each frequency bin was also used in [9, 14], where an autoregressive model was employed. Speech dereverberation, under the CTF model, can essentially be performed in each frequency bin independently, by estimating the unknown speech signal and the CTFs. In [12], the microphone signals are represented in a state-space form, and an expectation-maximization (EM) algorithm was derived. The Kalman smoother was applied in the E step to estimate the speech signal and its covariance, and the remaining parameters are estimated in the M step. In a similar fashion, we formulate dereverberation in each frequency bin based on cost functions obtained from maximum a posteriori (MAP) and variational estimation [15]. Formally, the main difference between the two methods is that point estimates are used in MAP, while variational approaches estimate distributions for the unknown parameters. For the latter, we use a cost function that corresponds to an estimation procedure that is variational in the estimated speech signal only. The obtained cost functions are minimized using an alternating optimization strategy. In both cases, the speech signal is modeled using a time-varying Gaussian (TVG) model with unknown variances, as in [9, 12], that can be equivalently represented as a sparse prior on the speech signal [16]. The experimental results show that the obtained MAP estimation procedure is not suitable for dereverberation in the presented scenario, while the variational estimation reduces reverberation, as indicated by the evaluated measures.

The paper is organized as follows. In Section 2 we introduce the notation and formulate the problem of speech dereverberation using the CTF model. In Section 3 the proposed approaches are formulated, with experimental results given in Section 4.

2. PROBLEM FORMULATION

We assume that a speech signal $s(t)$ is recorded by a set of M microphones in an enclosed space. The microphone signals are typically corrupted by reverberation and additive noise. Assuming that the RIRs between the source and the microphones are time-invariant, the m-th microphone signal $y_m(t)$ can be modeled in the time-domain as

$$y_m(t) = \sum_{i=0}^{N_r-1} r_m(i)\, s(t-i) + w_m(t), \tag{1}$$

where N_r is the length of the RIR $r_m(t)$, and $w_m(t)$ is the additive noise. The time-domain convolution in (1) can be exactly represented in the STFT domain using the cross-band filter representation [13]. However, a reasonable model can be obtained if the cross-band filters are neglected, i.e., the time-domain convolution can be approximated using the convolutive transfer function model, e.g., as used in [7, 12]. Let $y_m(n,k)$ and $s(n,k)$ denote the STFT representations of the m-th microphone signal and the speech signal, with time frame index n and frequency bin index k. Using the CTF

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).

model, $y_m(n,k)$ can be written as

$$y_m(n,k) = \sum_{l=0}^{N_h-1} h_m(l,k)s(n-l,k) + v_m(n,k), \quad (2)$$

where $h_m(n,k)$ denotes the n-th temporal coefficient of the m-th microphone's CTF in the k-th frequency bin with length N_h, and $v_m(n,k)$ jointly represents the additive noise and the error of the CTF model approximation. The CTF model in (2) can be written more compactly, with $*$ denoting convolution across the frames, as

$$\mathbf{y}_m(k) = \mathbf{h}_m(k) * \mathbf{s}(k) + \mathbf{v}_m(k), \quad (3)$$

with the CTF $\mathbf{h}_m(k) = [h_m(0,k), \ldots, h_m(N_h-1,k)]^T$, the speech signal coefficients $\mathbf{s}(k) = [s(1,k), \ldots, s(N_s,k)]^T$, N_s denoting the length of $\mathbf{s}(k)$, and $\mathbf{y}_m(k), \mathbf{v}_m(k) \in \mathbb{C}^{N_y}$ defined similarly as $\mathbf{s}(k)$ with $N_y = N_s + N_h - 1$. The expression in (3) can be written as

$$\boxed{\mathbf{y}_m(k) = \mathbf{H}_m(k)\mathbf{s}(k) + \mathbf{v}(k) = \mathbf{S}(k)\mathbf{h}_m(k) + \mathbf{v}_m(k),} \quad (4)$$

where $\mathbf{H}(k) \in \mathbb{C}^{N_y \times N_s}$ and $\mathbf{S}(k) \in \mathbb{C}^{N_y \times N_h}$ are convolution matrices constructed using the CTF $\mathbf{h}_m(k)$ and the speech $\mathbf{s}(k)$, respectively. Blind speech dereverberation can now be formulated as estimating the clean speech coefficients $\mathbf{s}(k)$ and the CTFs $\mathbf{h}_m(k)$, given only the STFT coefficients of the reverberant and possibly noisy microphone signals $\mathbf{y}_m(k)$.

