Real-time Room Reverberation Estimation for Online Speech Intelligibility Monitoring

Jens Schröder[1], Jan Rennies[1], Feifei Xiong[1], Jörn Anemüller[2], Stefan Goetze[1]
[1] *Fraunhofer IDMT / Hearing, Speech and Audio Technology, Oldenburg, Email: jens.schroeder@idmt.fraunhofer.de*
[2] *University of Oldenburg, Institute of Physics, Medical Physics*

Introduction

Speech is an important element in daily life. Unfortunately, speech intelligibility is influenced and often reduced by distortions. One of these distortions is reverberation. Reverberation is caused by the reflections of an acoustic signal at objects and walls superposing at the position of a receiver. A common measure for reverberation is the reverberation time T_{60}. It is defined as the decrease of an impulse response by 60 dB. In [1] methods were described that are able to blindly estimate the reverberation time T_{60} from speech signals. In this paper, these estimators are used to estimate the speech intelligibility based on the speech transmission index (STI) [2]. The influence of different noises in different signal-to-noise ratio (SNR) conditions on the STI estimation in real and artificial environments is presented.

Blind Speech Intelligibility Transmission Estimation

In this paper two different estimation methods for estimating the STI are used. Both use the estimation methods of the reverberation time T_{60} as presented in [1]. One of the procedures is based on cepstral mean estimation using the redundancy of speech to eliminate it in the cepstral domain and thus deconvolve the room impulse response (RIR). The second proposed procedure, the autocorrlation method, as well exploits the redundancy of speech to gain an autocorrelation function decaying in the same manner as the underlying RIR [3].
The T_{60} is a technically motivated measurement describing the decreasing time of the RIR by 60 dB. A more perceptually motivated meaurement for reverberation and its influence on speech intelligibility is the speech transmission index (STI). Steeneken and Houtgast empirically developed a formula for the influence of reverberation on the STI [4]. The modulation transfer function for reverberation, being part of the STI computation, is given by

$$m_f = \left[1 + \left(\frac{2 \cdot \pi \cdot f \cdot T_{60}}{13.8}\right)^2\right]^{-\frac{1}{2}} \quad , \qquad (1)$$

where f are 14 modulation frequencies ranging from 0.63 Hz to 12.5 Hz [2]. Thus, with knowledge of T_{60} the reverberation part of the STI can be computed directly (cf. [4]).

Experimental Setup

To blindly estimate the reverberation part of the STI from an unknown reverberant speech signal, two algorithms presented in [1] were used. To get valid estimations, a method with termination conditions was implemented [1]. In this paper the window lengths l_w are different from [1] because here the minimal observed T_{60} was longer: $l_w = \{0.2\text{ s}, 0.5\text{ s}, 1\text{ s}, 2\text{ s}, 4\text{ s}, 8\text{ s}\}$. All other parameters were chosen as suggested in [1]: overlap $= 7/8 \cdot l_w$; averaging time $= 5 \cdot l_w$; $\frac{N}{S} < 0.25$; $\frac{\Delta T_{60}}{\Delta l_w} < 0.1$.
From the deconvoluted impulse reponses, the reverberation part of the STI is estimated as described in [4]. These estimations are compared with the ground-truth T_{60} which is directly computed from the known RIR using the tool from [5].

To evaluate the estimators, clean speech was synthetically reverberated by convolution with RIRs. Six clean speech stimuli of 40 s were generated by randmoly concatenating sentences of the Oldenburg sentence test [6].
Two different setups of RIRs were used. Artifical ones were generated by the image method [7] using the implementation of [8] for a rectangular room $(4.5 \times 5.5 \times 2.5)$ m^3. Source and receiver were distributed randomly, but not directly at walls. The absorbtion coefficients were adjusted to generate RIRs with T_{60} times between 0.25 s and 6 s. The second RIR setup consisted of RIRs recorded in real environments. The T_{60} ranged from 0.14 s to 6.67 s.
The reverberant speech was tested with different SNRs. Gaussian white noise (GWN) and "olnoise" [6], that has a long-term speech spectrum, were used as disturbances. The noises were reveberated by RIRs differing from those of the speech signals but arising from the same room. By this, a room was simulated having different positions for speech and a noise sources.
The sampling frequency was $f_s = 44.1$ kHz.

Results

The stimuli setups were tested for each estimation method under nine different SNR conditions. In Figure 1 the estimated STI is plotted as function of the ground-truth STI for measured and artificial RIRs for the cepstral mean procedure disturbed with GWN at a SNR level of 0 dB. The estimations for the artificial RIRs are more accurate than the estimations for the measured ones. For the measured RIRs, the STI shows a tendency to be overestimated.

