EUSIPCO 2013 1569744095

MULTI-MICROPHONE SPEECH DEREVERBERATION USING EXPECTATION-MAXIMIZATION AND KALMAN SMOOTHING

Boaz Schwartz[1], *Sharon Gannot*[1], *and Emanuël A.P. Habets*[2]

[1] Faculty of Engineering
Bar-Ilan University
Ramat-Gan, 52900, Israel
boazsh0@gmail.com; sharon.gannot@biu.ac.il

[2] International Audio Laboratories Erlangen
University of Erlangen-Nuremberg
Am Wolfsmantel 33, 91058 Erlangen, Germany
e.habets@ieee.org

ABSTRACT

Speech signals recorded in a room are commonly degraded by reverberation. In most cases, both the speech signal and the acoustic system of the room are unknown. In this paper, a multi-microphone algorithm that simultaneously estimates the acoustic system and the clean signal is proposed. An expectation-maximization (EM) scheme is employed to iteratively obtain the maximum likelihood (ML) estimates of the acoustic parameters. In the expectation step, the Kalman smoother is applied to extract the clean signal from the data utilizing the estimated parameters. In the maximization step, the parameters are updated according to the output of the Kalman smoother. Experimental results show a significant dereverberation capabilities of the proposed algorithm with only low speech distortion.

1. INTRODUCTION

Microphones located within an enclosure capture a large number of reflections from the surrounding walls, ceiling, floor and other objects. These delayed arrivals, typically known as reverberation, can severely deteriorate the speech quality and intelligibility. Speech dereverberation was therefore the goal of numerous studies in the last decade, many of them exploit the advantages of multi-microphone schemes. Of special interest for the current contribution, are multi-microphone algorithms that utilize the acoustic systems relating the source and the microphones (or their respective inverse system). In practice, the acoustic systems of the room are unknown in advance, and should be estimated from the reverberant measurements. Examples of such estimation algorithms are the subspace methods in [1], and the multi-channel linear prediction methods in [2].

A Bayesian approach can also be adopted attributing a statistical model to the room impulse response (RIR). It was proposed in [3] to use the unscented Kalman filter for joint RIR estimation and speech dereverberation. In [4], the Kalman filter is used to estimate the dereverberated speech, and a particle filter is utilized to estimate the RIR of the reverberant room.

In [5], an EM scheme for retrieving the relevant speech and room parameters is presented. The reverberant speech is modelled as an auto-regressive (AR) process in each frequency band. The algorithm iterates until convergence. In the *E-step* the Wiener filter, calculated at the current values of the parameters, is applied to estimate the clean speech signal. In the *M-step*, the current estimated signal is used for updating the parameters. It was shown by Dempster, Laird and Rubin [6] that the EM algorithm is guaranteed to converge to a local maximum of the likelihood function. Note, that the algorithm presented in [5] is basically an expectation-conditional maximization (ECM) algorithm, a generalized version of EM algorithm. Another EM-based algorithm was proposed in [7] where the E- and M-steps objectives switch roles, namely the channel is identified at the E-step, and the clean speech at the M-step.

In this paper, we develop an EM algorithm for multi-microphone dereverberation in the short-time Fourier transform (STFT) domain. This choice is motivated by the length of the RIRs in the time-domain. In the STFT domain, the reverberation system can be approximated by the convolutive transfer function (CTF) model [8]. Under this approximation, the reverberation is modelled as a convolution with a finite impulse response (FIR) filter in each subband, which is much shorter than the original RIR.

The proposed algorithm aims at the ML estimation of the acoustic parameters of the room, and the dereverberated speech is actually estimated as a by-product of the parameter estimation procedure. For the parameter estimation we use the EM-Kalman scheme. This scheme was first formulated, in the time-domain, by Weinstein et al. in [9], and was later used for single microphone speech enhancement by Gannot et al. in [10]. In the E-step of the proposed algorithm, a Kalman smoother is used to estimate the dereverberated speech. In the M-step, the speech estimate is used to update the parameters. The algorithm iterates until convergence. Due to assumed signal model, the EM iterations in the proposed algorithm are simpler than the ECM iterations used in [5]. Simulation results show that, using the proposed algorithm, reverberation is significantly reduced, while speech quality increases.

