A robust speech preprocessing algorithm based on overlap-masking reduction

Julian Grosse, Steven van de Par

CvO Universität Oldenburg, AG Akustik, Cluster of Excellence "Hearing4all", email: julian.grosse@uni-oldenburg.de

Introduction

Since the invention of loudspeaker reproduction systems, the playback of speech in reverberant environments is often encountered in our daily life, e.g. during telephone conferences or announcements in public address systems. Since the clean speech signal is available in such a scenario, it allows to reduce the detrimental effects of reverberation on the speech signal before the signal is reproduced within a room. There are several implications for speech intelligibility when reverberation is present in a room, e.g., the time-smearing effect caused by late reflections that will potentially mask speech events [1] and a reduction of the modulation depth which is reported to be important for speech perception [2]. Instead of aiming to *post-process* enhancing the speech signal that is captured at the listener position like in a hearing aid scenario [3], the access to the clean speech allows to *pre-process* the speech to reduce the overlap-masking on consecutive speech events such as initially proposed by [4], [5], [6]. [5] showed that a reduction of overlap-masking increases the reduced modulation depth which results in an increased speech intelligibility [7]. Additionally, [8] has shown, that an artificially increased consonant-to-vowel intensity ratio can increase speech intelligibility when reverberation is present. Hence, a second approach is presented, that increases the consonant-to-vowel intensity ratio (onset-enhancement).

This study presents an extended pre-processing approach based on preliminary work by [4], that increases intelligibility based on human auditory processing to reduce the overlap-masking due to the reverberation that is present. An auditory model [9] is used to optimize the proposed algorithms to find the optimal parameter-settings that lead to the lowest predicted speech-receptions thresholds at 50% speech intelligibility (SRT$_{50}$). An objective and subjective evaluation shows the impact on the intrinsic modulation spectra and on speech intelligibility for a wide range of reverberation times even when a considerably mismatch of the impulse response (IR) is assumed.

Method

This section describes the basic idea to increase speech intelligibility in reverberant environment that is usually degraded when it is emmited into a reverberant environment. A more extensive derivation of the gain-functions that can be applied directly on the dry speech signal can be found in [10]. Because the algorithm makes use of knowledge about auditory processing, it does not rely on details of the exact fine-structure of the speech signal $s(t)$ but rather on the energy distribution $\phi_s^k(j)$ in auditory filters j in 32 ms time frames k.

If a speech signal $s(t)$ is rendered into a room it can be described as a convolution with a impulse response $h(t)$, formulated as the spectral energy distribution $\phi_{\mathrm{y}}^k(j)$. Because the direct sound has the most important contribution to speech intelligibility, we can assume that the best intelligibility can be achieved if only the direct component ($\phi_{s_d}^k(j)$) is present in a room. This will preserve the similarity of spectro-temporal shape without the disturbing masking component ($\phi_{s_m}^k(j)$):

$$\phi_{\mathrm{y}}^k(j) = \phi_{\mathrm{s_d}}^k(j) \tag{1}$$

The direct and masking contribution of the reverberated speech $\phi_y^k(j)$ is defined as follows:

$$\phi_y^k(j) = \phi_{s_d}^k(j) + \phi_{s_m}^k(j) \tag{2}$$

A closer look on Eq. 2 reveals, that the direct component $\phi_{s_d}^k(j)$ has only got contributions of a single frame k, whereas the masking component $\phi_{s_m}^k(j)$ has multiple contributions from previous frames ($k-1, k-2, ..., k-N$) contributing to frame k, because of the reverberant tail of the IR. The number of total frames N is defined by the total length of the IR.

Approach 1: Overlap-masking reduction (OMR)

The first approach aims to reduce the amount of overlap-masking to increase speech-intelligibility. After substituting Eq. 1 and Eq. 2 we receive a time and frequency dependent gain function which can be applied directly to the dry speech signal:

$$\alpha^k(j) = \frac{\phi_{s_d}^k(j) - \phi_{s_m}^k(j)}{\phi_{s_d}^k(j)} \tag{3}$$

The analysis of Eq. 3 shows, that $\alpha^k(j)$ will be 1 if no masking component ($\phi_{s_m}^k(j) = 0$) is present (or no reverberation). If the amount of reverberation increases ($\phi_{s_m}^k(j) > \phi_{s_d}^k(j)$), the direct contribution is masked by the reverberant tail and the respective spectro-temporal unit j is reduced ($\alpha^k = 0$) because it is assumed to be inaudible.

