Forum Acusticum 2011 - Aalborg - Denmark

Suppression of late reverberation at multiple speaker positions utilizing a single Room Impulse Response measurement

Alexandros Tsilfidis, Elias K. Kokkinis and John Mourjopoulos
Audio and Acoustic Technology Group, Department of Electrical and Computer Engineering, University of Patras, 26504, Greece

Summary
A spectral subtraction method for dereverberation at multiple speaker positions utilizing a single Room Impulse Response (RIR) measurement is presented. A running kurtosis approach is applied in order to define the RIR early/late reflections boundary. Then, the power spectrum of the late reverberant part of the measured RIR along with the excitation signal derived from the Linear Prediction (LP) analysis of the reverberant source are used to estimate the late reverberant power spectrum. A gain magnitude regularization step is also employed to compensate for overestimation errors and reduce musical noise artifacts. Objective and subjective performance measures show that a significant enhancement of the reverberant signals is achieved in all examined cases.

PACS no. 43.60.Dh, 43.55.Jz

1. Introduction

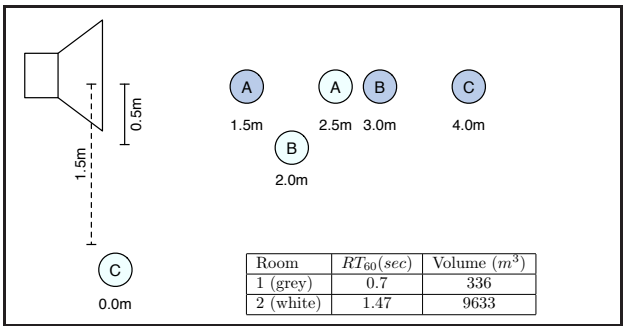
Room reverberation has a detrimental effect on the quality and intelligibility of speech signals recorded in enclosed spaces, while also deteriorating the performance of automatic speech recognition (ASR) systems. Consequently, dereverberation has been a challenging research topic for at least four decades. In theory, an ideal inverse filter of the Room Impulse Response (RIR) between the speech source and the microphone would remove the effects of reverberation; however a stable causal inverse filter cannot always be derived, especially when the exact source/receiver positions are not known [1]. In room acoustics the RIR is usually decomposed in an early and late part, namely the *early reflections* and the *late reverberation*. The early reflections mainly affect the signal's timbre and are perceived as coloration while the late reverberant tails produce a noise-like effect [2]. Most dereverberation techniques treat the effects of these two parts separately.

Several speech-oriented late reverberation suppression methods have been proposed as the degradation in ASR performance and speech intelligibility are to a large extent due to the effect of late reflections (e.g. [2, 3, 4, 5, 6, 7]). Many of these methods use spectral subtraction based either on RIR or signal modelling (e.g. [4, 8]) but very few take advantage of a measured RIR (e.g. [3]), even though the idea of exploiting the common features of RIRs in different room positions is not new (e.g. [9]).

Here, a method for suppressing late reverberation from a moving speaker utilizing a single RIR measurement is presented, assuming that the late part of the RIR approaches a wide-sense stationary process. Therefore, the spectral magnitude of late reverberation is estimated using the late part of the measured impulse response ("reference" response) and the excitation signal derived from the Linear Prediction (LP) analysis of the reverberant signal. Then, spectral subtraction is used to derive a clean signal estimation. Contrary to ASR-oriented techniques (e.g. [3]) the present approach introduces a low-complexity implementation and does not require database training either for the derivation of the early/late reflections boundary or for implementing the spectral subtraction. A final novel step that prevents overestimation errors and reduces musical noise artifacts through gain magnitude regularization is also introduced so that low Signal to Reverberation Ratio (SRR) signal regions are identified [10] and the suppression is dynamically constrained in order to avoid such processing distortions.

