DUAL-MICROPHONE SPEECH DEREVERBERATION USING GARCH MODELING

Ari Abramson[1], *Emanuël A. P. Habets*[1,2] , *Sharon Gannot*[2] *and Israel Cohen*[1]

[1] Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
{aari@tx,icohen@ee}.technion.ac.il

[2] School of Engineering
Bar-Ilan University
Ramat-Gan, 52900, Israel
{habetse,gannot}@eng.biu.ac.il

ABSTRACT

In this paper, we develop a dual-microphone speech dereverberation algorithm for noisy environments, which is aimed at suppressing late reverberation and background noise. The spectral variance of the late reverberation is obtained with adaptively-estimated direct path compensation. A Markov-switching generalized autoregressive conditional heteroscedasticity (GARCH) model is used to estimate the spectral variance of the desired signal, which includes the direct sound and early reverberation. Experimental results demonstrate the advantage of the proposed algorithm compared to a decision-directed-based algorithm.

Index Terms— speech dereverberation, spectral enhancement, GARCH modeling.

1. INTRODUCTION

In many speech communication systems the received signal is degraded by reverberation, as well as background noise. The reverberant signal consists of a direct sound, early reverberation, and late reverberation. Early reflections mainly contribute to coloration and tend to improve the intelligibility, whereas late reverberation causes a noise-like perception and degrades the fidelity and intelligibility of the speech signal.

Speech dereverberation algorithms can be divided into two classes. Algorithms in the first class are based on estimating and inverting the room impulse response (RIR), e.g., [1]. In the second class, algorithms try to suppress reverberation without estimating the RIR, e.g., [2]. Recently, Habets *et al.* [3] proposed a dual-microphone dereverberation system which is aimed at suppressing late reverberation that results from the tail of the RIR by applying a spectral enhancement approach. A direct path compensation (DPC) is applied to the late reverberant spectral variance estimate to enable better attenuation of the late reverberation with less distortion of the desired signal. However, the parameter of the DPC was evaluated directly from the RIR which is unknown in practice. In addition, the *a priori* signal to noise ratio (SNR) required for the spectral enhancement is estimated by using the traditional decision-directed approach. Recently, the generalized autoregressive conditional heteroscedasticity (GARCH) model with Markov regimes has been shown to be useful for speech enhancement applications [4, 5]. The model takes into account the strong correlation of successive spectral magnitudes, and is more appropriate than the decision-directed approach for speech spectral variance estimation in noisy environments.

This research was supported by the Israel Science Foundation (grant no. 1085/05)

In this paper, we develop an improved dual-microphone speech dereverberation algorithm which relies on a Markov-switching GARCH (MS-GARCH) modeling of the desired early speech component, which consists of the direct sound and early reverberation. The model is applied to distinctive frequency subbands and specifies the volatility clustering of successive spectral coefficients, while a speech-absence state is used for evaluating the speech presence probability. Furthermore, an adaptive approach is developed to estimate the parameter for the DPC directly from the observed signals. Experimental results show that using the MS-GARCH modeling rather than the decision-directed approach, improved results can be obtained. Furthermore, by using the proposed algorithm, the performance obtained with blindly estimated DPC parameter is comparable to that obtained with an optimal DPC parameter that is calculated from the actual RIR, which is unknown in practice.

The paper is organized as follows. In Section 2, we formulate the speech dereverberation problem and briefly review the algorithm proposed in [3]. In Section 3, we derive an adaptive estimator for the DPC parameter. In Section 4 we describe the MS-GARCH model which is used for the desired signal, and in Section 5 we present some experimental results which demonstrate the improved performance of the proposed algorithm.

2. DUAL-MICROPHONE DEREVERBERATION

Consider an $M-$microphone array located in a reverberant environment. Let $\mathbf{a}_m(n) = [a_{m,0}(n), ..., a_{m,L-1}(n)]^T$ denote the RIR at time n from the source signal $s(n)$ to the mth microphone, and let $d_m(n)$ denote the noise component received at the mth microphone. The observed signals are then given by

$$z_m(n) = \mathbf{a}_m^T(n)\,\mathbf{s}(n) + d_m(n) \tag{1}$$

where $\mathbf{s}(n) = [s(n), ..., s(n-L+1)]^T$. The RIR, $\mathbf{a}_m(n)$, can be divided into the direct path and early reflections, denoted by $\mathbf{a}_m^d(n)$, and late reflections, denoted by $\mathbf{a}_m^r(n)$. Accordingly,