3. PROPOSED METHODS

The formulated blind dereverberation problem is ill-posed, e.g. its solution is in general not unique, since neither the clean speech signal $\mathbf{s}(k)$, the CTFs $\mathbf{h}_m(k)$ or the noise signals $\mathbf{v}_m(k)$ are available. In this section we make assumptions about the speech and noise signals, and use them to formulate two different estimation procedures. Unknown parameters for the deconvolution problem in (4) are estimated using a maximum a posteriori procedure in Section 3.2, and using an alternative estimator based on a variational method in Section 3.3. Both procedures are formulated as optimization problems that are tackled using an alternating optimization strategy.

3.1. Speech and noise model

Many probabilistic speech dereverberation methods, e.g. [9, 12], assume that the speech signal can be modeled using a time-varying Gaussian (TVG) model, i.e., the coefficients of the speech signal in each time-frequency bin are modeled as independent zero-mean random variables with a circular complex Gaussian distribution with an unknown and time-varying variance. The probability density function (PDF) associated with the STFT coefficient $s(n,k)$ is then assumed to be

$$\mathcal{N}_{\mathbb{C}}\left(s(n,k); 0, \lambda(n,k)\right) = \frac{1}{\pi\lambda(n,k)} \mathrm{e}^{-\frac{|s(n,k)|^2}{\lambda(n,k)}}, \quad (5)$$

with the variance $\lambda(n,k)$ considered as an unknown parameter that needs to be estimated. Since this model assumes speech coefficients to be independent across frequencies, in the sequel the frequency bin index k is omitted and it is assumed that we operate in each frequency bin independently. In each frequency bin we then have a set of unknown speech coefficients $\mathbf{s}$ that are modeled as independent Gaussian random variables with unknown variances $\boldsymbol{\lambda} = [\lambda(1), \ldots, \lambda(N_s)]^T$. The variances $\boldsymbol{\lambda}$ can for instance be estimated by maximizing the likelihood [9, 12]. As noted in [16], this is equivalent to assigning an improper prior for the speech coefficients in the form

$$p\left(s(n)\right) = \max_{\lambda(n)>0} \mathcal{N}_{\mathbb{C}}\left(s(n); 0, \lambda(n)\right) \propto \frac{1}{|s(n)|^2}, \quad (6)$$

which strongly promotes sparsity of the coefficients in a single frequency bin [16]. In general, various sparse priors can be represented similarly as in (6) as maximization over scaled Gaussian densities [17]. Since the speech coefficients in different time-frequency bins are assumed to be independent, we can write

$$p\left(\mathbf{s}\right) = \prod_{n=1}^{N_s} p\left(s(n)\right) = \max_{\boldsymbol{\lambda}>0} \prod_{n=1}^{N_s} \mathcal{N}_{\mathbb{C}}\left(s(n); 0, \lambda(n)\right), \quad (7)$$

for each frequency bin independently. By plugging (5) into (7) and taking the negative logarithm, we obtain

$$-\log p\left(\mathbf{s}\right) = \min_{\boldsymbol{\lambda}>0} \mathbf{s}^H \boldsymbol{\Lambda}^{-1}\mathbf{s} + \sum_{n=1}^{N_s} \log\lambda(n) + \text{const}, \quad (8)$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, and the constant term is not important for minimization as it is not function of $\boldsymbol{\lambda}$. For the error term $\mathbf{v}_m$ in (3), we assume that it can be modeled as independent and identically distributed noise with probability density function given as $p\left(v_m(n)\right) = \mathcal{N}_{\mathbb{C}}\left(v_m(n); 0, \sigma_m^2\right)$, such that

$$-\log p\left(\mathbf{v}_m\right) = \frac{1}{\sigma_m^2}\|\mathbf{v}_m\|_2^2 + N_y \log\sigma_m^2 + \text{const}. \quad (9)$$

Using the speech and noise models in (8) and (9) we can now formulate estimators for the deconvolution problem in (4).

