SUBJECTIVE SPEECH QUALITY AND SPEECH INTELLIGIBILITY EVALUATION OF SINGLE-CHANNEL DEREVERBERATION ALGORITHMS

Anna Warzybok[1,5], *Ina Kodrasi*[1,5], *Jan Ole Jungmann*[2], *Emanuël Habets*[3], *Timo Gerkmann*[1,5], *Alfred Mertins*[2], *Simon Doclo*[1,5], *Birger Kollmeier*[1,4,5], *Stefan Goetze*[4,5]

[1] University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[2] University of Lübeck, Institute for Signal Processing, Lübeck, Germany
[3] International Audio Laboratories, Erlangen, Germany
[4] Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg, Germany
[5] Cluster of Excellence Hearing4all

ABSTRACT

In this contribution, six different single-channel dereverberation algorithms are evaluated subjectively in terms of speech intelligibility and speech quality. In order to study the influence of the dereverberation algorithms on speech intelligibility, speech reception thresholds in noise were measured for different reverberation times. The quality ratings were obtained following the ITU-T P.835 recommendations (with slight changes for adaptation to the problem of dereverberation) and included assessment of the attributes: *reverberant*, *colored*, *distorted*, and *overall quality*. Most of the algorithms improved speech intelligibility for short as well as long reverberation times compared to the reverberant condition. The best performance in terms of speech intelligibility and quality was observed for the regularized spectral inverse approach with pre-echo removal. The overall quality of the processed signals was highly correlated with the attribute reverberant or/and distorted. To generalize the present outcomes, further studies are needed to account for the influence of the estimation errors.

Index Terms— dereverberation, speech intelligibility, speech quality, perceptual validation

1. INTRODUCTION

In realistic conditions, speech intelligibility and perceived quality of speech utterances are mainly determined by background noise and reverberation. To decrease the detrimental effect of noise and reverberation on speech intelligibility and/or quality, a number of different noise reduction and dereverberation techniques have been proposed over the last decades. Most of these techniques, however, introduce temporal and spectral changes in the speech and noise components of the output signal, what may affect speech intelligibility and speech quality. The influence of the different types of distortions on speech intelligibility and perceived quality as well as the relationship between these two aspects is not yet entirely understood.

This work focuses on the perceptual evaluation of a selection of single-channel dereverberation algorithms. This encompasses speech intelligibility measurements in noise and quality assessment of processed signals for the evaluation dimensions *reverberant*, *colored*, *distorted* and *overall quality* [1]. To account for different types of distortions, different classes of dereverberation algorithms were included in the evaluation, i.e. (i) least-squares equalization [2], impulse-response reshaping by (ii) weighting of the error used for least-squares minimization [3] or by (iii, iv) aiming at *hiding* the equalized impulse response under the temporal masking threshold [4], as well as spectral suppression methods for direct dereverberation of the reverberant signal in the short-time Fourier domain, one (v) based on a statistical model of the room impulse response [5, 7] and one (vi) incorporating knowledge about the impulse response to be equalized in the spectral suppression scheme [6] (cf. also Section 2). Please note, that all algorithms besides [5, 7] are designed based on knowledge of the room impulse response (RIR) while [5, 7] only needs estimates of the room reverberation time (RT60) and the direct-to-reverberation ratio (DRR) which are much more easy to obtain in practical systems than a reliable estimate of the RIR. While this paper focuses on the subjective quality assessment for dereverberation algorithms, the results of the listening tests analyzed in this contribution are compared to ratings by objective quality measures in [8].

The remainder of this paper is organized as follows: the study design and methodology are introduced in Section 3. Section 4 describes the results which are then summarized in Section 5.

2. ALGORITHMS UNDER TEST

The most simple impulse response equalization technique is known as least-squares equalization [2] which is defined in a generalized form by

$$\mathbf{c}_{\mathrm{EQ}} = (\mathbf{WH})^{+}\mathbf{Wd}. \tag{1}$$

with $\mathbf{H}$ and $\mathbf{d}$ being the channel convolution matrix and the desired system response and $(\cdot)^{+}$ the Moore-Penrose pseudo inverse, respectively. An appropriate window function

$$\mathbf{W} = \mathrm{diag}\left\{\mathbf{w}_{\{\mathrm{I,II}\}}\right\} \tag{2}$$

may be chosen as

$$\mathbf{w}_{\mathrm{I}} = \mathbf{1}_{[N1+N2\times 1]} \tag{3}$$

The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the University of Erlangen-Nürnberg and Fraunhofer IIS.