$$a_{m,j}(n) = \begin{cases} a_{m,j}^d(n) & 0 \le j < t_r \\ a_{m,j}^r(n) & t_r \le j < L \end{cases}, \tag{2}$$

where t_r is the time where the late reverberation starts (about 40 to 80 ms). Hence, the reverberant signal can be divided into two signals

$$\mathbf{a}_m^T(n)\,\mathbf{s}(n) = x_m(n) + r_m(n)\ , \tag{3}$$

where $x_m(n)$ is the desired early speech component, and $r_m(n)$ denotes the late reverberant component. Applying the short-time

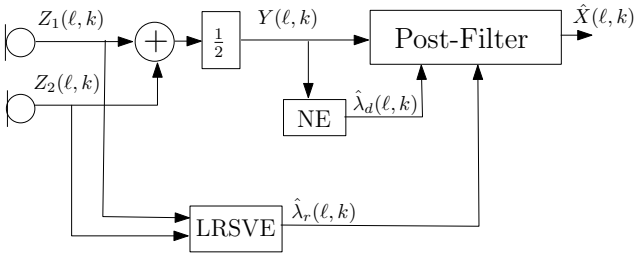

Fig. 1. Dual microphone speech dereverberation system.

Fourier transform (STFT) to the observed signals, we have

$$Z_m(\ell,k) = X_m(\ell,k) + R_m(\ell,k) + D_m(\ell,k)\ , \qquad (4)$$

where ℓ represents the frame index, and k the frequency bin index. At the output of a delay and sum beamformer (DSB) which is steered towards the desired source, we have the time-frequency signal

$$Y(\ell,k) = X(\ell,k) + R(\ell,k) + D(\ell,k)\ . \qquad (5)$$

Habets *et al.* [3] proposed a dual microphone dereverberation algorithm which is aimed at estimating the early speech component. In the system, shown in Figure 1, it is assumed that the arrival times of the direct speech signals are aligned. The lower branch is a late reverberant spectral variance estimator (LRSVE), $\hat{\lambda}_r(\ell,k)$, while the upper branch includes a beamformer, a background noise estimator (NE), $\hat{\lambda}_d(\ell,k)$, and a post-filter. The spectral variance of the noise signal, $\lambda_d(\ell,k)$, can be estimated, e.g., using [6]. The *a priori* SNR

$$\xi(\ell,k) = \frac{\lambda_x(\ell,k)}{\lambda_r(\ell,k) + \lambda_d(\ell,k)} \qquad (6)$$

is estimated using the decision-directed approach [7].

The desired spectral coefficients are estimated by minimizing the mean square error of the log-spectral amplitude (LSA) [8] by assuming two hypotheses, speech presence (H_1) and absence (H_0). The resulting optimally-modified LSA estimator is given by [9]

$$\hat{X}(\ell,k) = G_{H_1}(\ell,k)^{p(\ell,k)}\, G_{H_0}(\ell,k)^{1-p(\ell,k)}\, Y(\ell,k)\ , \qquad (7)$$

where $G_{H_1}(\ell,k)$ is the LSA gain under speech presence [8] and

$$G_{H_0}(\ell,k) = G_{\min}\frac{\hat{\lambda}_d(\ell,k)}{\hat{\lambda}_d(\ell,k) + \hat{\lambda}_r(\ell,k)} \qquad (8)$$

to allow reduction of the late reverberant signal down to the noise floor [3]. In the next subsection we derive an adaptive estimator for the late reverberant spectral variance, and in Section 4 we formulate the MS-GARCH modeling applied for the desired signal. The speech presence probability $p(\ell,k)$ is discussed in Section 4.2.

3. LATE REVERBERANT SPECTRAL ESTIMATION

The spectral variance of the late reverberation at each microphone, $\lambda_{r,m}(\ell,k)$, can be obtained based on Polack's statistical reverberation model of the RIR [3], using an estimate of the spectral variance of the reverberant signal, $\lambda_{b,m}(\ell,k) = E\left\{|X_m(\ell,k) + R_m(\ell,k)|^2\right\}$. Let $T_{60}(k)$ denote the reverberation time of the room in the kth frequency band, let $\delta(k) = 3\ln(10)/T_{60}(k)$, let R denote the frame rate of the STFT, and let $\alpha(k) = \exp\{-2\delta(k)R/f_s\}$. Then, the spectral variance of the late reverberant signal $\lambda_r(\ell,k)$ at the output of the DSB is estimated by

$$\hat{\lambda}_r(\ell,k) = \frac{1}{2}\sum_{m=1}^{2}\alpha(k)^{\frac{t_r}{R}}\,\hat{\lambda}_{b,m}\left(\ell - \frac{t_r}{R}, k\right)\ . \qquad (9)$$