2. Method Description

Assuming a noise-free and stationary environment the transfer function between a speaker and a microphone



Room	$RT_{60}(sec)$	Volume (m^3)
1 (grey)	0.7	336
2 (white)	1.47	9633

Figure 1. Illustration of the measurement setup for the RIRs used in Section 3.

in a room can be described by the corresponding impulse response $h_r(n)$ (n denotes the discrete time index). Then the reverberant speech signal captured by the microphone $y(n)$ is the convolution of the clean speech $s(n)$ and the room impulse response:

$$y(n) = \sum_{m=0}^{L_r} h_r(m)s(n-m) \tag{1}$$

where L_r is the length of the RIR. Using the LP analysis, a speech signal is modelled as the convolution of an excitation signal $u(n)$ and a speech production filter $h_s(n)$ describing the formant structure determined by the glottal, the vocal tract, and the lip radiation filters:

$$s(n) = \sum_{m=0}^{L_s} h_s(m)u(n-m) \tag{2}$$

From Eq. 1 and 2 the reverberant signal can be described as:

$$y(n) = \sum_{m=0}^{L_r}\sum_{l=0}^{L_s} h_r(m-l)h_s(l)u(n-m) \tag{3}$$

As mentioned, the impulse response of a reverberant room can be separated in two parts, that is the early reflections and the late reverberation:

$$h_r(n) = h_{early}(n) + h_{late}(n) \tag{4}$$

Therefore Eq. 3 can be written as:

$$y(n) = \sum_{m=0}^{L_b}\sum_{l=0}^{L_s} h_{early}(m-l)h_s(l)u(n-m) + \sum_{m=L_b+1}^{L_r}\sum_{l=0}^{L_s} h_{late}(m-l)h_s(l)u(n-m) \tag{5}$$

where L_b is the length of $h_{early}(n)$. Usually the length of the speech production filter can be assumed shorter than the length of $h_{early}(n)$ (i.e. $L_s < L_b$) [11]. Hence Eq. 5 is written as:

$$y(n) = \sum_{m=0}^{L_b}\sum_{l=0}^{L_s} h_{early}(m-l)h_s(l)u(n-m) + \sum_{m=L_b+1}^{L_r} h_{late}(m)u(n-m) \tag{6}$$

Consider now a setup with a fixed microphone and a moving source which represents a common hands-free communications scenario. The source has an initial position ρ_0 with a corresponding RIR $h^0(n)$. Assume that $h^0(n)$ is known and as in equation 4 it can be expressed as:

$$h_r^0(n) = h_{early}^0(n) + h_{late}^0(n) \tag{7}$$

In general, when the source moves to another position ρ_i, a different RIR defines the corresponding acoustical path:

$$h_r^i(n) = h_{early}^i(n) + h_{late}^i(n) \tag{8}$$

The early part of a RIR changes significantly with even small changes in the source-microphone position. On the other hand, during the late reverberation part, the energy is statistically equal in all regions of the room [12]. Hence, it can be assumed that the power spectral density (PSD) of $h_{late}^i(n)$ is approximately the same for all i and equal to the PSD of $h_{late}^0(n)$:

$$|H_{late}^i(\kappa,\omega)|^2 = |H_{late}^0(\kappa,\omega)|^2 \quad \forall i \tag{9}$$

where κ and ω are the time frame and the frequency bin index respectively, $H_{late}^0(\kappa,\omega)$ is the Short Time Fourier Transform (STFT) of the late part of the measured impulse response.

Based on the above assumption and the LP analysis, the principle of spectral subtraction can be used for the suppression of late reverberation from a moving speech signal. A fairly accurate estimation of the late reverberation power spectrum $R_{late}(\kappa,\omega)$ can be obtained in the STFT domain as:

$$R_{late}(\kappa,\omega) = |H_{late}^0(\kappa,\omega)|^2|U^i(\kappa,\omega)|^2 \tag{10}$$

with $U^i(\kappa,\omega)$ being the STFT of the LP residual of the reverberant signal. Hence, an estimation of the clean signal's power spectrum can be derived:

$$|\hat{S}^i(\kappa,\omega)|^2 = \frac{|Y^i(\kappa,\omega)|^2 - R_{late}(\kappa,\omega)}{|Y^i(\kappa,\omega)|^2}|Y^i(\kappa,\omega)|^2 = G(\kappa,\omega)|Y^i(\kappa,\omega)|^2 \tag{11}$$

where $|S^i(\kappa,\omega)|^2$ and $|Y^i(\kappa,\omega)|^2$ are the PSD estimations of the clean and the reverberant signals in position ρ_i respectively and $G(\kappa,\omega)$ is the derived gain magnitude function.