This work was partially supported by the project Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS, project no. 316969) funded by the European Commission (EC), as well as by the DFG-Cluster of Excellence EXC 1077/1 "Hearing4all".

to result in the conventional least-squares equalizer [2] or to [3]

$$\mathbf{w}_{\mathrm{II}} = [\underbrace{1, 1, ..., 1}_{N_1}, \underbrace{w_{\mathrm{II},0}, w_{\mathrm{II},1}, ..., w_{\mathrm{II},N_2-1}}_{N_2}]^T, \quad (4)$$

$$w_{\mathrm{II},i} = 10^{\frac{3\alpha}{\log_{10}(N_0/N_1)}\log_{10}(i/N_1)+0.5}, \quad (5)$$

to result in the so-called weighted least-squares equalizer that emphasizes the suppression of late parts of the equalized impulse response to prevent perceptually disturbing late echoes [1, 9]. In (4) and (5), the constants N_0, N_1 and N_2 are defined as follows: $N_0 = (t_0 + 0.2)f_s$, $N_1 = (t_0 + 0.004)f_s$ and $N_2 = L_h + L_{\mathrm{EQ}} - 1 - N_1$ with t_0, f_s, L_h and L_{EQ} being the time of the direct path of the impulse response, the sampling rate, and the lengths of the RIR and of the equalization filter, respectively. The factor α influences the steepness of the window. For $\alpha = 1$, the window corresponds to the masking found in human listeners [10]. It is known that impulse response shaping (e.g. by WLS equalization) is more robust regarding RIR estimation errors and spatial mismatch [9] than the conventional LS approach. Therefore, the third algorithm under test is the p-norm-based RIR shaping approach as described in [4], implemented here in two variants, i.e. (i) using the window function defined in (5) with $\alpha = 1$ (denoted here as p-norm standard) and (ii) using the same approach with a windows function limited to -60 dB (denoted here as p-norm adapted) [8]. The latter is motivated by the fact that it can be assumed that reverberation can not be perceived more than 60 dB below the main peak of the RIR. The algorithms described so far aim at reshaping of the room impulse response. They can be applied either in front of the loudspeaker for pre-equalization or as post-equalization in the microphone channel. Furthermore, a spectral reverberation suppression rule according to [5, 7] is assessed that aims at dereverberation of the reverberant microphone signal. In particular, the clean speech was estimated using the log-spectral amplitude estimator as described in [11] and the late reverberant spectral variance estimator was estimated using [7] assuming that the frequency-independent reverberation time and direct-to-reverberation ratio were known. The last dereverberation method under test calculates the regularized spectral inverse and then performs a post-processing to remove pre-echoes [6]. Table 1 summarizes the algorithms under test.

Table 1. Different dereverberation approaches and the respective acronyms.

Acronym	Method
LS-EQ	Least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}$ according to (1) without weighting of error signal ($\mathbf{w}_{\mathrm{I}} = \mathbf{1}$)
WLS-EQ	Least-squares equalizer $\mathbf{c}_{\mathrm{EQ}}$ according to (1) with window function according to (5) and $\alpha = 1$
Pnorm$_s$	Standard p-norm RIR shaping according to [4] using the window function according to (5) and $\alpha = 1$
Pnorm$_a$	Adapted p-norm RIR shaping according to [4] using the window function according to (5) with $\alpha = 1$, limited to a minimum of -60 dB [8]
Spec Sup	Spectral reverberation suppression according to [5, 7]
F-Inv	Regularized spectral inverse with pre-echo removal according to [6]