3.2. Estimation using a MAP cost function

We start by formulating the MAP estimation problem using a single microphone signal. The posterior distribution for $\mathbf{s}$ and the m-th CTF $\mathbf{h}_m$ can be written as

$$p\left(\mathbf{s}, \mathbf{h}_m, \sigma_m^2 | \mathbf{y}_m\right) \propto p\left(\mathbf{y}_m, \mathbf{s}, \mathbf{h}_m, \sigma_m^2\right) \propto p\left(\mathbf{y}_m | \mathbf{s}, \mathbf{h}_m\right) p\left(\mathbf{s}\right), \quad (10)$$

where we assumed that the speech, the CTF and the noise variance are independent, and no prior knowledge is available for the CTF and the noise variance, i.e., their PDFs are constant. The first term on the right-hand side is the likelihood $p\left(\mathbf{y}_m | \mathbf{s}, \mathbf{h}_m\right)$ that is determined by the noise model in (9), and the second term is determined by the speech model in (8). The MAP estimation problem for the unknowns $\mathbf{s}, \boldsymbol{\lambda}, \mathbf{h}_m, \sigma_m^2$ can be reformulated as minimizing the negative log-posterior, using (8), (9) and (10) , as

$$\min_{\substack{\mathbf{s},\boldsymbol{\lambda}>0 \\ \mathbf{h}_m,\sigma_m^2>0}} \frac{1}{\sigma_m^2}\|\mathbf{y}_m - \mathbf{h}_m * \mathbf{s}\|_2^2 + \mathbf{s}^H\boldsymbol{\Lambda}^{-1}\mathbf{s} + \\ + \sum_{n=1}^{N_s} \log\lambda(n) + N_y \log\sigma_m^2. \quad (11)$$

The optimization problem can be solved by applying an alternating optimization strategy. When multiple microphone signals are available, for simplicity assuming the error term is spatially white, the optimization problem can be obtained by averaging the cost function in (11) over the M microphones. This results in the following

optimization problem

$$\min_{\substack{\mathbf{s},\boldsymbol{\lambda}>0\\ \{\mathbf{h}_m,\sigma_m^2\}>0}} \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\sigma_m^2}\|\mathbf{y}_m-\mathbf{h}_m*\mathbf{s}\|_2^2+\mathbf{s}^H\boldsymbol{\Lambda}^{-1}\mathbf{s}+ \\ +\sum_{n=1}^{N_s}\log\lambda(n)+\frac{1}{M}\sum_{m=1}^{M}N_y\log\sigma_m^2, \quad (12)$$

where $\{\mathbf{h}_m,\sigma_m^2\}=\{\mathbf{h}_m,\sigma_m^2\}_{m=1}^{M}$ denotes the set of all CTFs and noise variances[1]. Finally, the obtained non-convex optimization problem is solved by applying an alternating optimization strategy. This results in the updates for all of the unknowns, as given in Algorithm 1, that are iteratively repeated for a fixed number of iterations or until convergence is attained. The two main subproblems for estimating the speech coefficients $\mathbf{s}$ and the CTFs $\{\mathbf{h}_m\}$ are

$$\hat{\mathbf{s}}=\arg\min_{\mathbf{s}}\sum_{m=1}^{M}\frac{1}{M\hat{\sigma}_m^2}\|\mathbf{y}_m-\hat{\mathbf{H}}_m\mathbf{s}\|_2^2+\mathbf{s}^H\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{s} \quad (13)$$

$$\hat{\mathbf{h}}_m=\arg\min_{\mathbf{h}_m}\|\mathbf{y}_m-\hat{\mathbf{S}}\mathbf{h}_m\|_2^2. \quad (14)$$

