A Speech Preprocessing Method Based on Perceptually Optimized Envelope Processing to Increase Intelligibility in Reverberant Environments

Ali Fallah 1, * , Steven van de Par 2

Abstract: Speech intelligibility in public places can be degraded by the environmental noise and reverberation. In this study, a new near-end listening enhancement (NELE) approach is proposed that using a time varying filter jointly enhances the onsets and reduces the overlap masking. For optimization, a some look ahead in the clean speech and prior knowledge of room impulse response (RIR) are required. In this method, by optimizing a defined cost function, the Spectro-Temporal Envelope of a reverb speech is optimized to be as close as possible to that of the clean speech. In this cost function, onsets of speech are optimized with increased weight. This approach is different from overlap masking ratio (OMR) and speech enhancement (OE) approaches (Grosse, van de Par, 2017, J. Audio Eng. Soc., Vol. 65(1/2), pp. 31-41) that only considers previous frames in each time slot for determining the time variant filtering. The SRT measurements shows that the new optimization framework in average enhances the speech intelligibility up to 2 dB more that OE.

Keywords: speech enhancement; NELE; reverberation; speech intelligibility; optimization.

1. Introduction

In conventional speech enhancement methods, the speech signal is recovered from a mixture of reverberation and noise. This type of processing can be used at the receiver side, for example in hearing-aids. For degradation of speech intelligibility in public places such as airports and train stations due to reverberation and noise, because of using one or multiple independent loudspeakers and lack of further processing in the listener side, speech modification is only possible at the source side and only on the clean speech before playback. In the literature, this type of clean speech modification is called near-end listening enhancement (NELE) [1] and is typically evaluated under an equal-level constraint. The modified signal must be more intelligible in the presence of reverberation and noise and also must be robust to different listener positions in a wide area.

NELE algorithms can be divided into three categories: rule-based, noise-dependent and reverberation-dependent. In the rule-based approaches, knowledge about psychoacoustics and speech perception is used to produce more intelligible speech preferably with low audible processing artifacts. However, this methods does not optimize a specific criteria and for this reason, the modification would only be sub-optimal in terms of speech intelligibility [2]. Generally, in NELE algorithms, the preprocessing of clean speech is performed on a time-frequency signal representation. A very well-known rule-based
approach is the Spectral-Shaping Dynamic Range Compression (SSDRC) method [2,3] for which the acoustic cues that are perceptually important are enhanced. In the time domain, a non-linear Dynamic Range Compression (DRC) amplifies lower-energy parts of speech, like consonants, which are known to be more susceptible to reverberation and noise [4]. Also in the frequency domain, by spectral-tilt flattening and formant shifting, the intelligibility is improved. Based on SSDRC, another successful method named Automatic Sound Engineer (ASE) is proposed that using equalization and broadband compression maximizes speech intelligibility while keeping a good sound quality [5].

In the second type of NELE methods, only the presence of noise is taken into account in the development of an enhancement algorithm. Speech modifications in these methods are usually based on an objective speech intelligibility measure, e.g., the speech intelligibility index (SII) [5] or the glimpse proportion (GP) [6] which is used as an optimization target. Similar to the SSDRC, most of the successful noise-dependent algorithms uses DRC in the time domain to enhance the consonants intelligibility. In adaptDRC [6,7], spectral modification and dynamic time compressing are performed to improve the SII in the presence of additive noise. In this method, although the impact of reverberation is not explicitly considered in the enhancement procedure, nevertheless considerable enhancement is reported in the intelligibility in the presence of reverberation [8]. Another noise-dependent NELE approach is based on improving the STOI score [9]. In some of the noise-dependent methods, deep neural networks (DNNs) are used to modify speech energy. In a method called iMetricGAN [10], the enhancement is performed by repeated predictions of the intelligibility score of modified speech and producing scale factors that multiplied on the unmodified spectrogram. The intelligibility-improving signal processing approach (IISPA) [11] is another DNN-based method that uses an automatic-speech recognition-based model of speech perception to optimize different parameters such as band-pass edge frequencies, spectral slope and curvature, and spectral modulation compression or expansion. Note that in these noise-dependent methods, the quality of speech can degrade strongly, specifically in the presence of non-stationary noise. In the third category of NELE methods named reverberation-dependent, room impulse response (RIR) data are explicitly considered in the modification procedure ([12], [13], [14], [15]). Grosse and van de Par (2017) [16] proposed two methods namely OE (Onset Enhancement) and Overlap Masking Ratio (OMR) which were inspired by previous studies in [17,18]. In these two approaches, having access to a RIR, time varying gains are calculated for each frame, based on the energy of current frame and that of the previous frames of speech.

In the current study, a reverberation-dependent approach is developed to optimized the Spectro-Temporal Envelope of a reverberated speech by onset enhancement and also by reducing the amount of overlap-masking. In contrast to the OMR and OE methods proposed in [16], in this approach also future frames are considered in determination of the weight of current frame. Considering the extension of the current frame and its overlapping with the upcoming frames, now an explicit cost function is defined and optimized in a way that the Spectro-Temporal Envelope of the reverberated speech is as close as possible to that of the clean speech signal.