3. PERCEPTUAL EVALUATION

The perceptual evaluation of the dereverberation algorithms included (i) speech intelligibility measurements in noise and (ii) subjective quality listening tests conducted according to the ITU-T P.835 recommendations [12] (with slight modifications, cf. [1]). The dereverberation algorithms were compared for 5 RIRs characterized by RT60s of 0.7 s, 1 s, 1.1 s, 1.6 s, and 3.8 s. To simulate the different RT60 conditions, the clean speech and noise signals were convolved with the respective RIRs. Four RIRs (0.7 s, 1.1 s, 1.6 s, 3.8 s) were generated by means of the image method [13] for a room size of 6 x 4 x 2.6 m^3. The RIR with RT60 of 1 s was measured in a real room having a size of 3.9 x 3.1 x 2.3 m^3. The source-receiver distance was fixed at 0.54 m for all RIRs. The reverberant speech signals (sampled at $f_s = 16$ kHz) were processed by the dereverberation algorithms described in Section 2. The filter lengths for LS and WLS equalizers were $L_{\mathrm{EQ}} = 8192$ and for the the p-norm approaches $L_{\mathrm{EQ}} = 16384$, respectively. Please note, that the algorithm performance not necessarily increases with the filter length [1]. The spectral suppression algorithm processed the reverberant speech signals in short-term spectral domain based on estimates of the RT60 and the DRR [5]. The regularized inverse filter F-Inv was computed using a discrete fourier transform (DFT) length of K = 262144 and a regularization parameter δ = 10 $^{-4}$ [6]. The re-synthesized signal was then processed by the speech enhancement scheme, where the spectral analysis is done using the DFT length K'=512 and an overlap of 50 %. As a reference, the reverberated unprocessed signals were also tested. The root mean square (RMS) values of the processed signals were set to the RMS of the original (clean) signals to enable the comparisons across the different algorithms.

3.1. Speech intelligibility measurements

9 normal-hearing listeners participated in the measurements. Speech intelligibility was measured adaptively in noise using speech material from the Oldenburg sentence test [14]. The signals were presented diotically over free-field equalized headphones (Sennheiser HDA200). The level of the speech-shaped noise was kept constant at 65 dB SPL. The speech level was varied and converged to the 50 % speech intelligibility (so-called speech reception threshold, SRT). Prior to the measurement, listeners were trained to account for the training effect and to familiarize themselves with the task. Two training lists were presented to each listener; the first list was presented at a fixed signal-to-noise ratio (SNR) of -2 dB. The second training list was presented adaptively. The training lists were disregarded from the further analysis. The order of listening conditions (RT60s and algorithms) was randomized across listeners.

To directly compare different algorithms, all results are shown as speech-weighted SNR which is a measure of an effective SNR taking into account the relative contributions of different regions of the frequency spectrum to speech intelligibility (cf. also Table 3 within the Speech Intelligibility Index standard [15]).

3.2. Subjective quality assessment

The quality assessment was conducted with 21 normal-hearing listeners, including all listeners participating in the speech intelligibility measurements. The listeners' task was to assess the speech quality regarding four attributes: *reverberant*, *colored*, *distorted*, and *overall quality*. The 5-point mean opinion score (MOS) scale was used as opinion rating method [12, 1]. Each category was assigned a numerical value between 1 (corresponding to bad overall quality or very reverberant, distorted or colored signals) and 5 (corresponding to excellent overall quality and not reverberant, colored or distorted signals). Quality assessment was possible in steps of 0.1. The speech samples, consisting of two sentences (a subset of the speech mate-

rial used in the speech intelligibility measurements), had a length of about 5 s and were scaled to have the same level. Prior to the actual measurements, listeners were trained to familiarize themselves with the task and the signals under test. Similarly to the speech intelligibility measurements, the order of listening conditions (RT60s and algorithms) was randomized across listeners.

4. RESULTS

4.1. Speech reception thresholds

Mean SRTs (averaged across listeners) and corresponding standard deviations for different dereverberation approaches are presented as a function of RT60 in Fig. 1.

The data were statistically analyzed by means of two-way repeated measures analysis of variance (ANOVA) with factors 'algorithm' and 'reverberation time'. The statistical analysis revealed the main effect of the factors 'algorithm' ($F(6,42.63) = 348.63$, $p < 0.001$), 'reverberation time' ($F(4,23.08) = 92.0$, $p < 0.001$) as well as the interaction between them ($F(24,79.67) = 12.45$, $p < 0.001$). To determine the sources of significance, the post hoc tests (with Bonferonni corrections) were conducted for each reverberation time separately. Generally, reverberation decreased speech intelligibility with increasing RT60 from -7 dB (RT60 = 0.7 s) to -2.8 dB (RT60 = 3.8 s). When comparing the SRTs for the measured and simulated RIR with similar RT60 of 1 and 1.1 s, respectively, significantly lower SRTs can be observed for the measured RIR. This can be related to the fact that the early (useful) to total energy ratio (so-called definition) was greater for the measured than for the simulated RIR.

