On the Application of Psychoacoustically-Motivated Dereverberation for Recordings taken in the German Parliament

Marco Jeub and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany

{jeub,vary}@ind.rwth-aachen.de

Abstract: In this paper, we discuss the application of speech dereverberation techniques for post-processing of recordings taken in the German parliament. Based on a novel psychoacoustically-motivated dereverberation concept, a significant improvement in terms of the perceived quality is obtained in comparison to a conventional dereverberation approach. Since time-varying changes of the acoustical environment are negligible, all required acoustical parameters such as reverberation time (RT) and direct-to-reverberant-energy ratio (DRR), are determined in an off-line procedure.

1 Introduction

Dereverberation of speech signals has received an increasing attention from the research community over the last years. Many authors suggested algorithms which are suitable for mobile phones, automatic speech recognition systems and digital hearing aids, cf. [1].

In this contribution, we show how such techniques can be used to enhance speech recordings taken in large conference halls or similar environments. For a case study, speech recordings from the German Bundestag, which is located in the Reichstag in Berlin, are investigated and suitable signal processing methods for a post-processing are proposed. Based on a conventional dereverberation algorithms, an improved concept which exploits masking properties of the human auditory system is developed.

In the remainder of this paper, the acoustical environment is analyzed in the next section which includes the considered signal model, followed by a short discussion on the reverberation time (RT) and direct-to-reverberant energy ratio (DRR). In Section 3, a conventional dereverberation algorithm is introduced and a psychoacoustically-motivated modification is discussion. Finally, in Sections 4 and 5 we show simulation results and draw conclusions.

2 Analysis of the Acoustical Environment

Room reverberation is usually caused by reflections of the emitted source, e.g., a speaker which stands far away from another speaker in an enclosure. In contrast to that, in a parliament discussion, the speaker is located at a lectern and the speech is captured by microphones at a small distance. This speech signal is then processed and emitted by a loudspeaker system to the audience. In very large rooms, this signal is then reflected on the walls and fed back into the microphones with a certain sound propagation delay. In order to avoid instability and overshoots, a feedback cancellation is employed which usually consists of a notch filter or an adaptive filter

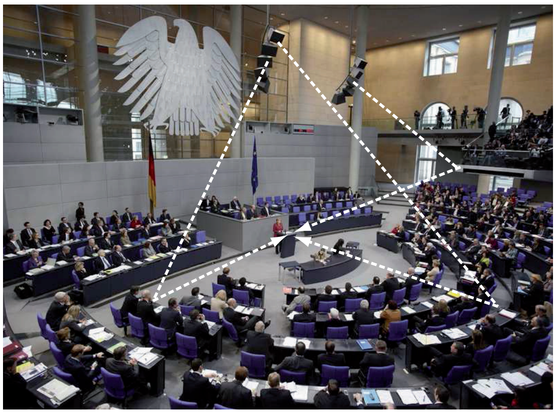

Figure 1 - Pictures of the German Bundestag[1]. (left) Audience where the lines mark possible sound propagation paths from the loudspeaker system to the microphones at the lectern; (right) Speaker at lectern with two microphones.

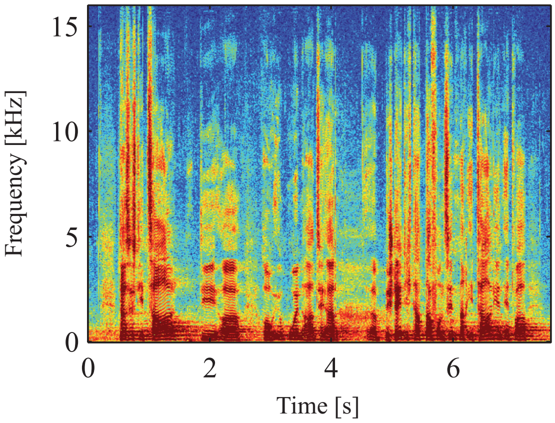

Figure 2 - Spectrogram of a recording without processing ($x(k)$).

with a short filter length. However, the captured reverberation remains, if no further countermeasures are employed. This acoustic situation is illustrated exemplarily for the German parliament in the left subplot of Fig.1 where simplified sound propagation paths are marked by the dashed lines. A speaker standing at the lectern is shown in the right subplot where the short distance to the capturing microphones can clearly be seen. Even though two microphones are mounted on the lectern, only one microphone is used in this setup since the second one is used as a microphone breakdown replacement.