Approach 2: Onset-enhancement (OE)

A second approach can be derived by using the direct to reverberant ratio of speech (DRRs) as a measure to control the gain-function:

$$\mathrm{DRRs}^k(j) = \alpha^k(j) = \frac{\phi_{s_d}^k(j)}{\phi_{s_m}^k(j)} \tag{4}$$

This function acts like an onset-enhancement, emphasizing the segments which rapidly change across time signaled by a masking component that is rather small with respect to the direct component (like for example consonants). The segments in which $\phi^k_{s_m}(j)$ is smaller than $\phi^k_{s_d}(j)$ are reduced like in the OMR approach.

Evaluation

For evaluation purposes, the speech material of the Oldenburg Sentence Test (OLSA, [11]) was used, both in the objective and subjective evaluation. Speech reception thresholds (SRT_{50}) at 50% speech intelligibility level was measured in the presence of speech shaped noise (SSN) that was presented at a level of 65 dB-SPL to the listener. The speech material and noise was convolved with the IR $h(t)$ to simulate the reproduction in a room and was either presented with processing (overlap-masking-reduction (OMR) or onset-enhancement (OE)) or without processing (unproc) to the listener. Eight listener participated in SRT measurements for the three different room-acoustical scenarios shown in table 1. Additionally, the algorithm was evaluated in two rooms that are similar to R1 and R3 to investigate the robustness of the proposed algorithms. For that, the algorithm was optimized on one (optimized)-position (IR) and was evaluated on a second (non-optimized)-position (IR) in the same room. Such a comparison allows to show if it is advantageous to consider only a coarse spectro-temporal representation of speech instead of using the exact fine-structure and phase of the IR.

Tabelle 1: Room-acoustical scenarios used in the objective and subjective evaluation.

properties	room-acoustical scenarios		
	room 1 (R1)	room 2 (R2)	room 3 (R3)
$T_{60}(s)$	0.8	1.2	3
length (m)	10.9	5	32
width (m)	10.8	3	11
height (m)	3.15	3	9.5

Objective Evaluation

Fig. 1 shows an example processing of a sentence, filtered with the OMR-approach. Illustrated are spectro-temporal energy distributions $\phi^k(j)$ of a) the dry speech signal $\phi^k_s(j)$, b) the reverberated speech signal $\phi^k_y(j)$, c) the direct component $\phi^k_{s_d}(j)$, d) the masking component $\phi^k_{s_m}(j)$ and e) the pre-processed speech signal $\phi^k_{\hat{s}}(j)$ after filtering with $\alpha^k(j)$. A closer look on panel a) shows well-defined characteristic of amplitude-modulation of speech which is strongly affected by reverberation b) when it is emitted into the room. We can assume that the best intelligibility can be achieved and both signals will have a high similarity when only the direct sound would be present, i.e. without the degrading masking component illustrated in d) which reduces modulation due to the time-smearing effect and overlap masking. The lowest panel shows the pre-processed speech signal. It can be seen that after a certain amount of reverberation, quasi stationary speech segments are reduced in amplitude to reduce the amount of energy emitted into the room to preserve the similarity of the envelopes in terms of spectro-temporal distributions and therefore to reduce the amount of overlap-masking.

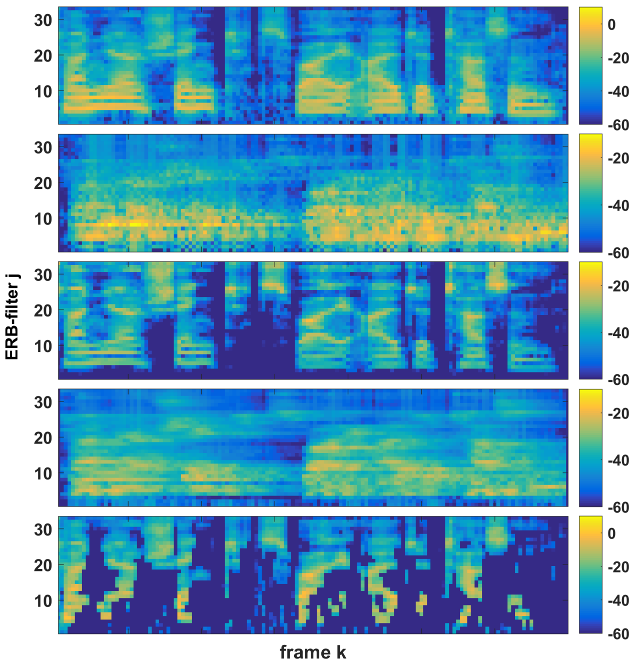

Abbildung 1: Example of the effect of processing of a speech sentence with the overlap-masking reduction approach. Illustrated in all panels is the energy within auditory bands j across frame number k. Panel a) shows the energy of a dry unprocessed speech signal $s(t)$, b) shows the reverberated signal caused by the convolution of a IR $h(t)$ with $s(t)$, c) shows the convolution of the direct contribution of the IR $h_d(t)$ with $s(t)$, d) shows the convolution of the reverberant (masking) contribution of the IR $h_r(t)$ with $s(t)$ and e) shows a pre-processed dry speech sentence calculated from the direct contributions from panel c) and the reverberant contributions from panel d).

