Room Transfer Function Estimation Using Cepstral Smoothing

Thomas Tomczyszyn, Benjamin Cauchi, Stephan Gerlach, Stefan Goetze
Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, 26129 Oldenburg, Germany
{thomas.tomczyszyn, benjamin.cauchi, stephan.gerlach, s.goetze}@idmt.fraunhofer.de

Introduction

Many state-of-the-art applications nowadays make use of a distant microphone to receive the speech signal of a user, e.g. hands-free communication systems, automatic speech recognition or acoustic speaker localization. In an enclosed space this signal is corrupted by reverberation caused by the influence of the room impulse response (RIR) which depends on the characteristics of the room and on the positions of the user and the microphone. As reverberation can decrease speech intelligibility as well as quality, several dereverberation algorithms have been developed in the past. Some dereverberation algorithms, e.g. channel equalization, need an estimate of the time domain RIR typically obtained using blind system identification (BSI) while others, e.g. reverberation suppression, can be fed only with parameters of the RIR such as the reverberation time (T_{60}) or the direct to reverberant ratio. An estimate of the amplitude of the room transfer function (ARTF) could be used to improve the performance of BSI algorithms, as a parameter within reverberation suppression methods or source localization algorithms. An adaptive, blind, single-channel estimator of the amplitude of the room transfer function based on cepstral smoothing is proposed in this contribution.

Notations

Consider an acoustic system with a single source and a single microphone. The input signal $z[n]$ received by the microphone can be expressed as

$$z[n] = \underbrace{s[n] * h[n]}_{y[n]} + v[n], \tag{1}$$

where $s[n]$ denotes the clean source signal, $h[n]$ denotes the RIR between the source and the microphone and $v[n]$ denotes the additive noise. In the short-time Fourier transform (STFT) domain, (1) becomes

$$Z[k,\ell] = Y[k,\ell] + V[k,\ell], \tag{2}$$

where $Y[k,\ell]$ and $V[k,\ell]$ denote the STFT representations of $y[n]$ and $v[n]$, respectively, while k and ℓ denote the frequency bin and the frame index. Assuming that the convolution theorem holds in the STFT domain $Y[k,\ell]$ becomes

$$Y[k,\ell] \approx S[k,\ell] H[k], \tag{3}$$

where $H[k]$ is the discrete Fourier transform (DFT) of $h[n]$. This paper aims to estimate the ARTF $|H[k]|$. In the absence of noise, the cepstrum $Z_{\text{cep}}[q,\ell]$ of $z[n]$ is expressed as

$$Z_{\text{cep}}[q,\ell] = \text{IDFT}\left\{\log\left(|Y[k,\ell]|\right)|_{k=0,1,\cdots,L}\right\}, \tag{4}$$

where q denotes the quefrency bin index, L the length of the inverse discrete Fourier transform (IDFT) and $\log(\cdot)$ the natural logarithm. The combination of (3) and (4) leads to

$$Z_{\text{cep}}[q,\ell] = S_{\text{cep}}[q,\ell] + H_{\text{cep}}[q], \tag{5}$$

which, assuming that $H_{\text{cep}}[q]$ is time invariant, leads to

$$\mathcal{E}\left\{Z_{\text{cep}}[q,\ell]\right\} = \mathcal{E}\left\{S_{\text{cep}}[q,\ell]\right\} + H_{\text{cep}}[q]. \tag{6}$$

In addition, $S_{\text{cep}}[q,\ell]$ is considered as a zero-mean process leading to

$$H_{\text{cep}}[q] = \mathcal{E}\left\{Z_{\text{cep}}[q,\ell]\right\}. \tag{7}$$

In practice, $\mathcal{E}\left\{Z_{\text{cep}}[q,\ell]\right\}$ can be estimated as the average over all frames of an occurrence or, in real applications, using temporal smoothing (TS) in the cepstral domain. TS leads to a time-dependant estimate $\hat{H}_{\text{cep}}[q,\ell]$,

$$\hat{H}_{\text{cep}}[q,\ell] = \beta \hat{H}_{\text{cep}}[q,\ell-1] + (1-\beta) Z_{\text{cep}}[q,\ell], \tag{8}$$

where $0 \leq \beta < 1$ is a forgetting factor which controls the delay in adapting $\hat{H}_{\text{cep}}[q,\ell]$ after potential changes in $h[n]$. Finally, the estimate $|\hat{H}[k,\ell]|$ of the ARTF can be obtained by transforming $\hat{H}_{\text{cep}}[q,\ell]$ back to the linear frequency domain as

$$|\hat{H}[k,\ell]| = \exp\left(\text{DFT}\left\{\hat{H}_{\text{cep}}[q,\ell]\right\}\right). \tag{9}$$

Experiment

In the following, the estimation error of the ARTF is measured as

$$\epsilon = 20\log_{10}\left(\frac{\sqrt{\sum_{k=0}^{L-1}\left(|H[k]| - |\hat{H}[k]|\right)^2}}{\sqrt{\sum_{k=0}^{L-1}|H[k]|^2}}\right).$$

All signals are sampled at $f_s = 16$ kHz. The STFT has been computed using a Hann window of length L_w and an overlap of 50%. First, the influence of L_w on the assumption from (3), i.e. application of the convolution theorem in the STFT domain, is investigated using artificial RIRs. Second, the influences of the forgetting factor β and the performance of the estimator using recorded RIRs are examined.

The research leading to these results has received funding from the EU Seventh Framework Programme project DREAMS under grant agreement ITN-GA-2012-316969.