In the spectrogram of a short recording in Fig.2, the smearing over time due to reverberation can clearly be seen. Since the acoustical scenario does not change, apart from small movements of the speaker and the audience, all required acoustical parameters for the considered speech enhancement algorithm such as reverberation time and direct-to-reverberant energy ratio can be estimated only once. This off-line procedure is carried out blindly from recorded data since no acoustical *room impulse response* (RIR) measurements are available.

[1] Photographic material provided by the digital image service of the German Bundestag. (c) Werner Schüring (left) and Thomas Trutschel/photothek.net (right).

2.1 Signal Model

The microphone signal $x(k)$ is the input of the considered single-channel speech enhancement system and is related to the clean speech $s(k)$ and a given RIR $h(k)$ by

$$x(k) = s(k) * h(k), \tag{1}$$

where $*$ indicates the convolution and k the discrete time index.

Within the DFT-domain speech dereverberation system, the input signal is first segmented into overlapping frames of length L. After windowing and zero-padding, these frames are transformed via FFT of length M into the short-term spectral domain. The corresponding spectra are denoted by $X(\lambda,\mu)$ where frame index and discrete frequency bin index are denoted by λ and μ. The enhanced spectral coefficients $\hat{S}(\lambda,\mu)$ are obtained by the multiplication of the coefficients $X(\lambda,\mu)$ with spectral weighting gains $G(\lambda,\mu)$. The enhanced time-domain signal $\hat{s}(k)$ is obtained by using the IFFT and overlap-add.

2.2 Reverberation Time (RT)

One fundamental parameter of room acoustics is the RT. It is defined as the time period a sound needs to decrease by 60 dB from its initial *sound pressure level* (SPL) after switch-off and is therefore also referred to as T_{60}. It is linked with the decay rate ρ by

$$\rho = \frac{3\ln(10)}{\mathrm{T}_{60}}. \tag{2}$$

If a time-continuous room impulse response $h(t)$ is available, e.g., based on measurements, the reverberation time can be measured with the Schroeder method [2]. Based on the *energy decay curve* (EDC), which can be obtained from the Schroeder integral by

$$\mathrm{EDC}(t) = \int_{t}^{\infty} h^2(\tau)\mathrm{d}\tau, \tag{3}$$

the RT can be determined directly from the time, where the EDC needs to drop by 60 dB from its initial energy level.

However, since the RIR is unknown from our recordings, the RT has to be estimated blindly from the reverberant input signal. In this contribution we use a modified *maximum likelihood* (ML) approach based on [3] in a preceding off-line procedure. The RT is estimated in speech offset periods only which are determined by a *voice activity detector* (VAD). To all obtained speech offset segments, which are larger than 200 ms, the ML procedure is applied and the results are averaged. From this estimation procedure, using 45 min of speech material, an average reverberation time of 0.86 s was obtained in the considered scenario.

2.3 Direct-to-Reverberant Energy Ratio (DRR)

A further very important characterization of a room impulse response are channel-based measures [4]. Among them, the DRR is the most important one and is defined as

$$\frac{\mathrm{DRR}}{[\mathrm{dB}]} = 10 \cdot \log_{10}\left(\frac{\sum\limits_{k=0}^{k_d} h^2(k)}{\sum\limits_{k=k_d+1}^{L_r} h^2(k)}\right), \tag{4}$$

where L_r denotes the length of the discrete impulse response and k_d the discrete time index where the direct sound ends. This value is usually chosen such that a few early reflections are included in the direct path, i.e., global maximum plus 2 ms. The DRR also determines the critical distance d_c of a sound event. The critical distance is defined as the distance from the source at which the sound energy due to the direct-path component is equal to the sound energy due to reverberation. Hence, the following two cases have to be distinguished:

- DRR $< 0\,$dB $\rightarrow$ Source outside the critical distance d_c,
- DRR $> 0\,$dB $\rightarrow$ Source within the critical distance d_c.

As for the RT, an off-line estimation procedure is used. Here, we determine the energy drop of the signal after a sharp speech offset by manual segmentation. The resulting DRR was determined as 18 dB, which indicates that the source is located within the critical distance.

3 Psychacoustically-Motivated Dereverberation

3.1 Conventional Dereverberation

One state-of-the art dereverberation algorithm is based on a statistical model of late reverberation [5]. The basic idea is to estimate the *power spectral density* (PSD) of the late reverberant speech components and to formulate a weighting rule that aims to suppress late reverberant components while leaving the direct path speech components and early reflections unaltered.

