SPEECH DEREVERBERATION USING BACKWARD ESTIMATION OF THE LATE REVERBERANT SPECTRAL VARIANCE

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

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

[2] Technion, Israel Institute of Technology
Department of Electrical Engineering
Technion City, Haifa 32000, Israel
icohen@ee.technion.ac.il

ABSTRACT

In speech communication systems the received microphone signals are degraded by room reverberation and ambient noise. This signal degradation can decrease the fidelity and intelligibility of the desired speaker. Reverberant speech can be separated into two components, viz. an early speech component and a late reverberant speech component. Reverberation suppression algorithms, that are feasible in practice, have been developed to suppress late reverberant speech or in other words to estimate the early speech component. The main challenge is to develop an estimator for the so-called late reverberant spectral variance (LRSV). In this contribution a generalized statistical reverberation model is proposed that can be used to estimate the LRSV. Novel and existing estimators can be derived from this model. One novel estimator is a so-called backward estimator that uses an estimate of the early speech component to obtain an estimate of the LRSV. Advantages and possible disadvantages of the estimators are discussed, and experimental results using simulated reverberant speech are presented.

1. INTRODUCTION

In general, acoustic signals radiated within a room are linearly distorted by reflections from walls and other objects. These distortions degrade the fidelity and intelligibility of the desired speaker, and the recognition performance of automatic speech recognition systems. In general, the degradation increases when the distance between the source and the microphone increases. One effect of reverberation on speech is the lengthening of speech phonemes. Consequently, reverberation of one phoneme overlaps subsequent phonemes. Evidence has been found that this phenomenon, which is referred to as *overlap-masking*, decreases speech intelligibility [1].

Reverberation reduction methods are generally divided into two categories. Methods of the first category are known as reverberation cancellation methods. In general, a linear filter operation is applied to the observed microphone signals to obtain an estimate of the anechoic signal. The filters are either estimated directly from the observed signals or indirectly using an estimate of the acoustic impulse responses (AIRs) of the acoustic channels between the source and the microphones. Methods in the second category are known as reverberation suppression methods. These methods commonly apply a non-linear operation to the observed microphone signals to suppress reverberation and require little or no *a priori* knowledge about the AIRs. In both categories single and multiple microphone signals are exploited. While multi-microphone cancellation methods can achieve perfect dereverberation, suppression methods can only achieve partial dereverberation.

Here we focus on a specific single microphone reverberation suppression technique that suppresses late reverberation or in other words estimates the early speech component. Reverberation is suppressed using spectral enhancement that is performed in the short-time Fourier transform (STFT) domain. In order to perform spectral enhancement an estimate of the short-term power spectral density (or in the context of statistical spectral enhancement methods, spectral variance) of the interference, i.e., the late reverberant speech component, is required. The main challenge is to estimate the spectral variance of the late reverberant signal from the reverberant microphone signal. In the last decade several late reverberant spectral variance estimators have been developed [2, 3, 4]. While some are very heuristic, others are based on statistical room acoustic models that are usually formulated in the time domain.

In this contribution we propose a generalized statistical reverberation model in the STFT domain. Using this model we can derive novel and existing estimators. Currently, all estimators are so-called forward estimators, i.e., they use the reverberant microphone signal to estimate the spectral variance of the late reverberant signal. Using the proposed model, we derive a so-called backward estimator that uses the estimated early speech component.

The paper is organized as follows: In Section 2 the problem is formulated. In Section 3 we propose a generalized statistical reverberation model in the STFT domain. This model is used in Section 4 to derive forward and backward estimators for the late reverberant spectral variance. In Section 5 we show how the early speech component can be estimated given the late reverberant spectral variance. Experimental results that demonstrate the performance of the forward and backward estimators are presented in Section 6. Finally, conclusions are provided in Section 7.