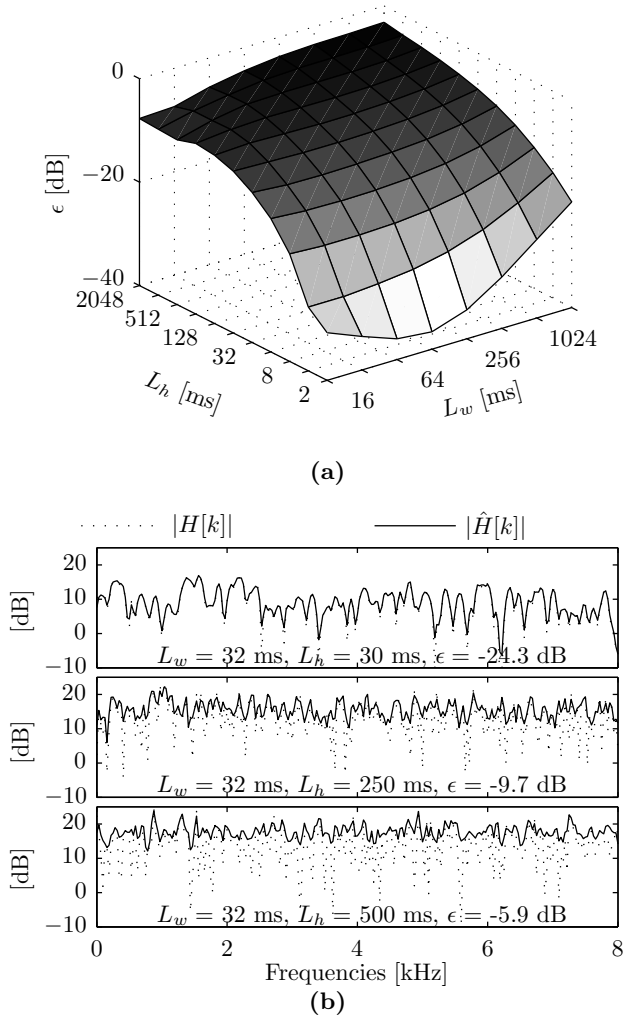

Figure 1: Influence of the window length on the estimation of the ARTF. (a) Estimation error for different combinations of window L_w and RIR lengths L_h. (b) True and estimated ARTF for $L_w = 32$ ms and L_h set to 30, 250 and 500 ms.

Convolution theorem

It is well known that the convolution theorem does not hold exactly in the STFT domain [1], therefore, the validity of (3) is to be verified. In addition, the influences of the windowing may be neglected if the length of the RIR L_h is short compared to L_w [2]. The accuracy of the estimation is analysed here on artificial signals for different values of L_w and lengths L_h. A Gaussian white noise signal of length 30 s is used here as the source signal in order to ensure stationarity of $S[k,\ell]$ and a zero-mean cepstrum $S_{\text{cep}}[q,\ell]$. The RIRs have been generated using the model presented in [3] in order to allow for very small values of L_h (down to 2 ms). The expected value of the input cepstrum has been estimated as the average over all $Z_{\text{cep}}[q,\ell]$. It can be seen in Fig. 1(a) that a low estimation error is obtained for short RIRs while the error increases with greater values of L_h. Furthermore, it shows that the use of a long analysis window does not provide an improvement when estimating the ARTF of a long RIR. This could be due to the fact that the simulation is based on a fixed input signal length and, thus, less

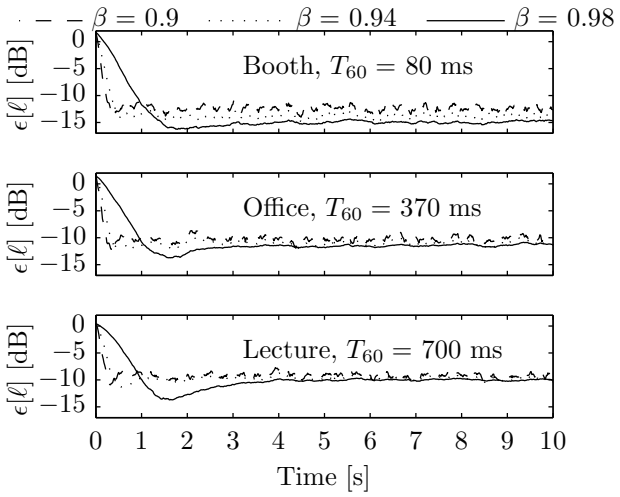

Figure 2: Influences of the forgetting factor β on temporal smoothing for different reverberation times T_{60} ($L_w = 32$ ms).

frames are available for the estimation of $\mathcal{E}\left\{Z_{\text{cep}}[q,\ell]\right\}$ for higher values of L_w. Fig. 1(b) shows the increasing deviation between the true and the estimated ARTF for higher values of L_h.

Temporal Smoothing

In real applications the entire input signal is not available and TS is used in order to estimate the expected value of the input cepstrum. In addition, L_w is often constrained by the period during which the source signal can be considered stationary. A window of length $L_w = 32$ ms is used here, as it is a typical value, e.g. in speech processing. RIRs from [4], recorded in a booth, a meeting room and a lecture room have been used. Fig. 2 illustrates how lower values of β allow for a faster adaptation of the estimate while higher values lead to a slightly better final estimation error. The value of $\epsilon[\ell]$, after convergence, is in the range of -15 to -10 dB and increases with T_{60}.

Conclusion

This paper investigates the use of TS in the cepstral domain to estimate the ARTF. The influences of the length of the analysis window of the STFT have been analysed. It is shown that the use of a long analysis window is not beneficial. However, the described estimator could be of interest for single-channel applications.

References

[1] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[2] J. Benesty, M. Mohan Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, 2008.

[3] J. D. Polack, "Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics," *Applied Acoustics*, vol. 38, no. 2, pp. 235–244, 1993.

[4] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Sydney, Australia, July 2009, pp. 1–4.