This subsection describes briefly an improved single-channel algorithm based on [5] which utilizes a generalized statistical model of the room impulse response according to [6]. This generalization allows to use the algorithm also in situations where the source is located within the critical distance, which is the case for the parliament recordings.

The considered room impulse response is described as a sequence of i.i.d. random variables $b(k)$ with zero mean and normal distribution, multiplied by an exponentially decaying function. In the generalized approach proposed in [6], the RIR is divided into two segments: one segment which corresponds to the direct path and early reflections and the second segment which describes late reverberation. Hence, this model can distinguish between early and late reverberation and is given by

$$h(k)|_{(\mathrm{Gen})} = \begin{cases} b_{\mathrm{d}}(k)\,e^{-\rho k/f_s} & \text{for } 0 \le k < k_d \\ b_{\mathrm{r}}(k)\,e^{-\rho k/f_s} & \text{for } k \ge k_d \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

with sampling frequency f_s in Hz and k_d which is chosen as in Eq.(4). $b_{\mathrm{d}}(k)$ and $b_{\mathrm{r}}(k)$ represent two uncorrelated noise sequences of early and late reverberation, respectively, which are both i.i.d. random variables with zero mean and normal distribution. The variances of $b_{\mathrm{d}}(k)$ and $b_{\mathrm{r}}(k)$ are denoted by σ_d^2 and σ_r^2 in the following. It is further assumed that $\sigma_d^2 \ge \sigma_r^2$ (high DRR, within the critical distance).

Based on the generalized statistical model, an improved estimator for the late reverberant PSD can be expressed by [6]

$$\sigma^2_{x_{\mathrm{late}}}(\lambda,\mu) = (1-\kappa(\mu))\cdot e^{-2\rho(\mu)\mathrm{T}_l}\cdot\sigma^2_{x_{\mathrm{late}}}(\lambda-1,\mu) + \kappa(\mu)e^{-2\rho(\mu)\mathrm{T}_l}\cdot\sigma^2_x(\lambda-N_l,\mu), \tag{6}$$

where T_l marks the time span after which the late reverberation begins. The PSD of the reverberant speech is denoted by $\sigma^2_x(\lambda,\mu)$ and N_l indicates the (integer) number of frames corresponding to T_l. The constant $\kappa(\mu)$ is inversely proportional to the direct-to-reverberant energy ratio [6].

Please note that for the special case $\kappa = 1 \,\forall\, \mu$, the estimator reduces to the approach by [5].

In order to estimate the a posteriori *signal-to-interference ratio* (SIR)

$$\eta(\lambda,\mu) = \frac{|X(\lambda,\mu)|^2}{\sigma^2_{x_{\text{late}}}(\lambda,\mu)}, \tag{7}$$

the spectral variance of the reverberant speech is calculated by recursive averaging

$$\sigma_x^2(\lambda,\mu) = \alpha \cdot \sigma_x^2(\lambda-1,\mu) + (1-\alpha)\cdot|X(\lambda,\mu)|^2, \tag{8}$$

with a smoothing factor $0 \leq \alpha \leq 1$. The weights for the suppression of the late reverberant components can be calculated, e.g., by a spectral magnitude subtraction rule

$$\widetilde{G}(\lambda,\mu) = 1 - \frac{1}{\sqrt{\eta(\lambda,\mu)}}. \tag{9}$$

It should be noted that a frequency-dependent DRR and RT could also be employed. However, for the desired recordings, no improvements could be observed and hence, the frequency-independent values are kept in order to reduce the computational complexity.

The direct application of the spectral gains to the reverberant DFT coefficients can lead to various artifacts such as musical noise. In order to counteract such artifacts, different methods are possible:

- Apply a high spectral floor to the gains which however, reduces the dereverberation performance or
- Smoothing of the spectral weights in the frequency domain [7] or cepstral domain [8].

3.2 Psychoacoustic Weighting

In this contribution we propose to apply a psychoacoustic weighing rule which was initially developed to reduce artifacts in noise reduction systems [9] and in a combination with acoustic echo control [10]. Related psychoacoustic dereverberation approaches are discussed in [11].