2. PROBLEM FORMULATION

The reverberant signal results from the convolution of the anechoic speech signal $s(n)$ and a causal AIR $h(n)$. Here we assume that the AIR is time-invariant and that its length is infinite. The reverberant speech signal at discrete-time n can be written as

$$z(n) = \sum_{n'=0}^{\infty} h(n)\, s(n-n'). \tag{1}$$

The observed microphone signal is given by

$$x(n) = z(n) + v(n), \tag{2}$$

where $v(n)$ denote the ambient noise. In this contribution we assume that $v(n) = 0$ for all n.

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

In the STFT domain the signal $s(n)$ is given by

$$S(\ell,k) = \sum_{n=-\infty}^{\infty} s(n)\,\tilde{\psi}(n-\ell R)\,\mathrm{e}^{-j\frac{2\pi}{N}k(n-\ell R)}, \qquad (3)$$

where ℓ is the frame index, k is the frequency band index, R is the discrete time shift, and $\tilde{\psi}(m)$ denotes the analysis window of length N. Subsequently we can express $z(n)$ in the STFT domain as [5]

$$Z(\ell,k) = \sum_{k'=0}^{N-1}\sum_{\ell'=-\infty}^{\infty} H(\ell',k,k')\,S(\ell-\ell',k'), \qquad (4)$$

where k and k' denote the band and cross-band frequency bin indices, respectively. The STFT response $H(\ell',k,k')$ is related to impulse response $h(n)$ by

$$H(\ell',k,k') = \big(h(n) * \vartheta(n,k,k')\big)\big|_{n=\ell' R}, \qquad (5)$$

where $*$ denotes convolution with respect to n. The function $\vartheta(n,k,k')$ is related to the analysis window $\tilde{\psi}(m)$ and the synthesis window $\psi(m)$ of length N:

$$\vartheta(n,k,k') \triangleq \mathrm{e}^{j\frac{2\pi}{N}k'n}\sum_{n'=-\infty}^{\infty}\tilde{\psi}(n')\psi(n'+n)\mathrm{e}^{-j\frac{2\pi}{N}n'(k-k')}. \qquad (6)$$

The STFT response $H(\ell',k,k')$ may be interpreted as a response to an impulse $\delta(\ell',k-k')$ in the time-frequency domain.

To simplify the following discussion, and without loss of generality, it is assumed that the direct sound arrives at time instance n. Since our objective is to suppress late reverberation we split the AIR into two components such that

$$H(\ell,k,k') = \begin{cases} 0 & \text{for } \ell < 0; \\ H_{\mathrm{e}}(\ell,k,k') & \text{for } 0 \le \ell < N_{\mathrm{e}}; \\ H_{\ell}(\ell,k,k') & \text{for } N_{\mathrm{e}} \le \ell \le \infty, \end{cases} \qquad (7)$$

where $H_{\mathrm{e}}(\ell,k,k')$ models the direct path and a few early reflections and $H_{\ell}(\ell,k,k')$ models all later reflections, and N_{e} ($N_{\mathrm{e}} \ge 1$) specifies the time instance (measured with respect to the arrival time of the direct sound) from where the late reverberation starts. This parameter can be specified by the design specifications or controlled by the listener depending on the subjective preference.

Using (7) we can write the microphone signal $X(\ell,k)$ as

$$X(\ell,k) = \sum_{k'=0}^{N-1}\sum_{\ell'=-\infty}^{N_{\mathrm{e}}-1} H(\ell',k,k')\,S(\ell-\ell',k') + \sum_{k'=0}^{N-1}\sum_{\ell'=N_{\mathrm{e}}}^{\infty} H(\ell',k,k')\,S(\ell-\ell',k') + V(\ell,k), \qquad (8)$$

where $V(\ell,k)$ denotes the additive ambient noise component. We can write (8) as

$$X(\ell,k) = Z_{\mathrm{e}}(\ell,k) + Z_{\ell}(\ell,k) + V(\ell,k), \qquad (9)$$

where $Z_{\mathrm{e}}(\ell,k)$ denotes the early spectral speech component and $Z_{\ell}(\ell,k)$ denotes the late reverberant spectral speech component.