The structure of the paper is as follows. The statistical model is formulated in Sec. 2. In Sec. 3, the algorithm derivation is presented. In Sec. 4, practical aspects concerning the implementation are discussed. Simulation results are given in Sec. 5, and conclusions are drawn in Sec. 6.

2. STATISTICAL MODEL

Let $x[n]$ be a clean speech signal in time-domain. The noisy and reverberant speech signal received by the jth microphone is given by

$$z_j[n] = x[n] * h_j[n] + v_j[n], \tag{1}$$

where $h_j[n]$ is the RIR between the speaker and the jth microphone, and $v_j[n]$ is an additive sensor noise.

In the STFT domain, $x(t,k)$ denotes the clean speech in time-frame t and frequency-bin k. Assuming $x[n]$ is short-term stationary

This research was supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development.

signal, and applying a proper STFT analysis intervals, $x(t,k)$ can be modelled as independent complex-Gaussian random variables

$$x(t,k) \sim \mathcal{N}_C\left\{0, \sigma_x^2(t,k)\right\}, \tag{2}$$

where $\sigma_x^2(t,k)$ denotes the short-time power spectra of $x[n]$. Now, (1) can be represented in the STFT domain as described by [8]

$$z_j(t,k) = \sum_{k'=0}^{K-1} \sum_{l=-\infty}^{\infty} h_{j,l}(k,k') x(t-l,k') + v_j(t,k), \tag{3}$$

where k and k' denote the band and cross-band frequency bin indices, respectively, K is the number of frequency bands, and $v_j(t,k)$ the additive noise. As in [11], we consider only the band-to-band filters ($k' = k$), i.e.

$$z_j(t,k) \approx \sum_{l=-\infty}^{\infty} h_{j,l}(k) x(t-l,k) + v_j(t,k). \tag{4}$$

Eq. (4) can be expressed in a vector form as

$$z_j(t,k) = \mathbf{h}_j^T(k)\mathbf{x}_t(k) + v_j(t,k), \tag{5}$$

where

$$\mathbf{h}_j(k) = \left[h_{j,L-1}(k),\ \ldots,\ h_{j,0}(k)\right]^T, \tag{6}$$

$$\mathbf{x}_t(k) = \left[x(t-L+1,k),\ \ldots, x(t,k)\right]^T, \tag{7}$$

and L is the CTF length that depends on the reverberation time. We further assume that $v_j(t,k)$ are stationary complex-Gaussian random variables:

$$v_j(t,k) \sim \mathcal{N}_C\left\{0, \sigma_{v_j}^2(k)\right\}. \tag{8}$$

Note that due to the approximation in (4), $v_j(t,k)$ may comprise an additional error component such that the variance of $v_j(t,k)$ may depend on $\left\{x\left(\bar{t},\bar{k}\right):\ \bar{t} \in \mathcal{T}_t,\ \bar{k} \in \mathcal{K}_k\right\}$, where $\mathcal{T}_t$ and $\mathcal{K}_k$ are close neighbours of t and k, respectively. In that case, $v_j(t,k)$ becomes a non-stationary signal. In the current study, this phenomenon is however neglected and the noise variance is assumed to be constant.

3. ALGORITHM DERIVATION

In the following section we develop an ML method for estimating the parameters. The problem is formulated in Sec. 3.1, the signal estimation is presented in Sec. 3.2, and the acoustic parameters estimation is developed in Sec. 3.3. The proposed algorithm is nicknamed Kalman-EM for dereverberation (KEMD).

3.1. Parameter Estimation Problem

Let $\mathcal{Z}$ be a set of measurements:

$$\mathcal{Z} = \left\{z_j(t,k):\ 1 \le j \le J,\ 1 \le t \le T,\ 0 \le k \le K-1\right\},$$

where T is the number of observed STFT frames, and J the number of microphones. Our goal is to maximize the likelihood function $f(\mathcal{Z};\mathbf{\Theta})$ with respect to the model parameters:

$$\mathbf{\Theta} \equiv \left\{\mathbf{\Theta}_X, \mathbf{\Theta}_H, \mathbf{\Theta}_V\right\} \tag{9a}$$

$$\mathbf{\Theta}_X \equiv \left\{\sigma_x^2(t,k)\right\} \tag{9b}$$

$$\mathbf{\Theta}_H \equiv \left\{\mathbf{h}_j(k)\right\} \tag{9c}$$

$$\mathbf{\Theta}_V \equiv \left\{\sigma_{v_j}^2(k)\right\} \tag{9d}$$

for each $1 \le j \le J,\ 1 \le t \le T,\ 0 \le k \le K-1$.