The obtained problems are quadratic and convex, with their closed-form solutions given in Algorithm 1. The subproblem in (13) is a non-blind deconvolution using the estimated CTFs, with an additional quadratic penalty on $\mathbf{s}$. The weights in the quadratic penalty are equal to the inverse of the previously estimated variances (cf. Algorithm 1), resulting in relatively larger weights for the coefficients in $\mathbf{s}$ with a small amplitude. This type of reweighting mimics the behiavior of ℓ_0-norm on $\mathbf{s}$, that is invariant to scaling of the coefficients, thus favoring a sparse estimate $\hat{\mathbf{s}}$ [18]. By solving the subproblem in (14) a new estimate for the m-th CTF is obtained, using the previously estimated speech signal $\hat{\mathbf{s}}$. In terms of computational complexity, the subproblem for estimation of $\hat{\mathbf{s}}$ is dominant, since its dimension depends on the length of the signal, and the matrix to be inverted is of size $N_s\times N_s$ (cf. Algorithm 1). However, efficient conjugate gradient solvers can be applied in case of long signals (i.e., large N_s), and the inverse in Algorithm 1 does not have to be explicitly calculated to solve the linear system. A small lower bound for variances λ_{lb} is included in Algorithm 1 to prevent the estimates from going to zero. The noise variance update in Algorithm 1 also includes a heuristically set lower bound σ_{lb}^2, to prevent σ_m^2 from going to zero. To avoid scaling ambiguity, the estimate $\hat{\mathbf{s}}$ is rescaled after each iteration to have the same norm as $\mathbf{y}_1$, ensuring that the average power spectrum of the estimated speech is the same as on the first microphone [12].

3.3. Estimation using a cost function based on variational estimation

Several blind deconvolution methods are based on a variational estimation, mainly in the context of image deblurring [15, 18–20]. In this section we apply an estimation strategy variational in $\mathbf{s}$ for speech dereverberation using the CTF model in (4). Again, we start by formulating the estimation problem for a single microphone signal. The main idea is to perform marginalization over $\mathbf{s}$, and then estimate the remaining parameters $\mathbf{h}_m,\boldsymbol{\lambda},\sigma_m^2$ by maximizing the obtained marginalized posterior [18]. To make the calculations

[1]Formally, in the direct MAP formulation for the multi-channel case, the likelihood term should not be weighted with $1/M$. However, omitting the scaling (and giving a larger weight to the likelihood) resulted in a worse performance for $M>1$.

Algorithm 1 Dereverberation based on MAP estimation, performed independently for each frequency bin. $(.)^\dagger$ denotes pseudoinverse, and $|.|$ denotes element-wise absolute value.

parameters N_h in (2), lower bounds $\lambda_{\mathrm{lb}},\sigma_{\mathrm{lb}}^2$, maximum number of iterations $i_{\max}$
input M signals $\{\mathbf{y}_m\}$
initialization $\hat{\mathbf{h}}_m,\hat{\boldsymbol{\lambda}},\hat{\sigma}_m^2$
repeat
$\hat{\mathbf{s}}\leftarrow\left(\sum_m\frac{\hat{\mathbf{H}}_m^H\hat{\mathbf{H}}_m}{M\hat{\sigma}_m^2}+\hat{\boldsymbol{\Lambda}}^{-1}\right)^{-1}\sum_m\frac{\hat{\mathbf{H}}_m^H\mathbf{y}_m}{M\hat{\sigma}_m^2}$
$\hat{\mathbf{s}}\leftarrow\hat{\mathbf{s}}\cdot\|\mathbf{y}_1\|_2/\|\hat{\mathbf{s}}\|_2$
$\hat{\boldsymbol{\lambda}}\leftarrow|\hat{\mathbf{s}}|^2+\lambda_{\mathrm{lb}}$
$\hat{\mathbf{h}}_m\leftarrow\hat{\mathbf{S}}^\dagger\mathbf{y}_m$
$\hat{\sigma}_m^2\leftarrow\frac{\|\mathbf{y}_m-\hat{\mathbf{H}}_m\hat{\mathbf{s}}\|_2^2}{N_y}+\sigma_{\mathrm{lb}}^2$
until convergence or $i_{\max}$ exceeded