Now our objective is to derive an algorithm that estimates the early speech component $Z_{\mathrm{e}}(\ell,k)$. We can estimate $Z_{\mathrm{e}}(\ell,k)$ using a spectral enhancement technique if we know the late reverberant spectral variance $\lambda_{z_\ell}(\ell,k) = \mathcal{E}\left\{|Z_{\ell}(\ell,k)|^2\right\}$. Ideally, we would require $H(\ell,k,k')$ to estimate $\lambda_{z_\ell}(\ell,k)$. In practice $H(\ell,k,k')$ is not ***a priori*** known and blindly estimating $H(\ell,k,k')$ remains a difficult task. In order to avoid the need of estimating $H(\ell,k,k')$ we propose a statistical model for $H(\ell,k,k')$ that depends on a small set of parameters $\Theta(\ell,k)$. This statistical model is then used to derive an estimator for $\lambda_{z_\ell}(\ell,k)$. A block diagram that describes the complete reverberation suppression system is depicted in Figure 1. It should be noted that the so-called spectral coloration that is caused by the early reflections cannot be reduced by the proposed suppression method.

3. GENERALIZED STATISTICAL REVERBERATION MODEL IN THE STFT DOMAIN

We propose a novel generalized statistical model in the STFT domain[1]:

$$H(\ell,k,k') = \begin{cases} B_{\mathrm{d}}(\ell,k,k') & \text{for } \ell = 0; \\ B_{\mathrm{r}}(\ell,k,k')\,\mathrm{e}^{-\alpha(k,k')\ell R} & \text{for } \ell \ge 1, \end{cases} \qquad (10)$$

where $\alpha(k,k')$ denotes the decay rate that is related to the reverberation time, and $B_{\mathrm{d}}(\ell,k,k')$ and $B_{\mathrm{r}}(\ell,k,k')$ are zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables. Let us define $\beta_{\mathrm{d}} = \mathcal{E}\{|B_{\mathrm{d}}(\ell,k,k')|^2\}$ and $\beta_{\mathrm{r}} = \mathcal{E}\{|B_{\mathrm{r}}(\ell,k,k')|^2\}$. Accordingly we have:

1. $\mathcal{E}\{H(\ell_1,k_1,k_1')H^*(\ell_2,k_2,k_2')\} = 0$ for $\ell_1 \neq \ell_2$ and $\forall k_1,k_1',k_2,k_2'$,
2. $\mathcal{E}\{H(\ell,k_1,k_1')H^*(\ell,k_2,k_2')\} = 0$ for $k_1 \neq k_2$ and $\forall k_1',k_2'$,
3. $\mathcal{E}\left\{|H(\ell,k,k')|^2\right\} = 0$ for $k \neq k'$,

where $(\cdot)^*$ denotes complex conjugation. It is extremely interesting to note that different realization of $H(\ell,k,k')$ can be interpreted as different spatial observation (i.e., at different source and/or microphone positions) in the enclosure.

Now we can calculate the spectral variance (also known as spectral envelope) in the STFT domain

$$\lambda_h(\ell,k) \triangleq \mathcal{E}\{|H(\ell,k,k')|^2\} = \begin{cases} \beta_{\mathrm{d}} & \text{for } \ell = 0; \\ \beta_{\mathrm{r}}\,\mathrm{e}^{-2\alpha(k)\ell R} & \text{for } \ell \ge 1, \end{cases} \qquad (11)$$

where and $\alpha(k)$ is linked to the frequency dependent reverberation time $T_{60}(k)$ through

$$\alpha(k) \triangleq \frac{3\,\log_{\mathrm{e}}(10)}{T_{60}(k)\,f_{\mathrm{s}}}, \qquad (12)$$

where f_{s} denotes the sampling frequency in Hz.

4. LATE REVERBERANT SPECTRAL VARIANCE ESTIMATORS

In the following we assume that the spectral coefficients of the speech signal can be modelled as a zero-mean i.i.d. complex random variable with a certain distribution and variance $\lambda_s(\ell,k)$. Using (8)

[1] This model is closely related to the time-domain model proposed in [4].

