Classification of Reverberant Acoustic Situations

Jens Schröder¹, Thomas Rohdenburg², Volker Hohmann¹, Stephan D. Ewert¹

¹ University of Oldenburg, Institute of Physics, Medical Physics, Germany,
² Fraunhofer Institute for Digital Media Technology, 26129 Oldenburg, Germany

Email: jens.schroeder@uni-oldenburg.de; stephan.ewert@uni-oldenburg.de

Introduction

In daily communication, speech intelligibility depends on the acoustic surrounding or acoustic situation. Particularly for hearing impaired persons, speech understanding is often problematic if speech is distorted by (room) reverb, noise or competing talkers. Acoustic situations are characterized by different dominating types of distortion. Hearing aids might provide appropriate algorithms to enhance speech intelligibility in the different acoustic situations. A robust and fast automatic classification of the acoustic situation should therefore select the appropriate hearing aid algorithm without requiring an action of the hearing aid wearer. This study is concerned with the automatic estimation of the reverberation time (T60) in natural situations and with unknown excitation signal. Acoustic test situations were generated by convolving speech signals with artificial and real room impulse responses with T60 times ranging from 0.05 to 4 s. Features derived from the cepstral mean, the autocorrelation function and from the distribution of modulation energy were used to blindly estimate different reverb times.

Impulse Response Model

In natural environments sound is often received as a superposition of direct and reflected sound from walls or objects in a room. Direct, early reflexions arrive first at the ear and after multiple reflexions and superpositions of reflexions from many objects they are diffuse and called reverberation. The level of reflexions in a room impulse response decreases due to attenuation effects and scattering. This decay is often assumed to be nearly exponential ([1],[6]). Thus, a reverberant impulse response can be approximated by an exponentially decaying part and if measurement noise or background noise occurred by a constant part [1]:

\[ h(t) = A_{\text{exp}}e^{-t/\tau}n_1(t) + A_{\text{noise}}n_2(t), \]  

(1)

where \( A_{\text{exp}} \) und \( A_{\text{noise}} \) are scalar, \( \tau \) is the decay parameter in seconds, \( t \) is the time in seconds and \( n_1(t) \) and \( n_2(t) \) present two independent noise processes.

A common measure for reverberation is the time until the impulse response has decreased by 60 dB. In (1), the reverberation time, T60, can be calculated directly from the decay parameter \( \tau \):

\[ T60 = -\ln(10^{-3}) \cdot \tau \approx 6.908 \tau. \]  

(2)

To calculate the T60 time from a measured impulse response different solutions exist. A procedure suggested in [1] is to fit the power by a least square fit. From equation (1) the instantaneous power can be derived as:

\[ a(t) = \sqrt{A_{\text{exp}}^2 e^{-2t/\tau} + A_{\text{noise}}^2}. \]  

(3)

The parameters \( A_{\text{exp}}, \tau \) and \( A_{\text{noise}} \) are evaluated by a least square fit of the form

\[ \min \int |a(t) - y(t)|^2 dt \]  

(4)

where \( s = 0.5 \) is a scaling factor to improve the results [1]. The T60 time was then calculated as stated in (2).

Blind Estimation Procedures

The goal of this study is to estimate the T60 time from a reverberated speech sample without having explicit information about neither the impulse response nor the underlying clean speech.

Three different methods are used in the following and their results are compared.

Cepstral Mean

To estimate the impulse response from an unknown reverberated signal there exists the theory of "blind homomorphic deconvolution" [3], [4], [5]. Here a reverberated speech signal is assumed as:

\[ \text{sr}(t) = s(t) \ast h(t), \]  

(5)

where \( \ast \) denotes the convolution product, \( s(t) \) clean speech and \( h(t) \) the impulse response. A Fourier transformation turns the convolution into a multiplication. The logarithm transforms this product into a sum. A further inverse Fourier transformation converts the sum into the cepstral domain where the additivity is preserved (see figure 1). If it is assumed that the cepstrum

\[ \text{sr}(t) \ast h(t) \overset{F}{\rightarrow} S(f) \cdot H(f) \overset{\log}{\rightarrow} \tilde{S}(f) + \tilde{H}(f) \overset{\mathcal{F}^{-1}}{\rightarrow} \hat{s}(q) + \hat{h}(q) \]

Figure 1: Calculation of the cepstrum from a convoluted input signal

of the clean speech is nearly uncorrelated in adjacent windows, averaging over a few cepstra estimates the mean cepstrum of the impulse response. By inverting \( h(t) \) can be derived.

To keep the inverse cepstrum stable and causal, only the minimum phase part of the deconvolved impulse response is taken [3]. The whole deconvolution scheme by cepstral mean is shown in figure 2. The T60 time can then be estimated from the resulting impulse response by the above mentioned least square fit.
In comparison to [7], the weighting of the high and low SRMR (Speech to Reverberation Modulation energy ratio) respective energies was calculated. This ratio is called high modulation frequency regions and the ratio of the modulation spectrogram was separated into low and frequencies. Thus [7] suggested a comparison of energies of white noise. The higher the reverberation time, the consequence of the whitening effect by the impulse response was estimated due to equation (4). Typically, clean speech shows the strongest modulation frequencies were defined to be smaller than 28.9 Hz or smaller than 10% of the corresponding bandwidth of the auditory filter.