tractable the mean field approximation is employed [21], i.e., the posterior $p(\mathbf{s},\boldsymbol{\lambda},\mathbf{h}_m|\mathbf{y}_m)$ is approximated by a factorized distribution $q(\mathbf{s})\,q(\boldsymbol{\lambda})\,q(\mathbf{h}_m)$, with $q(\mathbf{s})\,q(\boldsymbol{\lambda})=\prod_n q(s(n))\,q(\lambda(n))$, and the latter factorization resulting in a posterior for $\mathbf{s}$ with a diagonal covariance [15, 19]. In [18] it has been shown that the estimation procedure can be equivalently formulated in a penalized form, similar to the MAP estimation problem in (11). Using the results in [18], the following optimization problem is formulated

$$\min_{\substack{\mathbf{s},\boldsymbol{\lambda}>0\\ \mathbf{h}_m,\sigma_m^2>0}} \frac{1}{\sigma_m^2}\|\mathbf{y}_m-\mathbf{h}_m*\mathbf{s}\|_2^2+\mathbf{s}^H\boldsymbol{\Lambda}^{-1}\mathbf{s}+ \\ +\sum_{n=1}^{N_s}\log\left(\sigma_m^2+\lambda(n)\|\mathbf{h}_m\|_2^2\right). \quad (15)$$

The main difference when compared to the MAP problem in (11) is that the noise variance, the speech variance and the CTF are coupled in the last penalty term. When multiple microphone signals are available, assuming the error term is spatially white, the optimization problem can be obtained by averaging the cost function in (15) over the M microphones, similarly as in [20]. This results in the following optimization problem

$$\min_{\substack{\mathbf{s},\boldsymbol{\lambda}>0\\ \{\mathbf{h}_m,\sigma_m^2>0\}}} \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\sigma_m^2}\|\mathbf{y}_m-\mathbf{h}_m*\mathbf{s}\|_2^2+\mathbf{s}^H\boldsymbol{\Lambda}^{-1}\mathbf{s}+ \\ +\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_s}\log\left(\sigma_m^2+\lambda(n)\|\mathbf{h}_m\|_2^2\right). \quad (16)$$

Again, the obtained non-convex optimization problem in (16) can be solved by applying an alternating optimization strategy. However, the variables are coupled and convenient upper bounds have to be used, as derived in [18]. This results in the updates as given in Algorithm 2, that are iteratively repeated for a fixed number of iterations or until convergence is attained. The main two subproblems for estimating $\mathbf{s}$ and $\mathbf{h}_m$ are given as

$$\hat{\mathbf{s}}=\arg\min_{\mathbf{s}}\sum_{m=1}^{M}\frac{1}{M\hat{\sigma}_m^2}\|\mathbf{y}_m-\hat{\mathbf{H}}_m\mathbf{s}\|_2^2+\mathbf{s}^H\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{s} \quad (17)$$

$$\hat{\mathbf{h}}_m=\arg\min_{\mathbf{h}_m}\|\mathbf{y}_m-\hat{\mathbf{S}}\mathbf{h}_m\|_2^2+\hat{\alpha}_m\|\mathbf{h}_m\|_2^2, \quad (18)$$

where $\hat{\alpha}_m$ is a regularization parameter for the ℓ_2-regularized estimation of the CTF $\mathbf{h}_m$. The subproblem for estimating $\mathbf{s}$ in (17)

has the same structure as the one in (13), but with different estimation of the variances $\hat{\boldsymbol{\lambda}}$ (cf. Algorithm 2), that are now calculated using the estimate $\hat{\mathbf{s}}$ and the vectors $\mathbf{c}_m$ (that give a diagonal approximation of the covariance of $\hat{\mathbf{s}}$). As a result, the penalty term is less concave when the covariance of the estimate $\hat{\mathbf{s}}$ is large, while it becomes more similar to the reweighting in (13) when the covariance is decreased [18]. Intuitively, a less aggressive sparse penalty helps to avoid local minima when the estimate $\hat{\mathbf{s}}$ is not reliable (in terms of a large covariance). By comparing (18) with (14), it can be seen that the subproblem in (18) includes an additional quadratic penalty on the CTF, with the amount of regularization $\hat{\alpha}_m$ automatically determined (cf. Algorithm 2). Also, the noise variance update has an additional term dependent on the CTF estimate, that prevents a fast reduction of the noise variance estimate. Similar as for MAP estimation, the final estimate of the speech is scaled to have the same average power spectrum as the first input. The cost function in (16) is invariant in the sense that replacing $\mathbf{s}, \boldsymbol{\lambda}, \{\mathbf{h}_m\}$ with $c\mathbf{s}, c^2\boldsymbol{\lambda}, \{c^{-1}\mathbf{h}_m\}$, $c \in \mathbb{R}$, does not change the cost function value. Therefore, only the final estimate is rescaled to have the desired norm $\|\mathbf{y}_1\|_2$. The obtained estimation procedure is similar to the EM-based estimation in [12], where a Kalman smoother was applied to estimate the mean and covariance of the speech, that are used to update the remaining parameters.