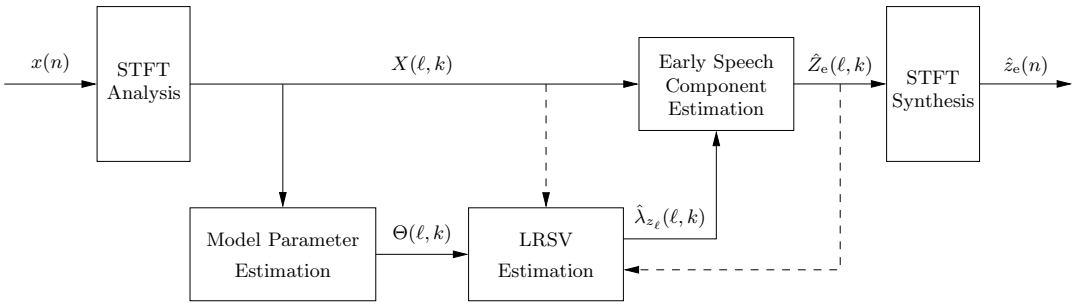


Fig. 1. Block diagram of the complete reverberation suppression system, which consists of a STFT analysis block, model parameter estimator, late reverberant spectral variance (LRSV) estimator, early speech component estimator, and STFT synthesis block.

and the statistical model proposed in (10), we can express the late reverberant spectral variance $\lambda_z(\ell,k) = \mathcal{E}\left\{|Z(\ell,k)|^2\right\}$ as

$$\begin{aligned}\lambda_z(\ell,k) &= \mathcal{E}\left\{\left|\sum_{k'=0}^{K}\sum_{\ell'=0}^{\infty} H(\ell',k,k')\,S\left(\ell-\ell',k\right)\right|^2\right\}\\ &= \sum_{\ell'=0}^{\infty}\lambda_h(\ell',k)\,\lambda_s(\ell-\ell',k)\\ &= \sum_{\ell'=1}^{\infty}\beta_{\mathrm{r}}\,\mathrm{e}^{-2\alpha(k)R\ell'}\,\lambda_s(\ell-\ell',k)+\beta_{\mathrm{d}}\,\lambda_s(\ell,k). \quad (13)\end{aligned}$$

The spectral variance $\lambda_z(\ell,k)$ can also be divided into two components, viz. the spectral variances of the early and late reverberant speech components:

$$\begin{aligned}\lambda_z(\ell,k) &= \sum_{\ell'=0}^{N_{\mathrm{e}}-1}\lambda_h(\ell',k)\,\lambda_s(\ell-\ell',k)\\ &\quad+\sum_{\ell'=N_{\mathrm{e}}}^{\infty}\lambda_h(\ell',k)\,\lambda_s(\ell-\ell',k) \quad (14)\\ &= \lambda_{z_{\mathrm{e}}}(\ell,k)+\lambda_{z_\ell}(\ell,k).\end{aligned}$$

The critical distance is defined as the distance at which the energy of the direct path is equal to the energy of all reflections. When the source-microphone distance is larger than approximately twice the critical distance the contribution of the direct-path energy can be neglected. In this case we can further simplify (11) using $\beta \triangleq \beta_{\mathrm{d}} = \beta_{\mathrm{r}}$. We can then rewrite (13) as

$$\lambda_z(\ell,k) = \mathrm{e}^{-2\alpha(k)R}\,\lambda_z(\ell-1,k)+\beta\,\lambda_s(\ell,k). \quad (15)$$

4.1. Forward Estimator

Using (14) and (15) we can derive the estimator for $\lambda_{z_\ell}(\ell,k)$:

$$\hat{\lambda}^{\mathrm{FE}}_{z_\ell}(\ell,k) = \mathrm{e}^{-2\alpha(k)RN_{\mathrm{e}}}\hat{\lambda}_z(\ell-N_{\mathrm{e}},k). \quad (16)$$

The spectral variance $\lambda_z(\ell,k)$ is estimated by

$$\hat{\lambda}_z(\ell,k) = \eta\,\hat{\lambda}_z(\ell-1,k)+(1-\eta)\,|Z(\ell,k)|^2, \quad (17)$$

where η ($0 \le \eta < 1$) denotes the smoothing factor. In case $V(\ell,k) \neq 0$ we first need to estimate the spectral variance $\lambda_z(\ell,k)$ before we can estimate the late reverberant spectral variance.

