SINGLE-MICROPHONE LATE-REVERBERATION SUPPRESSION IN NOISY SPEECH BY EXPLOITING LONG-TERM CORRELATION IN THE DFT DOMAIN

Jan S. Erkelens and Richard Heusdens

Delft University of Technology, Dept. of Mediamatics, Mekelweg 4, 2628 CD, Delft, The Netherlands
Email: {j.s.erkelens, r.heusdens}@tudelft.nl

ABSTRACT

We consider blind late-reverberation suppression in speech signals measured with a single microphone in noisy environments. We exploit that reverberant speech shows correlation over longer time spans than clean speech by predicting the contribution of reverberant energy to the current observed spectrum from the enhanced spectra of previous frames. The prediction parameters are recursively updated with estimates of the correlation coefficients between the current reverberant spectrum and enhanced previous spectra. The contributions of late reverberation and noise are suppressed by a standard noise reduction algorithm. The algorithm is shown to decrease the long-term correlation. It achieves significant improvements in segmental speech-to-interference ratio and Bark spectral distortion for typical reverberation times and noise levels, while almost no distortions are introduced in clean speech.

Index Terms— Speech enhancement, echo suppression.

1. INTRODUCTION

It is well-known that noisy and/or reverberant speech is harder to understand than clean speech [1–3]. Generally also the performance of automatic speech recognition systems decreases [4, 5]. It is therefore of much interest to develop processing algorithms that enhance speech degraded by additive and convolutive distortions. Best results are achieved using multiple microphones, but often there is only one microphone available. Recently, several single-microphone blind dereverberation techniques have been proposed in the literature, e.g., [5–10].

The HERB technique [5] starts with constructing an initial estimate of the harmonic speech components corresponding to the direct path of the Room Impulse Response (RIR). This is done for a large database of speech filtered by the same RIR. The dereverberation filter is found in the frequency domain such that it, on the average, best turns the reverberant speech into the direct harmonic signal estimates. When the initial harmonic signal estimates are sufficiently good, HERB is capable of providing precise dereverberation even for reverberation times (T_{60})[1] as long as 1 second.

Filters that maximize the kurtosis of the Linear Prediction (LP) residuals have been proposed [11] for multi-microphone dereverberation. This technique also can be applied for a single microphone [9]. However, convergence of the adaptive filter is then quite slow and additional processing is required to suppress long-term reverberation effects.

Practical reverberant signals are often contaminated by nonstationary additive background noise as well. This may deteriorate the performance of methods that are designed to combat convolutive distortions only. Habets *et al.* [10] proposed a single-microphone processing technique that uses a statistical model to suppress late reverberation and noise together and, in a second stage, spatiotemporal averaging of the LP residual to reduce early reverberation and residual late reverberation. The method shows promising results. It requires blind estimation of T_{60} in noisy conditions, however, which is not a trivial problem [2].

The Discrete Fourier Transform (DFT) magnitudes of clean speech are highly correlated over time spans of about 50 ms [12]. In reverberant speech, this correlation length will be extended. Esch and Vary [13] exploited the correlation in clean speech DFT coefficients by means of complex Linear Prediction to improve noise reduction performance. The prediction error signal still contains the noise that is subsequently suppressed. Any reverberation will remain, however, because it is included in the predicted signal.

In this paper, we will exploit the increased correlation length of reverberant speech to make a prediction of the contribution of late reverberation to the spectral variance of the complex DFT coefficients of the current frame, and suppress it by applying a standard spectral gain function. We do not need to estimate T_{60}.

This paper is organized as follows. Section 2 describes the modeling assumptions and introduces some quantities of interest. In Section 3 our proposed method is presented in detail. Experimental results can be found in Section 4 and concluding remarks follow.

2. MODELING ASSUMPTIONS AND DEFINITIONS

2.1. Time-domain model

We assume the observed noisy and reverberant speech signal x to be the sum of a source speech signal s convolved with an RIR h and additive noise d, independent of s:

$$x(n) = \sum_{l=0}^{\infty} h(l)s(n-l) + d(n) = y(n) + d(n), \qquad (1)$$

where n is the discrete-time sample index and y is the noise-free reverberant signal. RIRs generally consists of a number of impulses for the early reflections and an exponentially decaying tail with a more noise-like appearance giving rise to the late reverberation.