In order to solve this maximization problem, we adopt the EM approach. The latent data in this problem is defined to be the clean speech signal

$$\mathcal{X} = \left\{x(t,k):\ 1 \le t \le T,\ 0 \le k \le K-1\right\}.$$

In the E-step, the following function is calculated:

$$Q\left(\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(\ell)}\right) \equiv E\left\{\log f(\mathcal{Z},\mathcal{X};\mathbf{\Theta})\middle|\mathcal{Z};\widehat{\mathbf{\Theta}}^{(\ell)}\right\}, \tag{10}$$

where $\widehat{\mathbf{\Theta}}^{(\ell)}$ is the parameter estimate at iteration ℓ. For conciseness, the frequency index k will be omitted in the rest of the derivation. In the M-step, $\widehat{\mathbf{\Theta}}^{(\ell+1)}$ is derived by solving:

$$\widehat{\mathbf{\Theta}}^{(\ell+1)} = \arg\max_{\mathbf{\Theta}} Q\left(\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(\ell)}\right). \tag{11}$$

The statistical model in Sec. 2 assumes independence between adjacent time frames, and between speech and noise signals. Therefore, the log-likelihood of the complete data is:

$$\log f(\mathcal{X},\mathcal{Z};\mathbf{\Theta}) = C - \frac{1}{2}\sum_{t=1}^{T}\left[\log\sigma_x^2(t) + \frac{|x(t)|^2}{\sigma_x^2(t)}\right] - \frac{1}{2}\sum_{j=1}^{J}\left[T\log\sigma_{v_j}^2 + \frac{1}{\sigma_{v_j}^2}\sum_{t=1}^{T}\left|z_j(t) - \mathbf{h}_j^T\mathbf{x}_t\right|^2\right], \tag{12}$$

where C is a constant value independent of the parameters, the first summation term is the log-likelihood of clean speech signal, and the second summation term is related to the noise signal.

3.2. E-Step: Kalman Smoother

Following (10), we need to calculate the expected value of (12) given the measurement at hand, and the current parameter estimate. Given $\mathcal{Z}$ and $\widehat{\mathbf{\Theta}}^{(\ell)}$, the expected values of $\mathbf{x}_t$ and $\mathbf{x}_t\mathbf{x}_t^\dagger$ should be calculated, where † is the conjugate transpose operator. To obtain the expected value, the minimum mean square error (MMSE) estimator should be applied. In order to efficiently calculate the MMSE estimates, we formulate the signal model in state-space and apply the Kalman smoother [9, 10]:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t,\\ \mathbf{z}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{v}_t,\end{aligned} \tag{13}$$

where $\mathbf{x}_t$ was defined in (7), the innovation process is given by

$$\mathbf{w}_t \equiv \left[0,\ \ldots,\ x(t)\right]^T,$$

the measurement and noise vectors are equal to

$$\mathbf{z}_t \equiv \left[z_1(t),\ \ldots,\ z_J(t)\right]^T,$$
$$\mathbf{v}_t \equiv \left[v_1(t),\ \ldots,\ v_J(t)\right]^T,$$

and the process and measurement matrices are respectively equal to

$$\mathbf{\Phi} \equiv \begin{pmatrix} 0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1\\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix},$$

and $\mathbf{H} \equiv \left[\mathbf{h}_1,\ \ldots,\ \mathbf{h}_J\right]^T$, where $\mathbf{h}_j$ were defined in (6).

Note, that unlike the time-domain state-space representation in [9, 10], here the process is not modelled as an AR signal, as evident from the absence of regression parameters in $\mathbf{\Phi}$. In our model, the statistical dependency of adjacent time-frames of $x(t)$ can be discarded if the overlap between STFT frames is sufficiently small, as will be discussed in Sec. 5.