We will refer to the estimator in (16) as the ***forward estimator*** since it depends only on the spectral variance $\lambda_z(\ell,k)$ of the received reverberant microphone signal. The same estimator was derived in [3, 6] using a statistical reverberation model in the time-domain. Since we have derived the estimator directly in the STFT domain the assumptions are clear and the derivation has become significantly simpler.

From (16) we can see that the late reverberant spectral variance will be overestimated when the decay rate $\alpha(k)$ is underestimated (i.e., the reverberation $T_{60}(k)$ is overestimated). When the overestimated spectral variance $\hat{\lambda}^{\mathrm{FE}}_{z_\ell}(\ell,k)$ is used to estimate the early speech component, audible distortions might be introduced. Underestimation of $\lambda_{z_\ell}(\ell,k)$ results in less reverberation suppression but will not cause any audible distortions.

4.2. Backward Estimator

Using the previous expressions we note that $\lambda_{z_\ell}(\ell,k)$ is also given by

$$\begin{aligned}\lambda_{z_\ell}(\ell,k) &= \sum_{\ell'=N_{\mathrm{e}}}^{\infty}\lambda_h(\ell',k)\,\lambda_s(\ell-\ell',k)\\ &= \sum_{\ell'=0}^{\infty}\beta\,\mathrm{e}^{-2\alpha(k)R(\ell'+N_{\mathrm{e}})}\lambda_s(\ell-\ell'-N_{\mathrm{e}},k)\\ &= \mathrm{e}^{-2\alpha(k)R}\lambda_{z_\ell}(\ell-1,k)+\beta\,\mathrm{e}^{-2\alpha(k)RN_{\mathrm{e}}}\lambda_s(\ell-N_{\mathrm{e}},k). \quad (18)\end{aligned}$$

Obviously, $\beta\lambda_s(\ell-N_{\mathrm{e}},k)$ is unobservable. However, the spectral variance $\hat{\lambda}_{z_{\mathrm{e}}}(\ell-N_{\mathrm{e}},k)$ can be used as an estimate of $\beta\lambda_s(\ell-N_{\mathrm{e}},k)$ instead. Therefore, we have

$$\hat{\lambda}^{\mathrm{BE}}_{z_\ell}(\ell,k) = \mathrm{e}^{-2\alpha(k)R}\lambda_{z_\ell}(\ell-1,k)+\mathrm{e}^{-2\alpha(k)RN_{\mathrm{e}}}\hat{\lambda}_{z_{\mathrm{e}}}(\ell-N_{\mathrm{e}},k). \quad (19)$$

We will refer to the estimator in (19) as the ***backward estimator*** that depends only on the estimated spectral variance $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$ of the early speech spectral component $Z_{\mathrm{e}}(\ell,k)$. In the presence of ambient noise the early speech component can be estimated given an estimate of the spectral variance of the ambient noise. In the latter case the backward estimator might be advantageous since it circumvents the need to estimate the spectral variance $\lambda_z(\ell,k)$, which is required when the forward estimator is used.