2.2. Spectral modeling

Let $X(k,m)$, $Y(k,m)$, and $D(k,m)$ be complex-valued random variables representing the short-time DFT coefficients at frequency index k of the signal frame starting at sample index m from the noisy reverberant speech, reverberant noise-free speech, and noise process, respectively. In speech enhancement systems, m is a multiple rR of the frame hop size R. According to (1) we have $X(k,m) = Y(k,m) + D(k,m)$. $Y(k,m)$ can be written as

$$Y(k,m) = \sum_{l=0}^{\infty} h(l)S(k,m-l), \qquad (2)$$

[1] The reverberation time T_{60} is defined as the time taken for the sound to decay to 60 dB below its value at cessation [2].

where $S(k, m-l)$ is the short-time DFT of a windowed frame of s starting at sample index $m-l$. T_{60} can be defined frequency dependent. Interestingly, (2) shows that it is the length of the RIR in the *time* domain that determines, for *all* frequency bins, how many of the past clean DFT coefficients contribute significantly to the current observed DFT coefficient.

For ease of notation we may drop in the following time and/or frequency indices when this does not cause confusion. We can split the summation in (2) into a contribution Y_E from the direct path plus the early reflections, and the rest of the terms that constitute the late reverberation Y_L as follows

$$Y_E(k,m) = \sum_{l=0}^{L-1} h(l)S(k,m-l),\ Y_L(k,m) = \sum_{l=L}^{\infty} h(l)S(k,m-l). \tag{3}$$

2.3. Speech enhancement

Especially the reverberation arriving from about 50 ms after the direct signal degrades intelligibility [14]. As proposed in [15], we will model the current late reverberation term $Y_L(k,m)$ as an additive noise term that is uncorrelated with the current $Y_E(k,m)$. The late reverberation will be suppressed similar to the noise by means of a standard spectral gain function. Our task then is to estimate the spectral variances λ_L and λ_D of late reverberation and noise, respectively. For estimation of λ_D, we will use the method in [16], which can accurately track highly nonstationary noise sources. The algorithm for estimation of λ_L proposed in this paper differs from that in [15]. It exploits long-term correlation induced by the RIR and is detailed in the next section. Our final goal is to estimate the early-reverberance spectral DFT coefficients $Y_E(k,m)$. This is done in two steps. First noise reduction is applied[2] to $X(k,m)$ to estimate the reverberant DFT coefficient:

$$\hat{Y}(k,m) = G(\hat{\xi}_D(k,m), \hat{\zeta}_D(k,m))X(k,m), \tag{4}$$

where G is a spectral gain function and $\hat{\xi}_D$ and $\hat{\zeta}_D$ are estimates of the *prior SNR* and the *posterior SNR* parameters, defined as

$$\xi_D(k,m) = \frac{\lambda_Y(k,m)}{\lambda_D(k,m)}, \quad \zeta_D(k,m) = \frac{|X(k,m)|^2}{\lambda_D(k,m)}. \tag{5}$$

$\lambda_Y(k,m)$ is the variance of $Y(k,m)$. In the second step, $Y_E(k,m)$ is estimated as follows:

$$\hat{Y}_E(k,m) = G(\hat{\xi}_L(k,m), \hat{\zeta}_L(k,m))\hat{Y}(k,m), \tag{6}$$

with $\hat{\xi}_L = \hat{\lambda}_E/\hat{\lambda}_L$ and $\hat{\zeta}_L = |\hat{Y}|^2/\hat{\lambda}_L$. The algorithm for estimating λ_L is described in Sections 3.2 and 3.3. The variance of Y_E is denoted as λ_E. The decision-directed approach [17] will be used for estimation of the prior SNRs, with a bias correction [18]. The $\hat{Y}_E(k,m)$ are transformed back into the time domain and the enhanced speech is formed by an overlap-add procedure.

3. SPECTRAL VARIANCE ESTIMATION

3.1. Spectral autocorrelation functions

For a given frequency bin, the $S(k,m-l)$ in (2) are highly correlated for consecutive values of l, because the corresponding time frames (of length N samples) overlap by $N-1$ samples. Clean speech DFT coefficients can have significant correlation in time between them even when the frames overlap by less than 50% [12].