Finally, the second-order statistics matrices are defined as:

$$\mathbf{Q}_t \equiv E\left\{\mathbf{w}_t\mathbf{w}_t^\dagger\right\} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_x^2(t) \end{bmatrix}$$

$$\mathbf{R} \equiv E\left\{\mathbf{v}_t\mathbf{v}_t^\dagger\right\} = \begin{bmatrix} \sigma_{v_1}^2 & \cdots & \cdots & 0 \\ 0 & \sigma_{v_2}^2 & & \\ & & \ddots & \\ 0 & \cdots & \cdots & \sigma_{v_J}^2 \end{bmatrix}.$$

The Kalman smoothing procedure is summarized in Algorithm 1.

The outcome of the smoothing recursion is the state-vectors estimators, and the respective estimation covariance matrices of the entire observation interval, namely $\left\{\widehat{\mathbf{x}}_{t|T}, \mathbf{P}_{t|T} : 1 \le t \le T\right\}$. In the M-step, given in the sequel, the following first- and second-order statistics terms are used [10]:

$$E\left\{\mathbf{x}_t \middle| \mathcal{Z}; \mathbf{\Theta}\right\} = \widehat{\mathbf{x}}_{t|T}, \tag{14a}$$

$$E\left\{\mathbf{x}_t\mathbf{x}_t{}^\dagger \middle| \mathcal{Z}; \mathbf{\Theta}\right\} = \widehat{\mathbf{x}}_{t|T}\widehat{\mathbf{x}}_{t|T}^\dagger + \mathbf{P}_{t|T}. \tag{14b}$$

3.3. M-Step: Parameter Estimation

The solution of (11) is obtained by setting the partial derivatives with respect to the parameters to zero, resulting in:

$$\widehat{\sigma_x^2}^{(\ell)}(t) = \widehat{|x(t)|^2}, \tag{15}$$

Algorithm 1: The Kalman Smoother.

Forward recursion (Kalman filter):
for $t = 1$ **to** T **do**
  Predict:
  $\widehat{\mathbf{x}}_{t|t-1} = \mathbf{\Phi} \cdot \widehat{\mathbf{x}}_{t-1|t-1}$
  $\mathbf{P}_{t|t-1} = \mathbf{\Phi} \cdot \mathbf{P}_{t-1|t-1} \cdot \mathbf{\Phi}^T + \mathbf{Q}_t$
  Update:
  $\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}^\dagger \left[\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^\dagger + \mathbf{R}\right]^{-1}$
  $\mathbf{e}_t = \mathbf{z}_t - \mathbf{H}\widehat{\mathbf{x}}_{t|t-1}$
  $\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t \cdot \mathbf{e}_t$
  $\mathbf{P}_{t|t} = \left[\mathbf{I} - \mathbf{K}_t\mathbf{H}\right]\mathbf{P}_{t|t-1}$
end

Backward recursion (smoothing):
for $t = T$ **to** 2 **do**
  $\mathbf{S}_{t-1} = \mathbf{P}_{t-1|t-1}\mathbf{\Phi}^T\mathbf{P}_{t|t-1}^{-1}$
  $\mathbf{e}_{t|T} = \mathbf{x}_{t|T} - \mathbf{\Phi}\widehat{\mathbf{x}}_{t-1|t-1}$
  $\widehat{\mathbf{x}}_{t-1|T} = \widehat{\mathbf{x}}_{t-1|t-1} + \mathbf{S}_{t-1}\mathbf{e}_{t|T}$
  $\mathbf{P}_{t-1|T} = \mathbf{P}_{t-1|t-1} + \mathbf{S}_{t-1}\left[\mathbf{P}_{t|T} - \mathbf{P}_{t|t-1}\right]\mathbf{S}_{t-1}^T$
end

$$\left(\widehat{\mathbf{h}}_j^{(\ell)}\right)^T = \left(\sum_{t=1}^{T} \widehat{\mathbf{x}_t\mathbf{x}_t{}^\dagger}\right)^{-1} \times \left(\sum_{t=1}^{T} z_j(t) \cdot \widehat{\mathbf{x}}_t^\dagger\right), \tag{16}$$