Modulation improvement

Fig. 2 shows the signal-to-noise envelope energy ($\mathrm{SNR}_{\mathrm{env}}$) across auditory center frequency and center modulation frequency seen by an auditory model for speech intelligibility prediction [9]. The model is able to predict speech intelligibility in the presence of an interferer for non-linear processing applied on reverberated speech. Illustrated is the mean envelope energy at a global SNR of -3 dB for the unprocessed reverberated speech (upper panel) and for the pre-processed reverberated speech with the OE-algorithm (lower panel). The comparison between the unprocessed speech and pre-processed speech shows an overall improvement of the modulation energy due to the pre-processing. This effect is observable for low modulation frequencies which are important for speech perception.

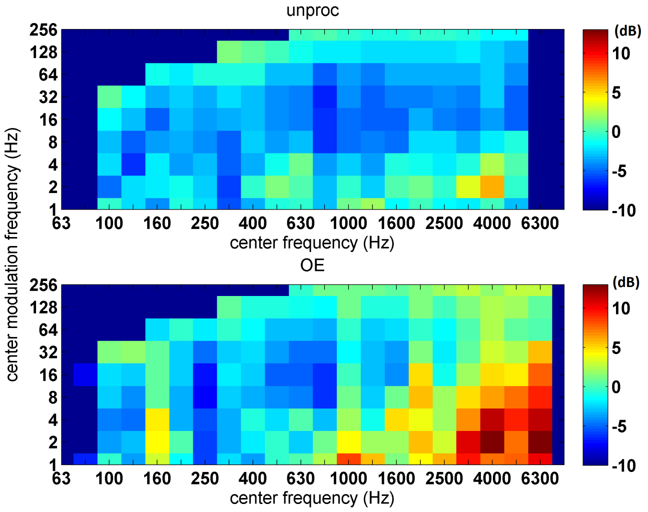

Abbildung 2: Signal-to-noise envelope energy $\mathrm{SNR}_{\mathrm{env}}$ across center frequency and center modulation frequency. The upper panel shows the $\mathrm{SNR}_{\mathrm{env}}$ of the unprocessed and the lower panel the pre-processed energy.

Subjective Evaluation

Figure 3 shows the SRT_{50} for the subjective evaluation obtained with the OLSA test for R1, R2 and R3. Illustrated are mean values with between-subject variability in terms of standard deviation. Each subject made two repetitions. Considering the left-panel in figure 3, it can be seen that both algorithms (OE and OMR) improve speech intelligibility by about 1 to 2 dB. At a reverberation time of $\mathrm{T}_{60} = 1.2$ s, the mid-panel shows an improvement of about 6 dB for the OMR condition and a slightly lower improvement of about 5 dB for the OE conditions. The right-most panel shows for the OMR only a slightly improved SRT and for the OE condition an improvement of about 2 dB is seen.

Robustness

To investigate the robustness of the algorithms in terms of a change in the listener position, the subjective evaluation was conducted at the position on which the algorithm was optimized and on a non-optimized position (distance between microphones approximately 1 m and 5 m). The results in figure 4 show that the improvements of 1 dB to 2 dB in speech reception thresholds across both rooms are comparable to one another which indicates a rather robust behavior of the algorithms against an IR mismatch.

Summary

This study presented two pre-processing approaches on basis of preliminary work by [4] that reduces the amount of overlap-masking of speech that is emitted in a reverberant environment. On basis of auditory motivated transfer functions derived separately for the direct and the reverberant/masking components, the algorithms use both contributions to control gain functions and to decide if a speech segment will be inaudible and does not contribute to speech intelligibility anymore and can be reduced. Because the first processing approach acts like an onset enhancement, a second approach was derived in which onsets are directly emphasized and steady-state components are reduced in amplitude. An optimization and analysis with an auditory model showed improvements of the degraded modulation spectra that are important for speech intelligibility. A speech intelligibility test indicated that such an algorithm allows to improve speech intellgibility between 1 dB and 6 dB depending on the pre-processing strategy and the amount of reverberation (T_{60}) that is present in the room. The listening test also indicated that the proposed algorithms are robust when the IR had a considerably mismatch. The rather low complexity and the causal framework of the algorithm allows for a real-time application to speech that is rendered via loudspeakers in reverberant environments.