[2]One reason for applying noise suppression first is that noise signals may also be reverberant, with RIRs different from that of the speech. This step already suppresses some of the late reverberation (see Section 4).

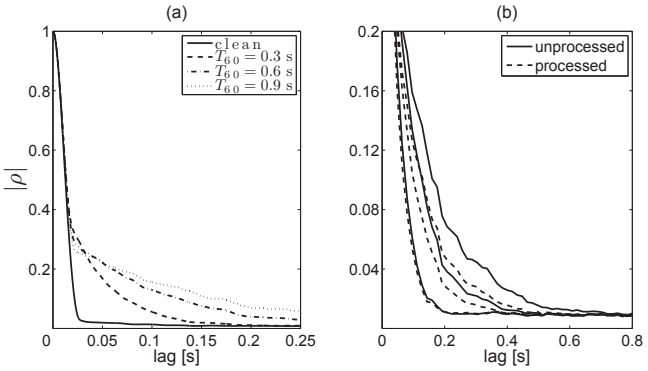

Fig. 1. (a) Absolute value of autocorrelation functions ρ of speech DFT coefficients for clean speech and reverberant speech with reverberation times of 0.3, 0.6, and 0.9 seconds. (b) Effect of the proposed algorithm on the long-term correlation in the reverberant speech.

In reverberant speech the correlation times are even larger. This is illustrated in Figure 1 (a). The absolute value of the (normalized) autocorrelation function of the complex DFT coefficients is shown for clean speech and for examples of reverberant speech with T_{60}s of 0.3, 0.6 and 0.9 seconds, respectively, obtained from about 5 minutes of speech[3]. Results are averages for the frequencies between 300 and 3400 Hz. The solid and dashed lines in Figure 1 (b) show the autocorrelation functions before and after application of our proposed algorithm to the reverberant speech signals, respectively. We see that our algorithm decreases the long-term correlation. This happens because we are *selectively* suppressing the reverberation: the DFT coefficients which are most affected by reverberation are most strongly suppressed by the spectral gain function.

3.2. Algorithm outline

It is possible to approximate the late reverberation term in (3) by a weighted sum $\tilde{Y}_L$ of speech DFTs, spaced P samples apart:

$$\tilde{Y}_L(k,m) = \sum_{j=0}^{J} c_j(k)S(k,m-\Delta-jP), \tag{7}$$

where Δ is introduced to skip the early reverberation part. Because of the correlation in the $S(k,m-l)$, each term in the summation in (7) will account for part of the contributions of the $P-1$ omitted terms in its neighborhood. The optimal coefficients $c_j(k)$ are complex and frequency dependent since the correlations are complex with frequency dependent phases. We would like to use $|\tilde{Y}_L|^2$ for estimation of λ_L. However, two modifications are needed. Firstly, we expect $|\tilde{Y}_L|^2$ to be biased low, because of the subsampling in (7). From the correlation function of clean speech shown in Figure 1 (a), we can calculate how much we expect to underestimate λ_L on the average. The resulting bias correction factor is given by

$$B = P/\sum_{l=-\infty}^{l=+\infty} |\rho(l)|^2, \tag{8}$$

where $\rho(l)$ is the correlation coefficient at lag l. The bias correction factor is roughly the same for all frequencies and we therefore use a frequency independent value.The second modification comes

[3]Frames with an energy more than 40 dB below the maximum frame energy of all the (reverberant) speech where omitted in the computation of the correlation coefficients.

about because in an enhancement system, we don't have available the clean speech DFT coefficients. We will therefore use the previously enhanced DFT coefficients $\hat{Y}_E(k, m-\Delta-jP)$ to make a prediction $\check{Y}_L$ as follows:

$$\check{Y}_L(k,m) = \sqrt{B}\sum_{j=0}^{J} \hat{c}_j(k)\hat{Y}_E(k, m-\Delta-jP). \qquad (9)$$

Section 3.3 explains how we determine the prediction coefficients $\hat{c}_j(k)$. An estimate $\hat{\lambda}_L$ of the late-reverberance spectral variance can now be made as follows

$$\hat{\lambda}_L(k,m) = \eta\hat{\lambda}_L(k,m-R) + (1-\eta)\left|\check{Y}_L(k,m)\right|^2, \qquad (10)$$

where η is a small smoothing parameter.