Due to the recursive nature of the backward estimator and the fact that $\hat{\lambda}_{z_{\mathrm{e}}}(\ell,k)$ depends on $\hat{\lambda}^{\mathrm{BE}}_{z_\ell}(\ell,k)$, the effect of underestimating $\alpha(k)$ is more complex compared to the forward estimator. Here

we make the following observations: Firstly, when $\alpha(k)$ is underestimated the spectral variance $\lambda_{z_\ell}(\ell,k)$ will be overestimated after a speech offset. Secondly, it should be noted that it is likely that $\hat{\lambda}_{z_\mathrm{e}}(\ell,k)$ becomes smaller than its true value in case $\lambda_{z_\ell}(\ell,k)$ is overestimated, therefore the error in the last term of (19) is reduced. Thirdly, during periods of silence the error reduces to zero. Due to the sparseness of the signal in the STFT domain these periods of silence occur more frequently in the subband signals than in the fullband signal. In practice, the overestimation can be detected and counter measures can be taken to adjust $\alpha(k)$.

5. EARLY SPEECH COMPONENT ESTIMATOR

Using statistical signal processing, the spectral enhancement problem can be formulated as deriving an estimator $\hat{Z}_\mathrm{e}(\ell,k)$ for the speech spectral coefficients such that the expected value of a certain distortion measure is minimized [7].

We can calculate an estimator for $Z_\mathrm{e}(\ell,k)$ which minimizes the expected value of the distortion measure given the estimated early speech spectral variance $\hat{\lambda}_{z_\mathrm{e}}(\ell,k) = \mathcal{E}\{|\hat{Z}_\mathrm{e}(\ell,k)|^2\}$, the estimated late reverberant spectral variance $\hat{\lambda}_{z_\ell}(\ell,k) = \mathcal{E}\{|\hat{Z}_\ell(\ell,k)|^2\}$ and the spectral coefficient $X(\ell,k)$:

$$\hat{Z}_\mathrm{e}(\ell,k) = \underset{\hat{Z}_\mathrm{e}(\ell,k)}{\operatorname{argmin}}\, \mathcal{E}\left\{ d\left(Z_\mathrm{e}(\ell,k), \hat{Z}_\mathrm{e}(\ell,k) \right) \right\}. \tag{20}$$

One frequently used distortion measure is the squared error distortion measure, i.e.,

$$d\left(Z_\mathrm{e}(\ell,k), \hat{Z}_\mathrm{e}(\ell,k) \right) = \left| g(\hat{Z}_\mathrm{e}(\ell,k)) - g\left(Z_\mathrm{e}(\ell,k)\right) \right|^2, \tag{21}$$

where $g(\cdot)$ is a specific function that determine the fidelity criterion of the estimator. For the squared error distortion measure, the estimator $\hat{Z}_\mathrm{e}(\ell,k)$ is calculated from

$$g(\hat{Z}_\mathrm{e}(\ell,k)) = \mathcal{E}\left\{ g(Z_\mathrm{e}(\ell,k)) \middle| X(\ell,k), \hat{\lambda}_{z_\mathrm{e}}(\ell,k), \hat{\lambda}_{z_\ell}(\ell,k) \right\}. \tag{22}$$

While there are many fidelity criteria it was recently found that the MMSE of the root spectral amplitude provides a good tradeoff between speech distortion, musical noise and noise reduction [8]. The corresponding fidelity criteria is given by

$$g(Q(\ell,k)) = |Q(\ell,k)|^{0.5}, \tag{23}$$

with $Q(\ell,k) \in \{Z_\mathrm{e}(\ell,k), \hat{Z}_\mathrm{e}(\ell,k)\}$.

The MMSE estimator is obtained by substituting (23) into (22). Using a super-Gaussian model for the spectral coefficients, the so-called SuGAR gain function yields [8]

$$G(\ell,k) = \frac{\sqrt{\zeta(\ell,k)}}{\gamma(\ell,k)} \left[\frac{\Gamma(0.75)}{\Gamma(.5)} \frac{{}_1F_1(0.25,1;-\zeta(\ell,k))}{{}_1F_1(0.5,1;-\zeta(\ell,k))} \right]^2, \tag{24}$$

where $\Gamma(\cdot)$ denotes the complete Gamma function, ${}_1F_1(a,b;x)$ denotes the confluent hypergeometric function, $\xi(\ell,k)$ denote the ***a priori*** signal to interference ratio (SIR),

$$\xi(\ell,k) = \frac{\lambda_{z_\mathrm{e}}(\ell,k)}{\lambda_{z_\ell}(\ell,k)}, \tag{25}$$