However, to avoid over-estimation of $\lambda_r(\ell,k)$ when the source-microphone distance is smaller than the critical distance (i.e., the energy of the direct path is larger than the energy of all reflections) it was proposed to compensate the over estimation of the spectral variance of the reverberant signal using

$$\begin{aligned}\hat{\lambda}'_{b,m}(\ell) &= \frac{\kappa_m(\ell)}{1+\kappa_m(\ell)}\alpha(k)\hat{\lambda}'_{b,m}(\ell-1,k) \\ &\quad + \frac{1}{1+\kappa_m(\ell)}\hat{\lambda}_{b,m}(\ell,k)\,, \end{aligned} \qquad (10)$$

where $\kappa_m(\ell)$ is a compensation parameter which is related to the direct and reverberant energy at the mth microphone. The compensated estimate $\hat{\lambda}'_{b,m}(\ell)$ is then used in (9) as the spectral variance estimate of the reverberant signal. It was shown in [3] that applying this DPC prevents over-estimation of the late reverberant spectral variance and improves the quality of the output signal. However, the DPC parameter, κ_m, was calculated directly from the presumably known RIR. Here, we propose to estimate the parameter κ_m adaptively. In case κ_m is too large the spectral variance $\hat{\lambda}'_{b,m}(\ell,k)$ could become larger than $\hat{\lambda}_{b,m}(\ell,k)$, which indicates that over-estimation can occur and that the value of κ_m should be decreased. Furthermore, during the free-decay, which occurs after an offset of the source signal, $\hat{\lambda}'_{b,m}(\ell,k)$ should be equal to $\hat{\lambda}_{b,m}(\ell,k)$. Estimation of κ_m could therefore be performed after a speech offset. Unfortunately, the detection of speech offsets is rather difficult. However, we can conclude that κ_m should at least fulfill the following conditions: (i) $\hat{\lambda}_{b,m}(\ell,k) \geq \hat{\lambda}'_{b,m}(\ell,k)$, (ii) when speech is present and $\hat{\lambda}_{b,m}(\ell,k) < \hat{\lambda}'_{b,m}(\ell,k)$ the value of κ_m can be increased, (iii) when $\hat{\lambda}_{b,m}(\ell,k) > \hat{\lambda}'_{b,m}(\ell,k)$ the value of κ_m can be decreased slowly, and (iv) when $\hat{\lambda}_{b,m}(\ell,k) = \hat{\lambda}'_{b,m}(\ell,k)$ the value of κ_m is assumed to be correct. Therefore, we can update $\kappa_m(\ell)$ when speech is present using

$$\hat{\kappa}_m(\ell+1) = \max\left\{\hat{\kappa}_m(\ell) + \mu_\kappa\left(\frac{\sum_k \hat{\lambda}'_{b,m}(\ell,k)}{\sum_k \hat{\lambda}_{b,m}(\ell,k)} - 1\right),\ 0\right\}, \qquad (11)$$

where μ_κ ($0 < \mu_\kappa < 1$) denotes the step-size.

4. MODELING EARLY REVERBERATION USING GARCH

Speech signals are characterized by time-varying energy levels and volatility. The spectral coefficients of the speech signal can be effectively characterized using an MS-GARCH model [4, 5]. The GARCH parameters specify the volatility of the spectral coefficients, and the Markovian regimes allow the model to switch between different sets of GARCH parameters. Let $q_\ell \in \{0,...,Q\}$ denote the active state of a first-order Markov chain at frame ℓ with known state-transition probabilities. Let $\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1)$ denote the conditional spectral variance of the desired signal $X(\ell,k)$ conditioned on q_ℓ and on all information up to previous frame, and let $\{V(\ell,k)\}$ be iid complex Gaussian random variables with zero-mean and unit variance. We assume that the spectral coefficients of the desired signal follow an MS-GARCH model [4], i.e., given q_ℓ

$$X(\ell,k) = \sqrt{\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1)}\,V(\ell,k) \qquad (12)$$