PNorm$_a$, Spec Sup, and F-Inv algorithms improved speech intelligibility at each RT60 compared to the reverberant condition. The lowest (i.e. the best) SRTs were obtained by using the F-Inv algorithm, which showed significantly better speech intelligibility than all other algorithms at all RT60s. No algorithm decreased speech intelligibility compared to the reverberant case. PNorm$_a$, Spec Sup, and LS algorithms showed similar performance (with the exception of RT60 = 1.1 s at which statistically relevant differences can be found), which suggests that different classes of algorithms can result in quantitatively comparable improvement in speech intelligibility compared to the reverberant condition, however, of course with differences regarding robustness. The PNorm$_a$ approach did not result in better speech intelligibility than the PNorm$_s$ approach, however, in contrast to PNorm$_s$, PNorm$_a$ improved speech intelligibility compared to the reverberant conditions.

4.2. Subjective quality assessment

Results of the subjective quality assessment are shown by means of box-plots in Fig. 2. For each of the four attributes, the results are ordered in descending order of median value. Different colors depict different algorithms (magenta: reverberant signals, grey: LS, orange: WLS, blue: PNorm$_s$, black: PNorm$_a$, green: Spec Sup, and red: F-Inv). The digits from 1 to 5 (in the x-axis labels) indicate the different RT60s ranging from 0.7 s to 3.8 s, respectively. To determine which speech signal properties (*reverberation*, *distortions*, *coloration*) have an influence on the *overall quality*, the inter-attribute correlations $|r|$ of median MOS values were calculated and are summarized in Table 2.

As expected, the *overall quality* for reverberated, unprocessed signals was mainly determined by the *reverberation* as shown by the high correlation between these two attributes ($r = 0.942$*). The median of MOS for *overall quality* and reverberated signals ranged

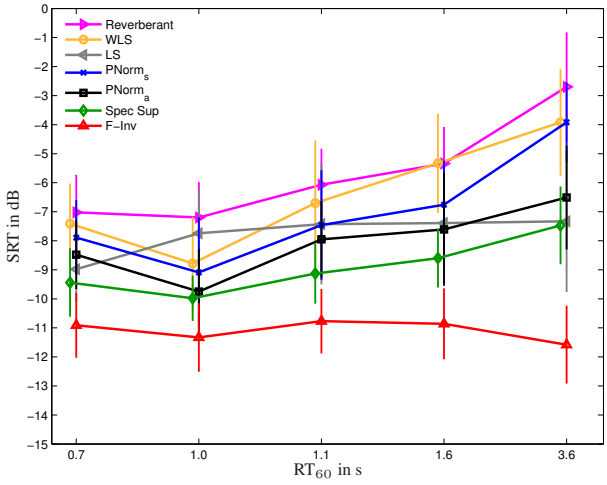

Fig. 1. Speech reception threshold as a function of reverberation time for ▷ reverberant signals and signals processed by ○ WLS, ◁ LS, × PNorm$_s$, □ PNorm$_a$, ◇ Spec Sup, △ F-Inv.

from 2 (RT60 = 3.8 s) to 3.2 for the shortest RT60. For the LS approach the median MOS scores for *overall quality* ranged from 2 to 2.4 which corresponds to poor *overall quality*. The WLS approach was assessed with higher median scores for *overall quality* than the LS approach but only for short RT60s. The median MOS for the WLS approach and attributes *reverberant* and *distorted* was on average 1.3 and 1.6 higher than for the LS approach. This indicates that better *overall quality* for the WLS approach than the LS approach at short RT60s was related to less *distortion* as well as less *reverberation*.

Both PNorm algorithms were qualitatively similarly assessed regarding *overall quality* with median MOS scores from 2.1 (RT60 =3.8 s) to 3.7 (RT60 = 1.1 s) for the PNorm$_s$ approach and from 2.4 (RT60 =3.8 s) to 3.7 (RT60 = 1.0 s) for the PNorm$_a$ approach. For the PNorm$_s$ approach, *overall quality* seems to be mainly determined by the amount of *reverberation* ($r = 0.958$*) and for the PNorm$_a$ approach by *distortion* ($r = 0.987$*). In terms of *overall quality*, PNorm algorithms were scored higher (i.e. better) than LS, WLS, and Spec Sup algorithms.