The correlation coefficient r between estimated and

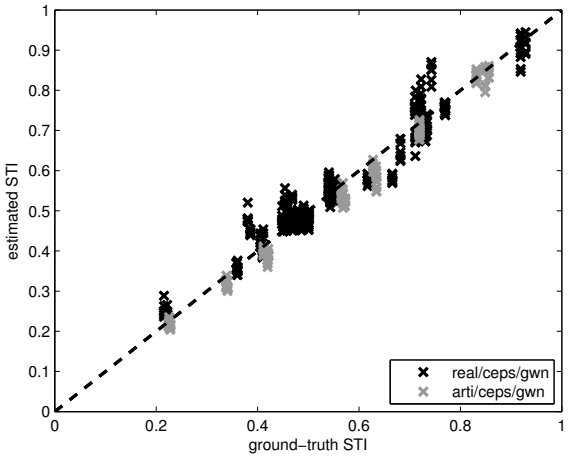

Abbildung 1: The estimated STI over the ground-truth STI for the measured (black) and artificial (gray) RIRs for the cepstral mean method and GWN at 0 dB SNR. The correlation coefficient between ground-truth and estimated STI is $r_{\mathrm{arti}} = 1.00$ for the artificial and $r_{\mathrm{real}} = 0.97$ for the measured RIR configuration.

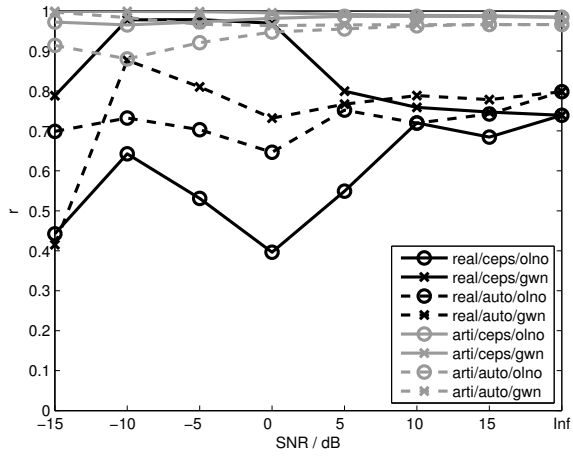

Abbildung 2: The correlation coefficient r between ground-truth and estimated STI over the SNR for eight different RIR/method/noise configurations. The configurations are indicated by different line styles (measured RIR: black | artificial RIR: gray / cepstral mean: solid line | autocorrelation: dashed line / Olnoise: circles | GWN: crosses).

ground-truth STI is shown as a function of the SNR in Figure 2. Results show that r is higher for artificial than for measured RIRs for every tested SNR condition. The correlation coefficient r of the artificial estimations range between 0.88 and 1.00, the measured ones between 0.40 and 0.98. The STI estimation for the artificial RIRs shows small dependency on the SNR especially for SNR below 0 dB. For GWN the the estimations get slightly better for low SNR, for olnoise slightly worse. Above 0 dB, the estimations by the autocorrelation and cepstral method are very similar. The estimations for the measured RIRs do not behave as clearly as the artificial across the SNR, especially for the cepstral mean method. While the cepstral mean works better than the autocorrelation method for all artificial RIRs and the same noise type, for measured RIRs the autocorrelation procedure produces better estimations for SNR above 10 dB. The dependency on the noise (and SNR) is presumable due to the assumption that the exciting signal (here speech) must be highly redundant. For artificial RIRs, olnoise seems to have decremental influence on the estimation and violates this assumption while GWN fullfils this redundany criterion perfectly. The autocorrelation method for meaured RIRs seems to dependent less on this assumption than the cepstral mean procedure.

Conclusion

In this paper, two blind estimation procedures (cepstral and autocorrelation based) for the reverberation part of the STI were evaluated. Two different types of RIRs (measured and artificial) and two noises (GWN and olnoise) were tested in different SNR conditions. It was shown that in artificial conditions, the estimation methods work better than in real ones, though in real conditions the estimations are also reasonable. For real RIRs the autocorrelation method leads to better results for SNR above 10 dB. For artificial environments, the cepstral feature works better for all SNR values.

Acknowledgement

The measured RIRs were kindly provided by "Akustikbüro Oldenburg".

Literatur

[1] Schröder, J., Rohdenburg, T., Hohmann, V. and Ewert, S. D., "Classification of Reverberant Acoustic Situations", Int. Conf. on Acoustics (NAG/DAGA 2009), Rotterdam, The Netherlands, Mar. 2009

[2] Steeneken, H.J.M. and Houtgast, T., "A physical method for measuring speech-transmission quality", J. Acoust. Soc. Am. 67, 318-326, 1980

[3] Schröder, J., "Klassifikation von Nachhall auf der Basis von einkanaligen Signalen", Diploma thesis, University of Oldenburg, 2009

[4] Steeneken, H.J.M. and Houtgast, T., "Basics of the STI-measuring method", Past. Present and Future of the Speech Transmission Index, edited by Sander J. van Wijngaarden,13-43, The Netherlands: TNO Human Factors, 2002

[5] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen and V. Välimäki, "Estimation of Modal Decay Parameters from Noisy Response Measurments", J. Audio. Eng. Soc., Vol. 50, No. 11, pp. 867-878, Nov. 2002

[6] Wagener, K., Kühnel, V. and Kollmeier, B., "Entwicklung und Evaluation eines Satztestes für die deutsche Sprache. I: Design des Oldenburger Satztestes", Zeitschrift für Audiologie 38(1):4-14, 1999

[7] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am., Vol. 65, No. 4, pp. 943-950, Apr. 1979

[8] Habets, E. A. P., "Room Impulse Response Generator", `http://home.tiscali.nl/ehabets/rir_generator.html`