where

$$\lambda_{x,q_\ell}(\ell,k\,|\,\ell-1)=\lambda_{\min,q_\ell}+\alpha_{q_\ell}\left|X(\ell-1,k)\right|^2 + \beta_{q_\ell}\left[\lambda_{x,q_{\ell-1}}(\ell-1,k\,|\,\ell-2)-\lambda_{\min,q_{\ell-1}}\right] \quad (13)$$

with $\lambda_{\min,q_\ell}>0$ and $\alpha_{q_\ell},\,\beta_{q_\ell}\geq 0$ for $q_\ell=0,...,Q$. As can be seen from (12) and (13), the conditional spectral variances of successive frames at a specific frequency bin are strongly correlated. However, given the sequence of the conditional spectral variances and the active states, the spectral coefficients $\{X(\ell,k)\}$ are statistically independent. It was shown that the spectral variance estimation resulting from this model is a generalization of the decision-directed estimator with improved tracking of the speech spectral volatility [4].

4.1. Spectral Variance Estimation

Let $\mathcal{Y}^\ell=\{Y(l,k)\mid l\leq\ell\}$ denote the set of the observed spectral coefficients up to frame ℓ. Given $\mathcal{Y}^\ell$ the set of conditional spectral variances can be recursively estimated using a propagation step

$$\begin{aligned}\hat{\lambda}_{x,q_\ell}&(\ell,k\,|\,\ell-1)\\ &=\lambda_{\min,q_\ell}+\alpha_{q_\ell}E\left\{\left|X(\ell-1,k)\right|^2\mid\mathcal{Y}^{\ell-1},q_\ell\right\}\\ &\quad+\beta_{q_\ell}E\left\{\lambda_x(\ell-1,k\,|\,\ell-2)\mid\mathcal{Y}^{\ell-1},q_\ell\right\}\\ &\quad-\beta_{q_\ell}E\left\{\lambda_{\min,q_{\ell-1}}\mid\mathcal{Y}^{\ell-1},q_\ell\right\} \end{aligned} \quad (14)$$

and an update step

$$\begin{aligned}&E\left\{\left|X(\ell-1,k)\right|^2\mid\mathcal{Y}^{\ell-1},q_\ell\right\}\\ &=\sum_{q_{\ell-1}}p\left(q_{\ell-1}\,|\,\mathcal{Y}^{\ell-1},q_\ell\right)E\left\{\left|X(\ell-1,k)\right|^2\mid\mathcal{Y}^{\ell-1},q_{\ell-1}\right\}\\ &\triangleq\sum_{q_{\ell-1}}p\left(q_{\ell-1}\,|\,\mathcal{Y}^{\ell-1},q_\ell\right)\hat{\lambda}_{x,q_{\ell-1}}(\ell-1,k\,|\,\ell-1)\,. \end{aligned} \quad (15)$$

A detailed estimation algorithm is given in [4]. The estimate of the spectral variance of the desired signal is then obtained by

$$\hat{\lambda}_x(\ell,k)=\sum_{q_\ell}p\left(q_\ell\,|\,\mathcal{Y}^\ell\right)\hat{\lambda}_{x,q_\ell}(\ell,k\,|\,\ell)\,. \quad (16)$$

Note that although the spectral variance is specified for each frequency bin independently, the Markovian state is frequency-independent. However, since different frequency bands of speech signals are characterized by different energy level and volatility, it was proposed in [5] to apply the model independently to distinctive subbands. Furthermore, a simple model estimation approach was proposed such that each state represents different energy level, and a specific state specifies signal absence. However, in our case the desired signal contains early reverberation such that the spectral variance at speech offsets has smoother decay than in case of a nonreverberant signal. Consequently, an immediate transition from a state which represents high spectral energy to a state which represents very low energy would not be expected. Therefore, the state transition probabilities are set such that the probability for a progressive state-transition is much higher than the probability for an immediate transition from the higher energy level to the lower.