**Stimuli and Methods**

To analyse the three estimation methods, clean speech material of four speaker sets sampled at 16 kHz was used: Two male, German speakers, one female, German speaker and a set of female, English speakers. The clean speech material was reverberated by convolution with room impulse responses. To characterize the impulse responses, the T60 times were estimated from the impulse responses using the method described in (4) and are referred to as real T60 times in the following.

For the first test setup, referred to as artificial impulse response setup (artificial IR setup), impulse responses were generated by source imaging [2] to achieve a wide range of T60 times with controllable spectral (white) properties. For these impulse responses a rectangular room was simulated and its size and reflection coefficients were adjusted to produce T60 times ranging from 50 ms to 4 s, increasing at a factor of two. The sound source and the receiver were distributed randomly in the room and the positions differed between the impulse responses. Independent of the T60 time the length of each impulse response was 2 s with 16 kHz sampling frequency. All reverberant speech samples were derived from the same underlying clean speech. In this paper, the results for one of the male German speakers are shown. The second test setup, called real impulse response setup (real IR setup), consisted of real impulse responses selected from a commercial impulse response library. Three groups of T60 times were defined: a "dry" one (0.16 - 0.36 s) with a mean T60 = 0.3 s, a medium reverberated one (0.72 - 1.02 s) with a mean T60 = 0.9 s and a reverberant one (1.71 - 1.98 s) with a mean T60 = 1.9 s. Each group consisted of four impulse responses. Each of the four impulse responses of a group are convolved with different speech material from one of the four speaker sets.

To test all three features, an estimation of the T60 time respectively the SRMR was done every 0.5 s for a total time of 100 seconds for the artificial IR setup and 40 seconds for each speaker/impulse response of a T60 group for the real IR setup.

For the cepstral mean and the autocorrelation feature the window lengths were 0.05 s, 0.2 s, 0.5 s, 1 s, 2 s, 3 s, and 4 s. The overlap between two windows was 7/8 and the averaging time 5 times the window length (≈ 33 windows). The first 6.3 ms (≈ 100 time samples at 16 kHz sampling frequency) of the deconvolved impulse responses were skipped for the fitting. The window length for the SRMR feature was 1 s.

**Results**

**Cepstral Mean**

The means of 200 T60-time estimates for the artificial IR setup are plotted in Fig. 4 (solid lines) as a function of the analysis window duration for three different real T60 times indicated by the dotted lines. The T60-time estimates depend on the analysis window duration,
starting at small values for short window durations and asymptoting against the real T60 times with increasing window duration. The T60-time curves flatten off at a window duration corresponding to about two times real T60 time. Since the estimates depend on the window duration, the proper window for a good T60-time estimation has to be chosen blindly. To do so, the algorithm suggested here successively calculates T60-time estimates for increasing analysis window durations. Following the slope analysis given above, the validity of the estimates is judged blindly by monitoring the differences of the T60-time estimates derived for two successive analysis window durations. If the slope calculated from the two last estimates is smaller than an empirically adjusted criterion of 0.1, the last estimate is considered to be valid.

A second criterion to judge the validity of the estimate is to monitor the ratio between the energy of the fitted noise and the energy in the fitted exponential decay calculated over the duration of the current T60-time estimate:

$$\frac{\sum_{t=0}^{T60} A_{noise}^2}{\sum_{t=0}^{T60} (A_{exp} \exp(-\tau t))^2} = \frac{N}{S}, \quad (8)$$

Again, an empirically adjusted criterion of $\frac{N}{S} < 0.25$ has to be met in order to judge the estimate as valid. If both criterions are met, a valid T60-time estimate was calculated by the algorithm.

The means of the valid T60-time estimates are plotted in Fig. 5 as a function of the real T60 times. The left panel shows the results for the artificial IR setup and the right panel for the real IR setup. The numbers indicate the percentage of the valid estimates in relation to all estimates. Left panel: artificial IR set up, 200 estimations per impulse response. Right panel: real IR set up, 320 estimations per impulse response group.

For the autocorrelation feature, the same sound material and parameters as for the cepstral mean feature were used (see above).

The means of 200 T60-time estimates per window and impulse response of the artificial IR setup are plotted in Figure 6, comparable to Figure 4. Comparable to the cepstral mean feature, the estimated T60 times of the autocorrelation feature asymptote against the real T60, though the deviations are much higher here than in case of the cepstral mean feature. The prolonging of the analysis window duration was again used and valid estimates were derived when the same slope and $\frac{N}{S}$ criteria as in case of the cepstral mean feature were met. The results for the valid T60-time estimates are shown in Figure 7. The results are similar to those from the cepstral mean. Again, there is a lower limit at about 200 ms for the T60-time estimates. Above 200 ms the means of estimated T60 times are matching the real ones well, though the deviations are larger than those of cepstral mean. The number of valid T60 times decreases heavily for all impulse responses.
responses. The numbers show the percentage of valid T

Figure 8: Autocorrelation: Means and standard deviations of calculated SRMRs per impulse response over the real T60 time. The large standard deviations indicate that even a meaningful classification in rough T60-time categories would fail without averaging over a large number of SRMRs.

Conclusions
Three different methods for the estimation of the reverberation time T60 were presented. It was shown that the cepstral mean and the autocorrelation feature are useful in combination with the cepstral mean or autocorrelation feature. In the next step all three features are combined in a classifier with a gaussian mixture model (GMM) for robustness.

Acknowledgements
This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) project ”Modellbasierte Hørsysteme”.

References