The overall block diagram of the new system is depicted in Fig. 3. The main idea is to perform a pre-dereverberation of the input spectra and to calculate the masking thresholds from the pre-enhanced signal. Based on an estimate of the late reverberant PSD and the masking threshold, the final spectral weights are calculated and applied to the reverberant signal. The processing steps are as follows:

1. Estimate the late reverberant PSD $\sigma^2_{x_{\text{late}}}(\lambda,\mu)$ using the method described above Eq.(6).
2. Calculate preliminary spectral gains $\widetilde{G}(\lambda,\mu)$ by means of the spectral subtraction rule in Eq.(9).
3. Compute a pre-dereverberated signal by

$$\widetilde{X}(\lambda,\mu) = X(\lambda,\mu)\cdot\widetilde{G}(\lambda,\mu). \tag{10}$$

4. Estimate masking threshold $\overline{X}(\lambda,\mu)$ based on $\widetilde{X}(\lambda,\mu)$ using the ISO model [12].

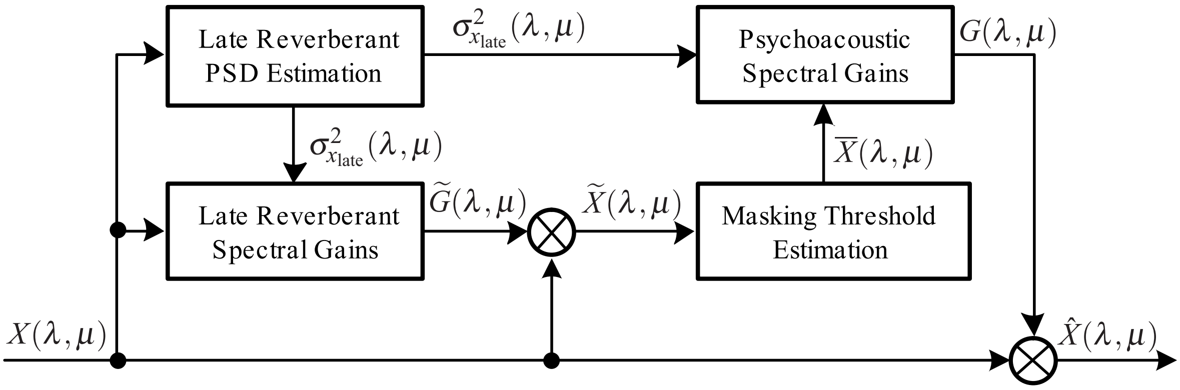
Late Reverberant PSD Estimation
$\sigma^2_{x_\text{late}}(\lambda,\mu)$
Psychoacoustic Spectral Gains
$G(\lambda,\mu)$
$\sigma^2_{x_\text{late}}(\lambda,\mu)$
$\overline{X}(\lambda,\mu)$
Late Reverberant Spectral Gains
$\widetilde{G}(\lambda,\mu)$
$\widetilde{X}(\lambda,\mu)$
Masking Threshold Estimation
$X(\lambda,\mu)$
$\hat{X}(\lambda,\mu)$

Figure 3 - Block diagram of the considered single-channel speech dereverberation system using a psychoacoustically motivated spectral weighting rule.

5. Calculate psychoacoustic weighting gains $G(\lambda,\mu)$ by [9]:

$$G(\lambda,\mu) = \min\left(\sqrt{\frac{\overline{\text{X}}(\lambda,\mu)}{\sigma^2_\text{x}(\lambda,\mu)}} + \zeta, 1\right), \tag{11}$$

   with an interference attenuation factor ζ. Additionally, a lower bound G_min is applied to the weighting gains.

6. Perform the final dereverberation by applying the psychoacoustic gains to the reverberant input spectra by

$$\hat{S}(\lambda,\mu) = X(\lambda,\mu)\cdot G(\lambda,\mu). \tag{12}$$

4 Simulations

In a first step the above mentioned off-line procedure to determine the RT and DRR was carried out. In a second step the single-channel recordings were processed with and without the psychoacoustical extension of the conventional dereverberation algorithm using Eq.(11) and Eq.(9), respectively. The corresponding time-domain signals are termed $\tilde{x}(k)$ and $\hat{x}(k)$ (see Fig. 3). Further important simulation parameters are listed in Table 1. The corresponding audio files are available online [1].

Since neither the room impulse response nor any anechoic reference signal is available, only non-intrusive objective quality measures can be used. Table 2 shows the results in terms of the *Speech- to-Reverberation Modulation energy Ratio* (SRMR) [13], which is a suitable indicator for the reverberation suppression.

The corresponding spectrograms of the enhanced signals are shown in Fig.4. It can be seen that due to the psychoacoustic weighting (right subplot), random fluctuations, i.e. musical tones, could be reduced significantly. Besides, the reduction of musical tones and further artifacts was confirmed by the subjective listening impression.

[1] http://www.ind.rwth-aachen.de/~bib/jeub11b

Table 1 - Main simulation parameters.