2.1. Defining the early-late reverberation boundary

For the sake of simplicity, the boundary between early reflections and late reverberation of a RIR is often defined as a fixed time interval [12] or in relation to the volume of the room [13]. However, the precise definition of the early/late reflections boundary is a challenging and open research issue. In this work, the method proposed in [13, 14] will be used. The measured RIR $h^0(n)$is partitioned in non-overlapping frames of length L_κ and for each frame $\boldsymbol{h}^0(k)$ the normalized kurtosis is calculated as follows:

$$Kurt[\boldsymbol{h}^0(k)] = \frac{E[\boldsymbol{h}^0(k) - \mu]^4}{\sigma^4} - 3 \tag{12}$$

The boundary is defined as $L_b = k_{min} L_\kappa$ where L_κ is the frame size and k_{min} is given by:

$$k_{min} = argmin(Kurt[\boldsymbol{h}^0(k)]) \tag{13}$$

2.2. Musical noise reduction through Gain Magnitude Regularization

When the term $R_{late}(\kappa, \omega)$ is overestimated, spectral subtraction methods tend to generate musical noise artifacts. Many methods have been proposed in order to compensate this effect (e.g. [6, 10]). Here a low-complexity approach based on Gain Magnitude Regularization (GMR) is proposed. High SRR spectral regions such as signal steady states are less affected by late reverberation [6, 15] and thus an overestimation of the late reverberation is less likely to happen. On the other hand, artifacts are expected in low SRR regions, hence a low SRR detector is used [10]. Then, a GMR technique is introduced in order to constraint only the low gain parts. This can be beneficial when compared to moving-averaging approaches as it is not affecting the high gain bins. The PSD estimation of the clean signal is derived as in Eqs 14 and 15, where θ is the threshold for applying the gain constraints, r is the regularization ratio, ζ is the power ratio between the enhanced and the reference signal, ζ_{th} the threshold of the low SRR detector and Ω is the frame size.

$$\zeta = \frac{\sum_{\omega=1}^{\Omega} G(\kappa, \omega) |Y^i(\kappa, \omega)|^2}{\sum_{\omega=1}^{\Omega} |Y^i(\kappa, \omega)|^2} \tag{15}$$

3. Tests and Results

For the evaluation of the proposed method sixteen phrases uttered by both male and female speakers of the TIMIT database were convolved with real RIRs measured in (a) a lecture hall (Room 1) and (b) a large auditorium (Room 2) of a conference center. The measurement setup, as well as the room acoustical properties are shown in Fig. 1. The performance of the proposed method was examined both with and without the GMR step (when the GMR module was not implemented a -20 dB threshold was applied to the suppression). The speech signals and the RIRs were sampled at 16 kHz with 16 bit precision and the LP analysis order was 13. The frame size was 1024 samples with a 25% overlap, the thresholds θ and ζ_{th} were set at 0.4 and the value of the regularization ratio r was 6. Corresponding audio demos can be found at the authors' website.[1]

Figures 2 and 3 show the averaged segmental Signal to Reverberation Ratio (SRR) improvement. For each room, the method was evaluated at three different positions. For each position the method was also tested three times, each time assuming a RIR measured at each of the three test positions. An improvement in terms of SRR was noticed in all tested cases with a relatively consistent performance regardless of the reference RIR. In both rooms the improvement was greater for more distant room positions (position C in both rooms) where the reverberant signals contained more late reverberation. The use of the GMR step resulted in a small reduction of the achieved late reverberation suppression in all positions.