Similar to the LS and the WLS algorithms, a relatively low *overall quality* was observed for the Spec Sup algorithm with the median scores ranging from 1.5 (RT60 = 3.8 s) to 2.4 (for RT60 = 0.7 s and 1.0 s). A strong correlation between attributes *overall quality* and *reverberant* ($r = 0.923$*) as well as *distorted* ($r = 0.976$*), and between *reverberant* and *distorted* ($r = 0.98$*) was found for the Spec Sup approach. Very low median scores for the attribute *distorted*, ranging from 1.3 (RT60 = 3.8 s) to 2.2 (RT60 = 1.1 s), indicate that the poor *overall quality* was mainly determined by high amount of *distortion*. For all four attributes, the highest rating scores (median in range from 3.5 to 5) were observed for the F-Inv algorithm indicating that this algorithm provides the highest signal quality.

5. DISCUSSION AND CONCLUSION

In this paper, single-channel dereverberation algorithms were subjectively evaluated in terms of speech intelligibility and peech qual-

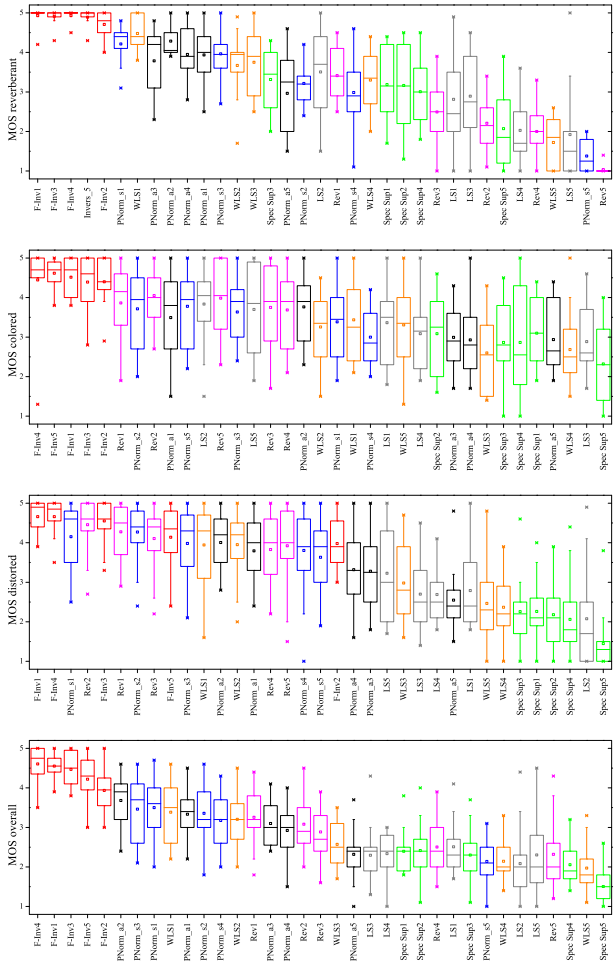

Fig. 2. Subjective rating of speech samples for attributes: reverberant, colored, distorted and overall. Different colors depict different algorithms; magenta: reverberant signals, grey: LS, orange: WLS, blue: PNorm$_s$, black: PNorm$_a$, green: Spec Sup, and red: F-Inv. The numbers 1 to 5 in the x-axes labels denote the RT60s ranging from 0.7 s to 3.8 s, respectively.

ity. The F-Inv algorithm which incorporates knowledge about the impulse response to be equalized to spectral inversion showed improved speech intelligibility and resulted in a very good or even excellent speech quality. The LS and Spec Sup algorithms significantly improved speech intelligibility but introduced noticeable distortions and due to this led to lower speech quality even for short RT60s. For the LS approach, an insufficient overall quality seems to be related to two different aspects: for short RT60s bad overall quality is determined by distortions (e.g. late- and ringing-echoes [1]), however, with increasing RT60 the influence of reverberation which is present in speech signals increases and probably masks the distortions perceived as detrimental at short RT60s. This is supported by correlation analysis which has shown a strong, negative correlation between the attributes reverberant and distorted ($r = -0.951$*). For

Table 2. Inter-attribute correlations $|r|$ of MOS values of subjective ratings. Stars indicate statistically significant correlations (* for $p < 0.05$ and ** for $p < 0.01$).