Parameter	Settings
Sampling frequency	$f_s = 32\,\mathrm{kHz}$
Frame length	$L = 640$ (20 ms)
FFT length	$M = 1024$ (including zero-padding)
Frame overlap	50% (Hann window)
Smoothing factors	$\alpha = 0.9$
Reverberation time	$\mathrm{T}_{60} = 0.86\,\mathrm{s}$
Late reverberant time span	$\mathrm{T}_l = 0.1\,\mathrm{s}$
DRR	$\mathrm{DRR} = 18\,\mathrm{dB}$
Interference attenuation factor	$\zeta = -15\,\mathrm{dB}$
Gain threshold	$G_{\min} = 0.1$

Table 2 - Dereverberation performance in terms of the non-intrusive quality measure SRMR.

	Reverberant $x(k)$	Pre-dereverberated $\tilde{x}(k)$	Dereverberated $\hat{x}(k)$
SRMR	7.96	8.66	8.79

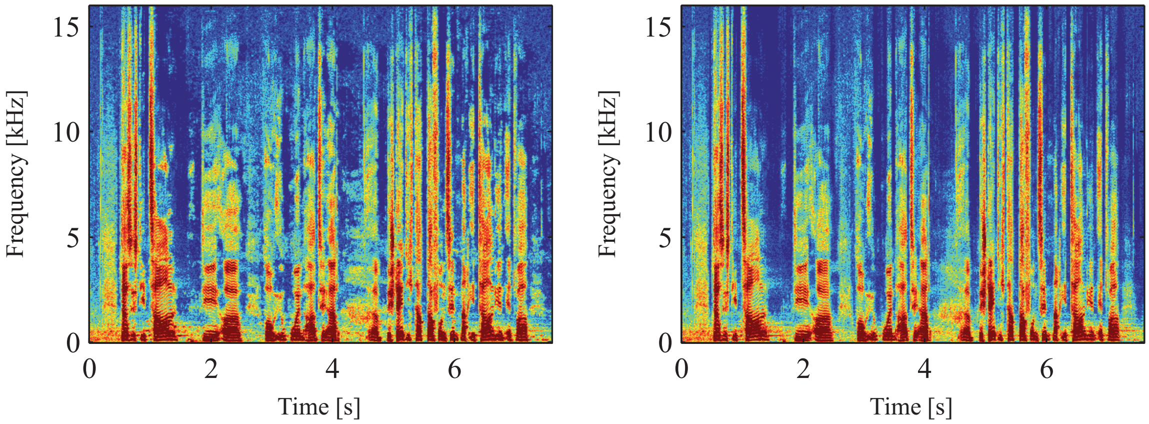

Figure 4 - Spectrogram of processed speech: (left) with pre-dereverberation only ($\tilde{x}(k)$); (right) with dereverberation using psychoacoustic weighting ($\hat{x}(k)$).

5 Conclusions

In this contribution, we have demonstrated the application of speech dereverberation techniques, which are commonly employed in hearing aids or hands-free speech communication systems, for recordings taken in the German parliament.

Based on a new psychoacoustically-motivated dereverberation concept, the drawbacks of the conventional system such as musical tones, could be reduced significantly. In future applications, the developed algorithm might be used for a post-processing of recorded data, e.g., for news broadcast or as a plug-in for video players such as the VLC. Besides, the enhanced recodings can be used for archival storage and documentation.

Acknowledgment

The authors would like to thank Steven Rösler from the German Bundestag parliament television for providing the audio recordings as well as Andreas Witte and Matthias Rüngeler from the IND for programming support and helpful discussions.

References

[1] P. Naylor and N. Gaubitch, Eds., *Speech Dereverberation*, Springer, London, 2010.

[2] M.R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America (JASA)*, vol. 37, no. 3, pp. 409–412, 1965.

[3] H.W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, 2010.

[4] H. Kuttruff, *Room Acoustics*, Spon Press, Oxon, 2009.

[5] K. Lebart, J. M Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

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

[7] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[8] T. Gerkmann, *Statistical Analysis of Cepstral Coefficients and Applications in Speech Enhancement*, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 2010.

[9] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, 1998.

[10] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[11] A. Tsilfidis, *Signal Processing Methods for Enhancing Speech and Music Signals in Reverberant Environments*, Ph.D. thesis, University of Patras, Patras, Greece, 2011.

[12] ISO/IEC 11172-3:1993, *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s - Part 3: Audio*, ISO/IEC, 1993.

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