Table I presents the improvement achieved by the proposed method in terms of the Perceptual Speech Quality measure (PESQ)[16], when comparing to the reverberant signals. PESQ implements a perceptual model in order to assess the quality of a processed speech signal and rate its quality according to the five grade Mean Opinion Score (MOS) scale. Again, an improvement is shown for all tested cases. Moreover, the application of the GMR resulted in significantly improved results for both rooms. Apparently the approximate nature of the late reverberation spectrum estimation presented here may produce artifacts and the GRM step is important in order to perceptually enhance the evaluated clean speech signal. It is interesting to note that the main assumption used here, i.e. the stationarity of the late reverberation, is largely supported by the presented results, as the method performs consistently regardless of the reference RIR used. The proposed technique has been also evaluated in comparison with known blind techniques (e.g. [8, 17]) and produces better results. However, these results are omitted here, as the proposed technique requires a RIR measurement and a comparison to such blind methods is beyond the scope of our work.

[1] http://www.wcl.ece.upatras.gr/audiogroup/tools/derev.html

$$|\hat{S}^i(\kappa,\omega)|^2 = \begin{cases} \left(\dfrac{G(\kappa,\omega)-\theta}{r}+\theta\right)|Y^i(\kappa,\omega)|^2 & \text{when } \zeta < \zeta_{th} \\ & \text{and } G(\omega,j) < \theta \\ G(\kappa,\omega)|Y^i(\kappa,\omega)|^2 & \text{otherwise} \end{cases} \tag{14}$$

Table I. PESQ improvement for various cases

Room 1						
Position	Reference Impulse Response					
	h_A	h_A+GMR	h_B	h_B+GMR	h_C	h_C+GMR
A	0.069	0.142	0.095	0.148	0.073	0.141
B	0.067	0.143	0.072	0.143	0.064	0.138
C	0.055	0.184	0.073	0.187	0.060	0.185
Room 2						
Position	Reference Impulse Response					
	h_A	h_A+GMR	h_B	h_B+GMR	h_C	h_C+GMR
A	0.082	0.195	0.098	0.199	0.116	0.190
B	0.026	0.163	0.051	0.160	0.055	0.162
C	0.049	0.156	0.077	0.143	0.063	0.150