We limit $\hat{\lambda}_L$ to values smaller than the estimated variance of the reverberant speech $\hat{\lambda}_Y$ which was obtained from the noise reduction step. That is, we take $\min[\hat{\lambda}_L(k,m), \hat{\lambda}_Y(k,m)]$ as our final estimate of λ_L.

The time span of $\Delta + JP$ should cover the maximum T_{60} considered. In this paper, we consider a maximum T_{60} of about 1 second. Larger values can be handled easily by increasing the value of J. Proper choices for all parameters are given in Section 3.4.

3.3. Estimation of the prediction coefficients

The prediction coefficients $\hat{c}_j(k)$ in (9) are found by estimation of the average correlation coefficients between $\hat{Y}(k,m)$ and the $\hat{Y}_E(k, m-\Delta-jP)$. The correlation coefficients are near-optimal least-squares solutions when P is chosen such that the $\hat{Y}_E(k, m-\Delta-jP)$ are weakly correlated. We use recursive smoothing to be able to adapt to slow changes in the RIR, as follows

$$\hat{c}_{j,r}(k) = \alpha_{j,r}(k)\hat{c}_{j,r-1}(k) + (1-\alpha_{j,r}(k))\hat{\rho}_{j,r}(k), \qquad (11)$$

where r is the frame index. The $\alpha_{j,r}(k)$ are smoothing parameters and $\hat{\rho}_{j,r}(k)$ are the current estimates of the correlation coefficients. The recursive smoothing effectively causes the parameters to be estimated from a limited amount of data (in the order of a few seconds). But since the sequence of speech spectral amplitudes is nonstationary with a very large dynamic range, we must be very careful with updating the prediction coefficients in order to keep their variance low. Large values of the $|\hat{Y}|$ and $|\hat{Y}_E|$ can cause jumps in the parameters. This can be avoided by computing the $\hat{\rho}_{j,r}(k)$ from normalized data, as follows:

$$\hat{\rho}_{j,r}(k) = \frac{\hat{Y}(k,m)}{\beta(k,m)|\hat{Y}(k,m)|}\frac{\hat{Y}_E^{\dagger}(k, m-\Delta-jP)}{|\hat{Y}_E(k, m-\Delta-jP)|}, \qquad (12)$$

where † means complex conjugation and $\beta(k,m)$ is a bias correction factor. It is given by

$$\beta(k,m) = \frac{\hat{\mu}_{|E|}(k,m)}{\hat{\mu}_{|Y|}(k,m)}, \qquad (13)$$

where $\hat{\mu}_{|Y|}(k,m)$ and $\hat{\mu}_{|E|}(k,m)$ are long-term estimates of the the mean of $|\hat{Y}(k,m)|$ and $|\hat{Y}_E(k,m)|$, respectively. The factor $\beta(k,m)$ is introduced to correct for the large bias that would result from simply normalizing with $|\hat{Y}(k,m)|$. The long-term mean estimates are computed by recursive smoothing of the $|\hat{Y}(k,m)|$ and $|\hat{Y}_E(k,m)|$, respectively, with a smoothing factor equal to 0.98.

The default value of the $\alpha_{j,r}(k)$ in (11) is 0.98. However, we set them to 1 (meaning that the $\hat{c}_j(k)$ are not updated) when the $\hat{\rho}_{j,r}(k)$ are deemed unreliable. The following conditions are used:

a) $\hat{\zeta}_D(k,m) < \tau$;

b) $\frac{|\hat{Y}_E(k,m-\Delta-jP)|^2}{\hat{\lambda}_L(k,m-\Delta-jP)+\hat{\lambda}_D(k,m-\Delta-jP)} < \theta$,

where τ and θ are threshold parameters. Conditions a) and b) aim at excluding from the adaptation the DFT coefficients that are affected strongly by the noise and reverberation. When a) happens, $\alpha_{j,r}(k)$ is set to 1 for all j, while b) is checked for each j individually.