$$\begin{aligned}\widehat{\sigma_{v_j}^2}^{(\ell)} &= \frac{1}{T}\sum_{t=1}^{T} \widehat{\left|z_j(t) - \left(\widehat{\mathbf{h}}_j^{(\ell)}\right)^T \widehat{\mathbf{x}}_t\right|^2} \\ &= \frac{1}{T}\sum_{t=1}^{T}\Bigg\{|z_j(t)|^2 - 2\,\mathrm{Re}\left(\left(\widehat{\mathbf{h}}_j^{(\ell)}\right)^T \widehat{\mathbf{x}}_t\right) \\ &\qquad + \left(\widehat{\mathbf{h}}_j^{(\ell)}\right)^T \widehat{\mathbf{x}_t\mathbf{x}_t{}^\dagger} \left(\widehat{\mathbf{h}}_j^{(\ell)}\right)^*\Bigg\},\end{aligned} \tag{17}$$

where

$$\widehat{(\cdot)} \equiv E\left\{(\cdot)\middle|\mathcal{Z}; \widehat{\mathbf{\Theta}}^{(\ell-1)}\right\}$$

is the MMSE estimator obtained from the application of Kalman smoother at the $(\ell-1)$th iteration, as given in (14), and

$$\widehat{\theta}^{(\ell)} = \underset{\theta}{\mathrm{argmax}}\, Q\left(\mathbf{\Theta}\middle|\widehat{\mathbf{\Theta}}^{(\ell-1)}\right)$$

is the updated parameter at the ℓth iteration.

4. PRACTICAL CONSIDERATIONS

The EM algorithm is known to be sensitive to initialization. In this work, no localization knowledge is considered. We suggest to initialize the acoustic systems $\mathbf{H}$ with the direct-path (represented in the STFT domain) of a source positioned at the broadside of the array, regardless of its true position. For comparison, we have also used another initialization procedure, for which the true direct-paths from the source to each of the microphones are assumed to be known a priori (which is equivalent to the assumption that the source location is known). It was experimentally verified that the prior information on the source location is not required.

An initial value for $\sigma_x^2(t,k)$ should also be set. We have tried two alternative initialization procedures. In the first, the variance of the reverberant and noisy signal z_1 is used, while in the second, z_1 is first preprocessed with a spectral enhancement (SE) dereverberation algorithm [11]. In Sec. 5, we show that while the latter alternative yields better speech quality, the former alternative might suffice in many practical scenarios.

The reverberant model in (4) suffers from an inherent *gain ambiguity* problem, which is evident from the following equation: $\mathbf{h}_j^T(k)\mathbf{x}_t(k) = \left(g(k)\mathbf{h}_j^T(k)\right)\left(\frac{1}{g(k)}\mathbf{x}_t(k)\right)$, where $g(k)$ is an arbitrary frequency-dependent gain. Since the algorithm is independently applied to each frequency bin, this can result in undesired fluctuations in the spectral envelope. As a practical cure to this problem, we have constrained the power profile of the system output to match the respective power at the input. Hence, at each iteration, the estimated parameter set (at the kth frequency band) is substituted by its normalized counterpart:

$$\widehat{\sigma_x^2}^{(\ell)}(t,k) \leftarrow b^2(k) \cdot \widehat{\sigma_x^2}^{(\ell)}(t,k)\ ,\ 0 \le t \le T-1 \tag{18}$$

$$\widehat{\mathbf{h}_j}^{(\ell)}(k) \leftarrow \tfrac{1}{b(k)} \cdot \widehat{\mathbf{h}_j}^{(\ell)}(k)\ ,\ 0 \le l \le L-1 \tag{19}$$

where

$$b^2(k) = \frac{\sum_{t=0}^{T-1}|z_1(t,k)|^2}{\sum_{t=0}^{T-1}\widehat{\sigma_x^2}^{(\ell)}(t,k)}.$$

Applying this procedure, guarantees the preservation of the average spectral profile of the input signal without affecting the convergence of the algorithm.

Finally, we note that in high signal to noise ratio (SNR) scenarios, estimation errors can result in a negative noise variance estimation (17). To circumvent this phenomenon, the noise variance estimate was confined to a small positive value. Details regarding the updated procedure are not given in this paper due to space constraints.