Method	Attribute	Colored	Distorted	Overall
all algos	Reverberant	0,339*	0.409*	0.84**
	Colored	-	0.767*	0.684**
	Distorted	-	-	0.775**
Rev	Reverberant	0.459	0.717	0.942*
	Colored	-	0.61	0.648
	Distorted	-	-	0.881*
LS	Reverberant	0.03	-0.97**	-0.052
	Colored	-	-0.205	-0.881*
	Distorted	-	-	0.152
WLS	Reverberant	0.282	0.773	0.884*
	Colored	-	-0.795	0.675
	Distorted	-	-	0.978**
PNorm$_s$	Reverberant	-0.418	0.805	0.969*
	Colored	-	-0.031	-0.372
	Distorted	-	-	0.688
PNorm$_a$	Reverberant	0.466	0.69	0.774
	Colored	-	0.939*	0.895*
	Distorted	-	-	0.987**
Spec Sup	Reverberant	0.828	0.942*	0.837
	Colored	-	0.809	0.772
	Distorted	-	-	0.968**
F-Inv	Reverberant	0.943*	0.938*	0.772
	Colored	-	0.933*	0.765
	Distorted	-	-	0.933*

the Spec Sup algorithm an overall quality was mainly determined by distortions which were detrimental even for short RT60s. This indicates that time variant distortions of the speech part affect speech quality. However, they are not necessarily detrimental to speech intelligibility. Thus, focus for development of future spectral suppression algorithm has to be on a processed speech signal with minimum distortions, if speech quality should be the main focus. The weighting window applied in the WLS algorithm improved overall quality for short RT60s compared to the LS algorithm. This improvement seems to be related to the reduction of the pre- and late echoes what is expressed by higher MOS scores for the attribute *distorted* for the WLS than for the LS algorithm. However, applying the weighting window did not improve speech intelligibility as well as speech quality for longer RT60s. PNorm$_a$ showed similar results as LS and Spec Sup algorithms in terms of speech intelligibility but additionally improved speech quality.

It should be stressed that all algorithms, except for [5, 7], were designed based on perfect knowledge of the RIR. The Spec Sup algorithm requires an estimate of the RT60 and the DRR which were also known in this study. In realistic conditions, the RIR, the RT60, and the DRR have to be estimated. It is generally known, that estimation of the RT60 and the DRR is easier than estimation of the full RIR. Furthermore, the errors in the RT60 and the DRR estimation have less influence on the algorithm performance than estimation errors that occur while estimating the full RIR [5]. To generalize the present outcomes for all algorithms, further studies have to be done to account for the influence of the estimation errors on the speech intelligibility and quality.

6. REFERENCES

[1] S. Goetze, E. Albertin, J. Rennies, E.A.P. Habets, and K.-D. Kammeyer, "Speech Quality Assessment for Listening-Room

Compensation," in *38th AES Conference*, Pitea, Sweden, July 2010, pp. 11–20.

[2] S. T. Neely and J. B. Allen, "Invertibility of a Room Impulse Response," *Journal of the Acoustical Society of America (JASA)*, vol. 66, pp. 165–169, July 1979.

[3] M. Kallinger and A. Mertins, "Room Impulse Response Shaping – A Study," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. V101–V104.

[4] A. Mertins, T. Mei, and M. Kallinger, "Room Impulse Response Shortening/Reshaping with Infinity- and p-Norm Optimization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, pp. 249–259, Feb. 2010, DOI:10.1109/TASL.2009.2025789.

[5] E.A.P. Habets, *Single and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, University of Eindhoven, Eindhoven, The Netherlands, June 2007.

[6] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-Domain Single-Channel Inverse Filtering for Speech Dereverberation: Therory and Practice," in *Proc. 2014 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[7] 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, Sep. 2009.

[8] S. Goetze, A. Warzybok, Kodrasi I, J. O. Jungmann, B. Cauchi, J. Rennies, E.A.P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A Study on Speech Quality and Speech Intelligibility Measures for Quality Assessment of Single-Channel Dereverberation Algorithms," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes, France, Sep. 2014.

[9] S. Goetze, *On the Combination of Systems for Listening-Room Compensation and Acoustic Echo Cancellation in Hands-Free Telecommunication Systems*, Ph.D. thesis, Dept. of Telecommunications, University of Bremen (FB-1), Bremen, Germany, 2013.

[10] L. D. Fielder, "Practical Limits for Room Equalization," in *Proc. AES Convention (Audio Engineering Society)*, New York, NY, USA, Sept. 2001, vol. 111, pp. 1 – 20.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[12] ITU-T P.835, "Subjective Test Methology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm, ITU-T Recommendation P.835," Nov. 2003.

[13] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[14] K. Wagener, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation einesSatztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests (In German language)," *Zeitschrift für Audiologie / Audiological Acoustics*, vol. 38, pp. 86–95, 1999.

[15] ANSI 1997, "Methods for Calculation of the Speech Intelligibility Index," 1997.