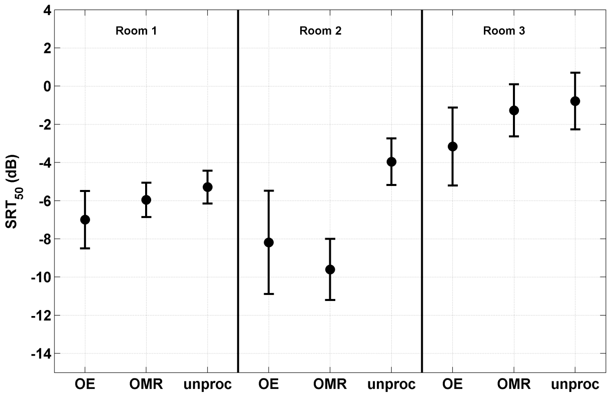

Abbildung 3: Illustrated are mean values and standard deviation of speech reception thresholds at 50% speech intelligibility (SRT_{50}) measured in speech-shaped noise. *Unproc* represents the unprocessed speech, onset-enhancement (OE) and overlap-masking reduction (OMR) represent both pre-processing approaches evaluated in the room-acoustical scenarios shown in table 1.

Literatur

[1] Bolt, RH and MacDonald, AD, “Theory of speech masking by reverberation,” *The Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580 (1949).

[2] Houtgast, T. and Steeneken, H. JM, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077 (1985).

[3] Cauchi, B. and Kodrasi, I. and Rehr, R. and Gerlach, S. and Jukić, A. and Gerkmann, T. and Doclo, S. and Goetze, S., “Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech,” *EURASIP*

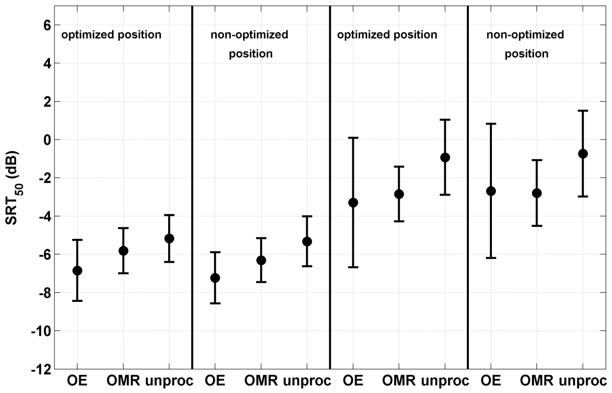

Abbildung 4: Illustrated are boxplots of speech reception thresholds at 50% speech intelligibility (SRT$_{50}$) measured in speech-shaped noise. *Unproc* represents the unprocessed speech, onset-enhancement (OE) and overlap-masking reduction (OMR) the both pre-processing approaches measured in the room-acoustical scenarios shown in table 1. The listening test was conducted at the optimized position and the non-optimized position when as mismatch between the IRs are assumed.

Journal on Advances in Signal Processing, vol. 2015 no. 1, pp. 1–12 (2015).

[4] Arai, T. and Kinoshita, K. and Hodoshima, N. and Kusumoto, A. and Kitamura, T., “Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoustical Science and Technology*, vol. 23, no. 4, pp. 229–232 (2002).

[5] Hodoshima, N. and Arai, T. and Kusumoto, A. and Kinoshita, K., “Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments,” *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4055–4064 (2006).

[6] Arai, T. and Hodoshima, N. and Yasu, K., “Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18 no. 7, pp. 1775–1780 (2010).

[7] Arai, T. and Murakami, Y. and Hayashi, N. and Hodoshima, N. and Kurisu, K., “Inverse correlation of intelligibility of speech in reverberation with the amount of overlap-masking,” *Acoust. Sci. & Tech.*, vol. 28 no. 6, pp. 438–441 (2007).

[8] Gordon-Salant, S., “Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing,” *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1599–1607 (1986).

[9] Jørgensen, S. and Ewert, S. D. and Dau, T., “A multiresolution envelope-power based model for speech intelligibility,” *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446 (2013).

[10] Grosse, J. and van de Par, S., “A Speech Preprocessing Method Based on Overlap-Masking Reduction to Increase Intelligibility in Reverberant Environments,” *J. Audio Eng. Soc.*, vol. 65 no. 1/2, pp. 31–41 (2017).

[11] Wagener, K. and Brand, T. and Kollmeier, B., “Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test,” *Zeitschrift Fur Audiologie*, vol. 38, pp. 86–95 (1999).