$\gamma(\ell,k)$ denote the ***a posteriori*** SIR,

$$\gamma(\ell,k) = \frac{|X(\ell,k)|^2}{\lambda_{z_\ell}(\ell,k)}, \tag{26}$$

and

$$\zeta(\ell,k) = \frac{\xi(\ell,k)}{1+\xi(\ell,k)}\ \gamma(\ell,k). \tag{27}$$

While the ***a posteriori*** SIR can be calculated directly using the forward or backward estimator, the ***a priori*** SIR cannot be estimated directly because the early speech spectral variance $\lambda_{z_\mathrm{e}}(\ell,k)$ in (25) is unobservable. The estimation of the ***a priori*** SIR is obtained using the so-called decision-directed approach proposed in [9].

To avoid speech distortions a lower bound, denoted by $G_{\min}$, is applied to $G(\ell,k)$. An estimate of the early spectral speech component $Z_\mathrm{e}(\ell,k)$ can now be obtained by applying the constraint gain function to the reverberant spectral coefficient $X(\ell,k)$, i.e.,

$$\hat{Z}_\mathrm{e}(\ell,k) = \max(G(\ell,k), G_{\min})\ X(\ell,k). \tag{28}$$

Finally, given the estimated spectral component $\hat{Z}_\mathrm{e}(\ell,k)$ the early speech component $\hat{z}_\mathrm{e}(n)$ can be obtained using the inverse STFT,

$$\hat{z}_\mathrm{e}(n) = \sum_\ell \sum_{k=0}^{N-1} \hat{Z}_\mathrm{e}(\ell,k)\psi(n-\ell R)\mathrm{e}^{j\frac{2\pi}{N}k(n-\ell R)}, \tag{29}$$

where $\psi(n)$ is a synthesis window that satisfy the so-called completeness condition:

$$\sum_\ell \tilde{\psi}(n-\ell R)\psi(n-\ell R) = \frac{1}{N} \quad \text{for all } n. \tag{30}$$

Given analysis and synthesis windows that satisfy (30) we can reconstruct $\hat{z}_\mathrm{e}(n)$ from its STFT coefficients $\hat{Z}_\mathrm{e}(\ell,k)$. In practice, a Hamming window is often used for the synthesis window. A reasonable choice for the analysis window, having minimum energy [10], is given by

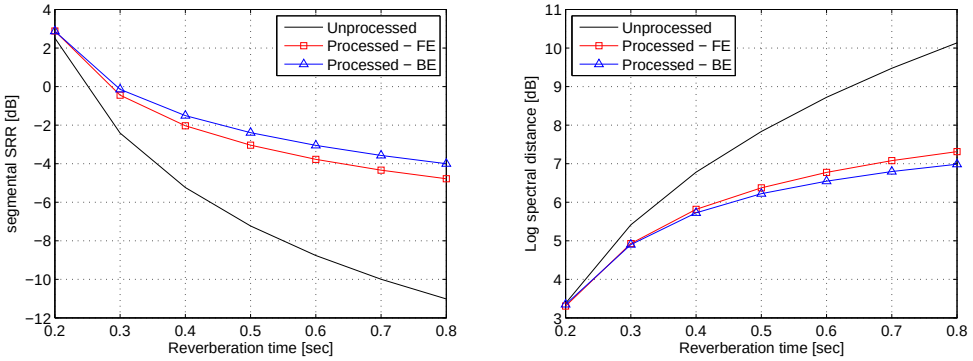
$$\tilde{\psi}(n) = \frac{\psi(n)}{N\sum_\ell \psi^2(n-\ell R)}. \tag{31}$$

The inverse STFT in (29) is efficiently implemented using the weighted overlap-add method [11].

6. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the reverberation suppression algorithm when using the forward and backward estimators. The algorithm was tested using reverberant speech (sampling rate was $f_\mathrm{s} = 8$ kHz) that was generated by convolving anechoic speech fragments from the APLAWD database [12] with various AIRs. The AIRs were generated using an efficient implementation of the celebrated image method [13]. The distance between the source and the microphone was 2.5 m. The reverberation time T_{60} ranges from 200 to 800 ms.

The length of the STFT analysis and synthesis window was $N = 256$, and an overlap between two successive STFT frames was 75% (i.e., $R = 64$). The parameter $G_{\min}$, which controls the maximum suppression, was set to -12 dB. The smoothing factor η was set to 0.9, and the weighting factor used in the decision-directed approach was set to 0.98. The time instance (measured with respect to the arrival time of the direct sound) at which the late reverberation starts was set to 48 ms, i.e., $N_e = 6$. The reverberation time $T_{60}(k)$ was determined for each octave band by applying Schroeder's method [14] to a bandpass filtered version of the AIR. The decay rate $\alpha(k)$ was calculated using (12). In practice one can use a blind estimation procedure as proposed in [15, 16].



(a) (b)

Fig. 2. Results of (a) segmental signal to reverberation ratio and (b) log spectral distance for the reverberant speech signals, processed signals using the forward estimator (FE) and processed signals using the backward estimator (BE) using different reverberation times.

The performance of the algorithm was evaluated using the segmental signal to reverberation ratio (SRR) and log spectral distance (LSD) [4]. For each reverberation time the results were averaged over 10 different source-microphone positions (with equal source-microphone distance), 10 (male and female) speech fragments and 5 sentences. The averaged results are shown in Figure 2 for (i) reverberant speech, (ii) processed speech using the forward estimator (FE) and (ii) processed speech using the backward estimator (BE). From these results we conclude that the backward estimator yields slightly better results in terms of the segmental SRR and LSD compared to the results obtained using the forward estimator. This might be caused by the approximation used to derive the backward estimator, since the early speech spectral variance $\lambda_{z_e}(\ell - N_e, k)$ is slightly larger than the direct speech spectral variance $\beta\lambda_s(\ell - N_e, k)$. The approximation becomes more accurate when $N_e \rightarrow 1$. It should be noted that similar results in terms of segmental SRR and LSD can be obtained by both estimators by using different values of N_e for each estimator.

7. CONCLUSIONS

One of the main challenges of developing reverberation suppression algorithms is the development of an estimator for the so-called late reverberant spectral variance. In this contribution a generalized statistical reverberation model is proposed that can be used to estimate this spectral variance. We have shown that both novel and existing estimators can be derived using this model. An alternative estimator that uses an estimate of the early speech component was derived, discussed and tested using simulated reverberant speech. It was shown that the backward and forward estimators provide comparable results. Further research is required to validate the model and the performance of the estimator using recorded reverberant speech and in the presence of ambient noise. In addition, the proposed statistical model might be used to derive other, more advantageous, estimators.

8. REFERENCES

[1] A. K. Nábělek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1259–1265, 1989.

[2] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[3] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

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

[5] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Processing*, vol. 28, no. 1, pp. 55–69, Feb. 1980.

[6] E. A. P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'04)*, Veldhoven, Netherland, Nov. 2004, pp. 250–254.

[7] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2007, ch. 45, part H.

[8] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, Apr. 2008, pp. 4037–4040.

[9] 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. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] J. Wexler and S. Raz, "Discrete Gabor expansions," *Speech Processing*, vol. 21, no. 3, pp. 207–220, Nov. 1990.

[11] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice-Hall, 1983.

[12] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Tech. Rep., Jun. 1987.

[13] E. A. P. Habets, "Room impulse response (RIR) generator," [Online] Available: http://home.tiscali.nl/ehabets/rir_generator.html, Oct. 2008.

[14] M. R. Schroeder, "A new method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

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

[16] H. W. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *International Workshop on Acoustic Echo and Noise Control (IWAENC'08)*, Sep 2008, pp. 1–4.