Because the $\hat{c}_j(k)$ are estimated from a limited amount of data, their variance can cause overestimation of $\hat{\lambda}_L$, leading to some distortions in clean speech. These distortions can be kept to a minimum by using only the terms in the summation in (9) for which the $\hat{c}_j(k)$ are statistically significant. We determined experimentally the standard deviation of the $\hat{c}_j(k)$ in clean speech (that is, without noise or reverberation). When the algorithm is applied in practice only the terms in (9) are summed for which the current value of $|\hat{c}_j(k)|$ is larger than 2.6 times the standard deviation of $\hat{c}_j(k)$ in clean speech.

3.4. Parameter settings

We use frames of length $N = 256$ samples (32 ms for a sampling frequency of 8 kHz). Square-root Hanning analysis and synthesis windows are applied. Common values of R are $N/4$ or $N/2$. We experimented with $R = N/4$ and $R = N/2$, and $P = R$ and $P = 2R$ (for convenience, P is taken as a multiple of R). We found that smaller values of R and P lead to more reverberation suppression, but also to more speech distortion. Therefore, we chose $R = N/2$ and $P = 2R$ for the experiments in Section 4. Δ should be taken larger than the correlation length in clean speech. From Figure 1(a), we see that a value corresponding to about 30 ms or more is appropriate, therefore we used $\Delta = N$. With these choices $J = 30$ is sufficient to cover a maximum T_{60} of 1 second. A bias correction factor $B = 1.65$ was applied in (9). A constant value of $\eta = 0.2$ was used in (10), and the default value of the $\alpha_{j,r}$ in (11) is 0.98. We set $\tau = 2$ and $\theta = 1$ in conditions a) and b) in Section 3.3, respectively, independent of frequency. The gain function G in (4) and (6) assumes a generalized Gamma prior for the speech spectral amplitudes with parameters $\gamma = 1$ and $\nu = 1$ [19].

4. EXPERIMENTAL RESULTS

We evaluated the algorithm on speech convolved with RIRs with reverberation times up to 0.9 seconds, in different noise conditions. All RIRs used in this paper were simulated with the image method, using the Matlab implementation provided by Habets [15, 20]. The RIRs were simulated for a room with dimensions 6x5x3 meters. We used a source-microphone distance of 1 meter.

The speech material consisted of 3 minutes of speech sentences from TIMIT, without intervening pauses. The noise signals were taken from the NTT monaural noise database. We used car interior noise and shopping mall noise. In addition, computer-generated stationary white Gaussian noise was used. All signals were limited to telephone bandwidth. The prediction coefficients were always initialized with zero values.

For evaluation of the algorithm, we use Segmental Signal-to-Interference Ratio ($SegSIR$). Let $\mathbf{s}_r$ be a time frame of the clean signal corresponding to the direct path of the RIR. Similarly, $\hat{\mathbf{s}}_r$ is a time frame of an enhanced signal. $SegSIR$ is now defined as

$$SegSIR = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} 10\log_{10}\left(\frac{||\mathbf{s}_r||^2}{||\mathbf{s}_r - \hat{\mathbf{s}}_r||^2}\right), \qquad (14)$$

where $||.||^2$ is the energy of a frame and $\mathcal{R}$ is an index set consisting of the non-silence frames of the clean direct signal. $|\mathcal{R}|$ is the cardinality of $\mathcal{R}$.

Table 1. $SegSIR$ [dB] values before processing (bp) and after noise suppression (ns) and subsequent dereverberation suppression (ds) for different reverberation times, noise sources, and SNRs.