Algorithm 2 Dereverberation based on variational estimation, performed independently for each frequency bin. $1/.$ denotes element-wise division, and $|.|$ denotes element-wise absolute value.

parameters N_h in (2), lower bounds $\lambda_{\text{lb}}, \sigma^2_{\text{lb}}$, maximum number of iterations $i_{\max}$
input M signals $\{\mathbf{y}_m\}$
initialization $\hat{\mathbf{h}}_m, \hat{\boldsymbol{\lambda}}, \hat{\sigma}^2_m$
repeat

$$\hat{\mathbf{s}} \leftarrow \left(\sum_m \frac{\hat{\mathbf{H}}_m^H \hat{\mathbf{H}}_m}{M\hat{\sigma}^2_m} + \hat{\boldsymbol{\Lambda}}^{-1}\right)^{-1} \sum_m \frac{\hat{\mathbf{H}}_m^H \mathbf{y}_m}{M\hat{\sigma}^2_m}$$

$$\hat{\mathbf{c}}_m \leftarrow \frac{1}{\left(\|\hat{\mathbf{h}}_m\|_2^2/\hat{\sigma}^2_m + 1/\hat{\boldsymbol{\lambda}}\right)}$$

$$\hat{\boldsymbol{\lambda}} \leftarrow |\hat{\mathbf{s}}|^2 + \frac{1}{M}\sum_m \hat{\mathbf{c}}_m + \lambda_{\text{lb}}$$

$$\hat{\alpha}_m = \sum_n \hat{c}_m(n)$$

$$\hat{\mathbf{h}}_m \leftarrow \left(\hat{\mathbf{S}}^H \hat{\mathbf{S}} + \hat{\alpha}_m \mathbf{I}\right)^{-1} \hat{\mathbf{S}}^H \mathbf{y}_m$$

$$\hat{\sigma}^2_m \leftarrow \frac{\|\mathbf{y}_m - \hat{\mathbf{H}}_m \hat{\mathbf{s}}\|_2^2 + \hat{\alpha}_m \|\hat{\mathbf{h}}_m\|_2^2}{N_y} + \sigma^2_{\text{lb}}$$

until convergence or $i_{\max}$ exceeded

$$\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} \cdot \|\mathbf{y}_1\|_2 / \|\hat{\mathbf{s}}\|_2$$

4. EXPERIMENTS

In this section we investigate the performance of the presented dereverberation approaches for several experimental conditions. We used a set of 10 sounds samples with utterances from different speakers, where the average length of the utterances was approximately 3.3 s and the sampling frequency was $f_s = 16$ kHz. The microphone signals were generated by convolving an utterance with RIRs that were measured in a room with a reverberation time $RT \approx 750$ ms, with the source positioned 2.3 m from the microphones. In the experiments we use a setup with $M = 1$ and $M = 4$ microphones. The obtained reverberant signals were further degraded with spatially white speech-shaped noise at a reverberant speech to noise ratio (RSNR) of $\{10, 20, 30\}$ dB. The STFT was computed using a 64 ms Hanning analysis window with 16 ms frame shift. In all experiments the CTFs length was set to $N_h = 47$, to approximately match the reverberation time. The initialization for both MAP and variational approaches was performed as follows. The first 4 taps of