4.2. Speech Presence Probability

The *posteriori* speech presence probability, $p(\ell,k)$, required for (7) is originally calculated [9] based on a Gaussian model from the *a priori* speech presence probability. The latter is evaluated based on the time-frequency distribution of the *a priori* SNR, $\xi(\ell,k)$. For a multi-sensor system, it was proposed in [3] to exploit the spatial information and to use additional parameter $P_{spatial}(\ell,k)$ for the *a priori* probability which is evaluated based on the spatial coherence between the microphone signals. In our case, the multi-state model for the speech spectral coefficients inherently results in a conditional probability for each state. Having a specific state for speech absence (say $q_\ell=0$), we obtain a speech presence probability for each subband in each frame, $p\left(q_\ell\neq 0\,|\,\mathcal{Y}^\ell\right)$. Accordingly, we define

$$P_{sb}(\ell,k)=\begin{cases}p_h & p\left(q_\ell\neq 0\,|\,\mathcal{Y}^\ell\right)>T_h\\ p_l & p\left(q_\ell\neq 0\,|\,\mathcal{Y}^\ell\right)<T_l\\ p\left(q_\ell\neq 0\,|\,\mathcal{Y}^\ell\right) & \text{otherwise}\end{cases} \quad (17)$$

where $p_l\leq T_l\leq T_h\leq p_h$ are constrain parameters for the subband speech presence probability. The subband probability, $P_{sb}(\ell,k)$, is employed as an additional multiplicative parameter for the evaluation of the *a priori* speech presence probability. Note that although we do not use a specific index for the subband, $p\left(q_\ell\neq 0\,|\,\mathcal{Y}^\ell\right)$ is calculated for each subband independently, and therefore $P_{sb}(\ell,k)$ includes also a frequency bin index.

5. EXPERIMENTAL RESULTS

In our experimental study, we consider synthetic RIRs which were generated using the *image* method. The speech signals, sampled at 8 kHz, include male and female speakers, each of 20 seconds. A moderate level of white Gaussian noise was added to each of the microphone signals. The distance between the two microphones is 0.15 meter, and the source-to-microphone distance was set to 0.5 and 1 meter (which are both smaller than the critical distance). While applying the MS-GARCH model, the model parameters are estimated from the noisy signal as proposed in [5].

Segmental signal to interference ratio (SegSIR) and log spectral distortion (LSD) are used to evaluate the performance of the proposed algorithm, as well as informal listening tests and inspection of spectrograms. For the quality measures, the direct sound signal was used as the reference signal. Figure 2 shows experimental results of the proposed algorithm as a function of the number of GARCH states, and for several reverberation times. The input SNR is 15 dB and the source to microphone distance is 0.5 m. It can be seen that the performance improves monotonically with the growth of the number of states, but, the most significant improvement is achieved by using up to 3 Markovian states.

Table 1 compares the performance of the proposed algorithm with that of the original algorithm [3] which employs a decision-directed estimator for the *a priori* SNR. The reverberation time is $T_{60}=0.5$ sec, and the proposed algorithm was applied with 3-state MS-GARCH model. In both algorithms, the DPC parameters κ_1 and κ_2 are blindly estimated adaptively, as proposed in Section 3, and the results shown in parentheses are obtained using the optimal values which are evaluated from the actual RIRs. It can be seen that the GARCH modeling is more advantageous than the decision-directed approach, and the blindly estimated DPC parameters yield results which are comparable to using the optimal value.

In Figure 3 spectrogram and waveform of a noisy signal are shown with input SNR of 20 dB and a source to microphone distance of 1 m. The smearing caused by the late reverberation and the background noise are reduced. Wave files are available online at: *http://siglab.technion.ac.il/~ari_a/Audio_demos.htm*.

Table 1. SegSIR and LSD obtained by using the decision-directed approach and the proposed MS-GARCH-based approach. $T_{60} = 0.5$ sec. In parentheses - results using optimal DPC parameters.

	d=0.5 m, SNR=15 dB		d=0.5 m, SNR=20 dB		d=1 m, SNR=15 dB		d=1 m, SNR=20 dB	
	SegSIR [dB]	LSD [dB]	SegSIR [dB]	LSD [dB]	SegSIR [dB]	LSD [dB]	SegSIR [dB]	LSD [dB]
Unprocessed	5.849	4.875	7.284	2.681	2.295	6.379	2.864	4.578
Decision-directed	8.359	1.995	8.745	1.744	4.289	3.583	4.385	3.482
	(8.783)	(1.825)	(9.230)	(1.535)	(4.452)	(3.455)	(4.578)	(3.333)
MS-GARCH	9.010	1.700	9.392	1.493	4.551	3.521	4.654	3.442
	(9.265)	(1.606)	(9.715)	(1.367)	(4.941)	(3.390)	(5.110)	(3.298)