T_{60} $[s]$		Noise free	Stat. WGN 10 dB	Stat. WGN 20 dB	Car int. 10 dB	Car int. 20 dB	Shop. mall 10 dB	Shop. mall 20 dB
0	bp	∞	3.97	14.0	5.50	15.5	4.80	14.8
	ns	39.5	9.18	16.2	10.2	17.6	7.81	15.9
	ds	35.5	9.00	15.7	10.1	17.2	7.74	15.6
0.3	bp	4.06	-1.01	2.64	-0.55	2.75	-0.68	2.74
	ns	4.18	2.67	3.76	2.68	3.78	1.91	3.59
	ds	4.69	3.10	4.18	3.07	4.23	2.30	4.03
0.6	bp	-1.60	-4.35	-2.15	-4.11	-2.11	-4.16	-2.12
	ns	-1.04	-1.08	-0.99	-1.19	-1.02	-1.47	-1.06
	ds	0.94	0.62	0.91	0.48	0.85	0.14	0.86
0.9	bp	-4.50	-6.46	-4.82	-6.30	-4.81	-6.33	-4.81
	ns	-3.53	-3.21	-3.39	-3.36	-3.45	-3.41	-3.44
	ds	-0.81	-0.82	-0.75	-1.00	-0.78	-1.12	-0.78

Table 1 shows the $SegSIR$ values for different reverberation times, and noise sources and Signal-to-Noise Ratios (SNRs). The SNRs are defined with respect to the reverberant speech. We also computed the EMBSD measure [21]. Its values are not shown because of lack of space, but they display clear improvements as well.

We observe that our algorithm causes almost no distortions on clean direct signals ($SegSIR = 35.5$ dB). The reverberant character of the other signals was clearly reduced by the processing, also in the noisy conditions. The artifacts that could be heard in these signals are typical of spectral suppression algorithms.

One can notice in the table that the noise reduction step suppresses some of the reverberation. This does not affect the overall performance, since it decreases the long-term correlation. Therefore, the subsequent reverberation suppression works on the remaining part of the reverberance energy.

5. CONCLUDING REMARKS

Considerable reduction in late reverberation is possible by exploiting long-term correlation in the DFT domain, even with a single microphone in noisy conditions. Combining our algorithm with techniques that reduce early-reverberation effects is of interest, for example those that process the linear prediction residual [6, 9, 10]. This may also further improve the suppression of the late reverberation, because the approximation made in (9) then becomes more accurate.

The prediction coefficients are continuously updated in our algorithm and we may be able to follow slow changes in the room impulse response. Robustness to such changes is currently under investigation.

6. REFERENCES

[1] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318 – 326, January 1980.

[2] R. Ratnam *et al.*, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, vol. 114, no. 5, pp. 2877–2892, November 2003.

[3] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," *Int. Workshop, Acoustic Echo and Noise Control*, 2005.

[4] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," *Proc. ICASSP*, vol. I, pp. 557 – 560, 2002.

[5] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 1, pp. 80 – 95, January 2007.

[6] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech, Audio Proc.*, vol. 8, no. 3, pp. 267 – 281, may 2000.

[7] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech, Audio Proc.*, vol. 11, no. 5, pp. 467 – 488, September 2003.

[8] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," *Proc. Interspeech*, pp. 3145 – 3148, 2005.

[9] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 3, pp. 774 – 784, May 2006.

[10] E. A. P. Habets, N. D. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," *Proc. ICASSP*, pp. 4577 – 4580, 2008.

[11] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP*, pp. 3701 – 3704, 2001.

[12] I. Cohen, "Relaxed statistical model for speech enhancement and *a priori* SNR estimation," *IEEE Trans. Speech, Audio Proc.*, vol. 13, no. 5, pp. 870–881, September 2005.

[13] T. Esch and P. Vary, "Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors," *Proc. ICASSP*, pp. 4877 – 4880, 2008.

[14] F. Santon, "Numerical prediction of echograms and of the intelligibility of speech in rooms," *J. Acoust. Soc. Am.*, vol. 59, no. 6, pp. 1399 – 1405, June 1976.

[15] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," *Ph.D. Thesis, Eindhoven Univ. of Techn., The Netherlands*, 2007.

[16] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 6, pp. 1112 – 1123, August 2008.

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

[18] J. S. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Comm.*, vol. 49, pp. 530–541, July–August 2007.

[19] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 6, pp. 1741–1752, August 2007.

[20] "Room Impulse Response generator for Matlab," `http://home.tiscali.nl/ehabets/rir_generator.html`.

[21] W. Yang, "Enhanced Modified Bark Spectral Distortion (EMBSD)," `http://www.temple.edu/speech_lab/Wonhos_Dissertation.pdf`, *Ph.D. Thesis, Temple Univ., Ft. Washington, USA*, 1999.