5. SIMULATION RESULTS

The proposed algorithm was evaluated with the following procedure. Clean speech utterances of the same speaker, were drawn from the TIMIT database [12] and concatenated to a sentence of length 32 s. Sampling rate is 16 kHz. These sentences were convolved with four room impulse responses (using an efficient implementation of Allen and Berkley's image method [13, 14]). The distance between adjacent microphones was set to 8 cm. The reverberation time, T_{60}, was set to 700 ms, and the impulse response length is 5000 samples.

The reverberant signals were contaminated by pink noise to obtain reverberated-signal to noise ratio (RSNR) levels of 0, 10, 20, and 30 dB. The RSNR is defined as the ratio of noise-free reverberant signal power and the additive noise power:

$$\text{RSNR} = 10\log_{10}\frac{\sum_{t,k}|z(t,k)-v(t,k)|^2}{\sum_{t,k}|v(t,k)|^2}\,. \qquad (20)$$

The procedure was repeated for four different speakers in all RSNR levels.

The STFT analysis window used was 32 ms Hamming window, with 50% overlap. Higher percentage of overlap will result in a significant dependency between adjacent frames, rendering the statistical model of Sec. 2 inaccurate, and hence leading to performance degradation. The system length L was set to 19 in accordance with the sampling rate, the length of h in the time-domain, the analysis window length, and the overlap percentage. The value of L is short enough to impose a reasonable computational load.

As mentioned in Sec. 4, two alternatives for initializing the clean speech variance $\sigma_x^2(t,k)$ were compared. In the first alternative (denoted Z-init), the instantaneous power of the noisy and reverberant signal z_1 was used . In the second alternative (denoted SE-init), we used the instantaneous power of $\hat{x}_1$ computed using the single-channel SE dereverberation algorithm proposed in [11]. The acoustic system $\mathbf{H}$ was initialized as discussed in 4. Ten EM iterations were performed.

We used two objective measures to evaluate the performance of the proposed algorithm, namely the speech to reverberation modulation energy ratio (SRMR) [15] and the log-spectral distortion (LSD). The LSD distance between x and $\tilde{z} \in \{z, \hat{x}\}$ in frame t is defined as:

$$\text{LSD}(t) = \frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\left(\frac{\max\{x(t,k),\epsilon(x)\}}{\max\{\tilde{z}(t,k),\epsilon(\tilde{z})\}}\right) \qquad (21)$$

where the minimum value is calculated as:

$$\epsilon(y) = 10^{-A_{\text{dB}}/10}\max_{t,k}\{y(t,k)\}$$

and A_{dB} is set to the desired dynamic range, which is chosen in our case to be 60 dB. While reduction in reverberation is measured by higher SRMR, better speech estimate would be indicated by lower LSD values. The LSD and SRMR average results are summarized in Tables 1 and 2, respectively. The clean, the noisy and reverberant, and the KEMD output signals for input RSNR of 20 dB are depicted in Figure 1. Independent of the initialization, the proposed algorithm is able to reduce the LSD and improve the SRMR. Moreover, we can conclude that the best results are obtained when initializing with the SE-init procedure. Unofficial listening tests indicate however that the Z-init procedure yields slightly less distorted output signal.

Table 1. LSD values for $T_{60} = 700$ ms

RSNR (dB)	Input	Z-init	SE-init
0	6.80	4.61	3.85
10	4.79	3.27	2.86
20	3.54	2.83	2.72
30	3.11	2.69	2.99

Table 2. SRMR values for $T_{60} = 700$ ms

RSNR (dB)	Input	Z-init	SE-init
0	1.18	2.88	3.72
10	1.93	2.90	3.47
20	2.16	2.83	3.29
30	2.19	2.80	3.25

6. CONCLUSIONS

An EM algorithm for multi-microphone speech dereverberation was presented. The algorithm converges to the ML estimate of the acoustic parameters. An estimate of the denoised and dereverberated speech signal is obtained (as a byproduct of the algorithm) at the E-step by applying the Kalman smoother. The iterative procedure converges in reasonably low number of iterations. The entire algorithm is applied in the STFT domain, enabling an efficient implementation. Simulation results show that a significant amount of reverberation is reduced, with low speech distortion, as indicated by two commonly used speech dereverberation measures and by the assessment of sample speech sonograms.