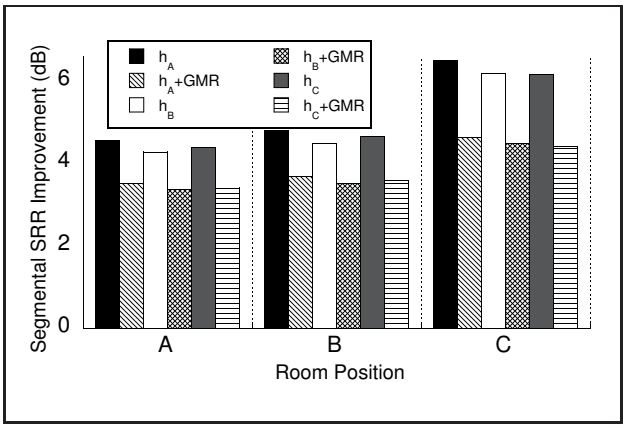


Figure 2. SRR improvement (in dB) for different cases in Room 1

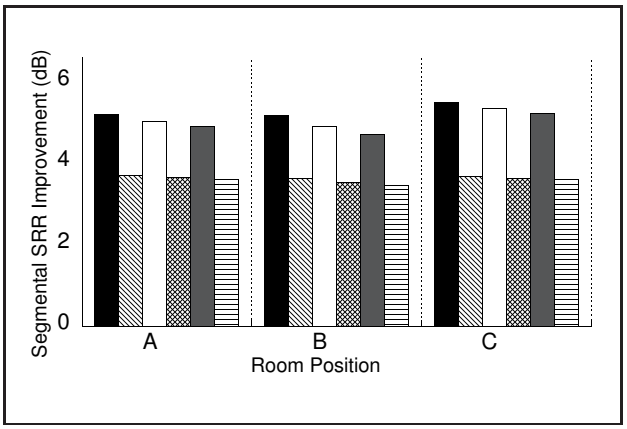


Figure 3. SRR improvement (in dB) for different cases in Room 2

4. Conclusions

The proposed technique extracts the room late reflections characteristics based on a single RIR measurement and adopts an efficient spectral subtraction approach in order to suppress late reverberation at multiple speaker positions. A Gain Magnitude Regularization technique is also used in order to compensate for any overestimation errors and to reduce the musical noise artifacts. Objective and subjective results show significant speech enhancement independent of the measurement position of the reference RIR.

Acknowledgement

The research activities that led to these results, were co-financed by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Strategic Reference Framework (ESPA) 2007-2013, according to Contract no. MICRO2-38/E-II-A.

References

[1] J. Mourjopoulos. On the variation and invertibility of room impulse response functions. *J. Sound Vib.*, 102:217–228, 1985.

[2] E. A. P. Habets, S. Gannot, I Cohen, and P. C. W Sommen. Joint dereverberation and residual echo suppression of speech signals in noisy environments. *IEEE Trans. Audio, Speech and Lang. Process.*, 16(8):1433–1451, 2008.

[3] R. Gomez and T Kawahara. Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. *IEEE Trans. Audio, Speech and Lang. Process.*, 18:1708–1716, 2010.

[4] J. S. Erkelens and R Heusdens. Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE Trans. Audio, Speech and Lang. Process.*, 18:1746–1765, 2010.

[5] K. Kinoshita, M Delcroix, T Nakatani, and M Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Trans. Audio, Speech and Lang. Process.*, 17:534–545, 2009.

[6] A. Tsilfidis and J. Mourjopoulos. Signal-dependent constraints for perceptually motivated suppression of late reverberation. *Signal Process.*, 90:959–965, 2010.

[7] A. Tsilfidis and J. Mourjopoulos. Blind single-channel suppression of late reverberation based on perceptual reverberation modeling. *J. Acoust. Soc. Amer.*, 129(3):1439–1451, 2011.

[8] K. Lebart and J. Boucher. A new method based on spectral subtraction for speech dereverberation. *Acta Acust. Acust.*, 87:359–366, 2001.

[9] Y. Haneda, S. Makino, and Y. Kaneda. Common acoustical pole and zero modeling of room transfer functions. *IEEE Trans. on Speech Audio Process.*, 2(2):320–328, 1994.

[10] M. Jeub, M Schafer, T Esch, and P Vary. Model-based dereverberation preserving binaural cues. *IEEE Trans. Audio, Speech and Lang. Process.*, 18:1732–1745, 2010.

[11] M Woelfel. Enhanced speech features by single-channel joint compensation of noise and reverberation. *IEEE Trans. Audio, Speech and Lang. Process.*, 17:312–323, 2009.

[12] B. Blesser. An interdisciplinary synthesis of reverberation viewpoints. *J. Aud. Eng. Soc.*, 49(10):867–903, 2001.

[13] G. Defrance and J.-D. Polack. Measuring the mixing time in auditoria. In *in Proc. of the Acoustics '08*, pages 3871–3876, Paris, France, June-July 2008.

[14] R. Stewart and M. Sandler. Statistical measures of early reflections of room impulse responses. In *in Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx-07)*, pages 1–4, Bordeaux, France, September 2007.

[15] B. Yegnanarayana and P. S. Murthy. Enhancement of reverberant speech using lp residual signal. *IEEE Trans. Audio, Speech and Lang. Process.*, 8(3):267–281, 2000.

[16] ITU-R P.862. Perceptual evaluation of speech quality (pesq), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. 2000.

[17] K. Furuya and A Kataoka. Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *IEEE Trans. Audio, Speech and Lang. Process.*, 15:1579–1571, 2007.