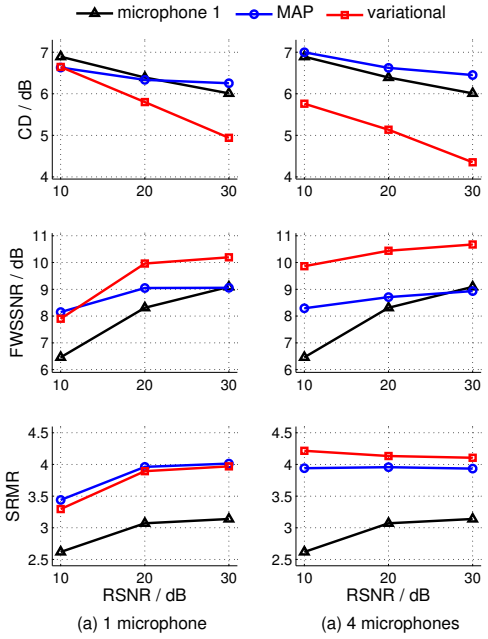

Fig. 1. Evaluated objective measures for the setup with (a) one microphone and (b) four microphones.

the CTFs were initialized to 1, and the remaining taps were set to zero, just to avoid a trivial solution. The variances were initialized using the signal from the first microphone as $\hat{\boldsymbol{\lambda}} = |\mathbf{y}_1|^2$ for each frequency bin, with the lower bound λ_{lb} set to 10^{-8}. The noise variance $\hat{\sigma}^2_m$ was initialized as the average power of $\mathbf{y}_1$ and the lower bound was set to 1% of the initial value. Ten iterations are performed for both approaches.

The performance was evaluated in terms of cepstral distance (CD), frequency-weighted segmental signal-to-noise ratio (FWSSNR), and speech-to-reverberant modulation energy ratio (SRMR) [22]. The measures were evaluated with the anechoic speech as reference, and the values reported in Figure 1 are obtained by averaging over all utterances. The presented results show that the proposed method based on variational estimation performs better than MAP estimation in terms of objective measures, both for single microphone and multiple microphones. The MAP estimation does not necessarily perform better when multiple microphones are available, since the noise variance update and a fixed speech penalty result in the optimization being trapped in a local minima. Both of these problems are handled better by the variational estimation procedure, that results in a better performance in terms of the evaluated measures when compared to the MAP-based procedure.

5. CONCLUSIONS

We have presented a framework for blind speech dereverberation in the STFT domain, using the CTF model approximation and blind deconvolution techniques. The presented algorithms are based on the cost functions obtained using MAP estimation and an estimation procedure variational in $\mathbf{s}$. Sparsity of the STFT coefficients of the speech signal is exploited in both approaches. The experimental results show that direct application of the proposed MAP estimation procedure is not suitable for dereverberation in the presented scenario, while variational estimation procedure results in improved speech quality in terms of the evaluated measures when compared to the reverberant and noisy microphone signal.

6. 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," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, July 2006.

[3] A. Sehr, *Reverberation Modeling for Robust Distant-Talking Speech Recognition*, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürenberg, Erlangen, Oct. 2009.

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

[5] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," vol. 36, no. 2, pp. 145–152, Feb. 1988.

[6] 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.

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

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

[9] 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.

[10] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 17–20.

[11] D. Schmid, S. Malik, and G. Enzner, "A maximum a posteriori approach to multichannel speech dereverberation and denoising," in *Proc. IWAENC*, Aachen, Germany, Sept. 2012, pp. 1–4.

[12] 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.

[13] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[14] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on Itakura-Saito distance minimization," in *Proc. ICASSP*, Kyoto, Japan, May 2012, pp. 245–248.

[15] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding blind deconvolution algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2354–2367, Dec. 2011.

[16] 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 on Hands-free Speech Communication and Microphone Arrays*, Nancy, France, May 2014.

[17] 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.

[18] D. Wipf and H. Zhang, "Analysis of Bayesian blind deconvolution," in *Proc. Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lund, Sweden, Aug. 2013, pp. 40–53.

[19] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, "Bayesian blind deconvolution with general sparse image priors," in *Proc. European Conference on Computer Vision*, Florence, Italy, Oct. 2012, pp. 341–355.

[20] H. Zhang, D. Wipf, and Y. Zhang, "Multi-image blind deblurring using a coupled adaptive sparse prior," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013, pp. 1051–1058.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2006.

[22] 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, pp. 1 – 4.