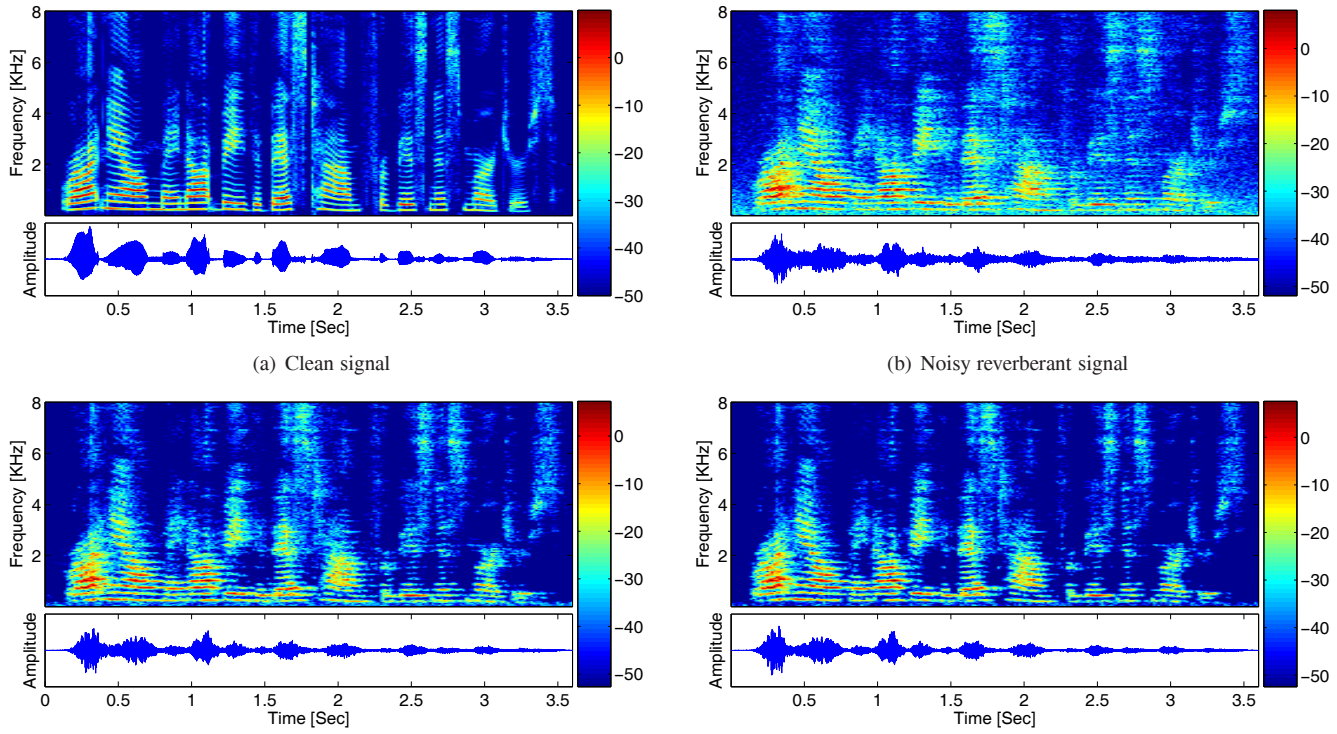
7. REFERENCES

[1] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, pp. 1074–1090, 2003.

[2] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

[3] S. Gannot and M. Moonen, "On the application of the unscented kalman filter to speech processing," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2001.

[4] C. Evers and J. R. Hopgood, "Marginalization of static observation parameters in a rao-blackwellized particle filter with application to sequential blind speech dereverberation," in *European Signal Processing Conference (EUSIPCO)*, 2009, pp. 1437–1441.



(a) Clean signal

(b) Noisy reverberant signal

(c) Output signal (Z-init)

(d) Output signal (SE-init)

Fig. 1. Sonograms and waveforms for $T_{60} = 700$ ms, and a RSNR of 20 dB. In (Z-init), σ_x^2 was initialized using $|z_1|^2$, and in (SE-init) $|\hat{x}_1|^2$ was used instead.

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

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[7] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 17–20.

[8] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[9] E. Weinstein, A. Oppenheim, M. Feder, and J. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Transactions on Signal Processing*, vol. 42, pp. 846–859, Apr. 1994.

[10] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

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

[12] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA*, 1988.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[14] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[15] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Speech and Audio Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.