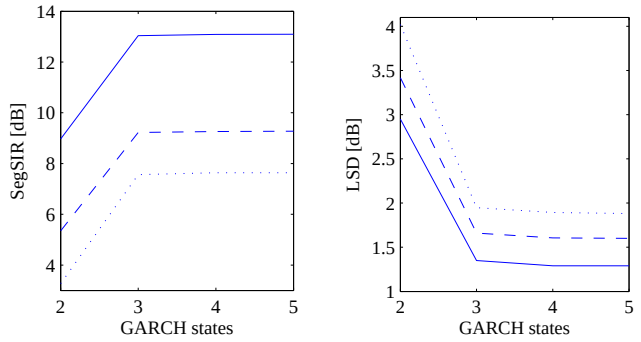

Fig. 2. SegSIR and LSD as functions of the number of GARCH states (solid line: $T_{60} = 0.25$ sec, dashed line: $T_{60} = 0.5$ sec, and dotted line: $T_{60} = 0.75$ sec).

6. CONCLUSIONS

We have developed a dual-microphone speech dereverberation algorithm for noisy environments which is based on MS-GARCH modeling of the desired early speech component. The spectral variance of the late reverberation is estimated from the observed signals while compensating for the energy of the direct path. The algorithm blindly operates in noisy and reverberant environments without any knowledge of the RIR, except for the reverberation time, which can be obtained blindly using, e.g., [10]. It is shown that compared to the original algorithm which employs the decision-directed estimator [3], improved performance is obtained with little distortion to the desired signal.

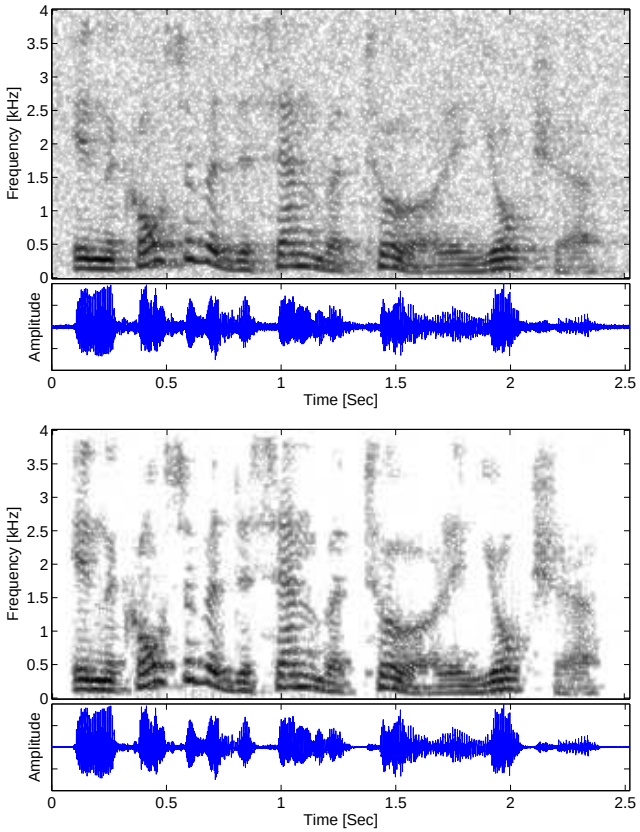

Fig. 3. Spectrograms and waveforms of a noisy and reverberated speech signal (top), and the processed signal (bottom).

7. REFERENCES

[1] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[2] N. Gaubitch and P. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *Proc. of the 15th International Conference on Digital Signal Processing (DSP 2007)*, July 2007, pp. 607–610.

[3] E. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. 6th IEEE Int. Symposium on Signal Process. and Information Technology, ISSPIT-2006*, Vancouver, Canada, Aug. 2006, pp. 651–655.

[4] A. Abramson and I. Cohen, "Recursive supervised estimation of a Markov-switching GARCH process in the short-time Fourier transform domain," *IEEE Trans. Signal Processing*, vol. 55, no. 7, pp. 3227–3238, July 2007.

[5] ——, "Markov-switching GARCH model and application to speech enhancement in subbands," in *Proc. Int. Workshop on Acoust. Echo and Noise Control, IWAENC-06*, Paris, France, Sept. 2006, paper no. 7, pp. 1–4.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[8] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary environments," *Signal Processing*, vol. 81, pp. 2403–2418, Nov. 2001.

[10] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien Jr., C. Lansing, and A. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.