The paper is organized as follows. Section 2 provides the structure of the proposed NELE algorithm. In this section, the cost function definition and the optimization procedure are described. Section 3 presents the results of simulations and measurements. Finally, Section 4 concludes the article with the discussion.
Figure 1. Block diagram of the proposed NELE approach. The pre-processing is including signal windowing, convolving short frames with the RIR to construct extended frames, FFT processing and bin separation to divide a speech into one-third octave bands and onset detection. In the cost function block, the extended frames and data of the onset detection unit are used to construct a cost function for each short frame. The constructed cost functions are iteratively optimized using the data of onset detection unit to compute the gains of short frames. Finally, the enhanced signal is reconstructed using the obtained weights by OLA and summation over one-third octave bands.

2. Proposed NELE Method

The block diagram of the proposed approach is depicted in Figure 1. It consists of preprocessing consisting of a FFT, onset detection, cost function calculation with optimization unit, and finally the signal modification that is including Overlap–Add (OLA) and summation over frequency bands. These parts are explained as follows.

2.1. Preprocessing

As a first step, the time-domain signal is transformed to the frequency domain on a frame-by-frame basis. The main goal of this preprocessing unit is the separation of speech signal into one-third octave bands. This frame-wise frequency analysis is needed for optimization and also for onset detection. First, the clean speech \( x(t) \) is framed using \( \tau = 30 \text{ms} \) Hann windows with 50% overlapping to construct \( N \) overlapping frames \( x_n(t) \):

\[
x(t) = \sum_{n=1}^{N} w(t - n\tau) x(t) = \sum_{n=1}^{N} x_n(t),
\]

in which \( N \) is determined according to the length of signal. Then, each frame is convolved with the RIR:

\[
x(t) * h(t) = \sum_{n=1}^{N} x_n(t) * h(t) = \sum_{n=1}^{N} y_n(t).
\]

Here, \( y_n(t) \) is called “extended frame” number \( n \). An extended frame can mask the upcoming frames dependent on the reverberation time \( T_{60} \). Each extended frame is analyzed using a Fast Fourier Transform (FFT) and subsequently the frequency bins are separated according to the one-third octave bands and finally are synthesized using the inverse FFT (IFFT).
\[
\begin{align*}
\text{FFT} & : y_n(t) \rightarrow Y_n(F) \\
\text{one-third octave band bin separation} & : Y_n,f(F) \rightarrow Y_{n,f}(F) \\
\text{IFFT} & : y_{n,f}(F) \rightarrow y_{n,f}(t).
\end{align*}
\]

(3)

where \( y_{n,f}(t) \) is the synthesized signal in one-third octave band number \( f \). The \( y_{n,f}(t) \) can be considered as the convolution of a short frame \( x_{n,f}(t) \) with the RIR that is filtered by the one-third octave band number \( f \). \( x_{n,f}(t) \) is named a “short frame”. The \( y_{n,f}(t) \) can also be described by smaller frames called “sub-frames”, such that the length of each sub-frame is equal, and temporally aligned to that of a short frame. The length of an extended frame is \( M \) times the length of a short frame:

\[
y_{n,f}(t) = \sum_{m=1}^{M} y_{n,m,f}(t), \quad M = \left[ \frac{l_w + l_h}{l_w} \right],
\]

(4)

where \( l_w \) and \( l_h \) are the length of a short frame and the RIR respectively. A sub-frame \( y_{n,m,f}(t) \) is defined as the frame number \( m \) of an extended frame \( y_{n,f}(t) \):

\[
y_{n,m,f}(t) = \begin{cases} y_{n,f}(t), & (m-1)\frac{\tau}{2} \leq t \leq m\frac{\tau}{2}, \\ 0, & \text{ow} \end{cases}
\]

(5)

The extended frame that is transformed into its corresponding subframes, allows to calculate the total signal power that can be observed within the reverberant environment within a particular time-frame and frequency band. Since in this approach, a time variant filtering is applied on a frame-by-frame basis to the input signal \( x \), the effect of the filtering can be evaluated in weighted summations of \( y_{n,m,f}(t) \) that is affected by all extended frames in a short frame interval. This resulting summation will be considered in a cost function. The weight of short frame \( x_{n,f}(t) \) and subsequently an extended frame \( y_{n,f}(t) \) within an one-third octave band is denoted by \( a_{n,f} \). The short frames and their respected weights, an extended frame and a sub-frame are schematically illustrated in Figure 2.

2.2. Construction of a Cost Function for a frame

According to Figure 2 and Equation (4), the sub-frame number one of \( y_{k,f}(t) \) denoted by \( y_{k,1,f} \) has overlaps with the previous \( M \) extended frames. All of these \( M \) extended frames have sub-frames that are overlapping with \( y_{k,1,f} \).

Figure 2. The schematic of three consecutive and overlapped short frames \( x_{1,f}, x_{2,f}, x_{3,f} \). The short frame and subsequently their extended frames \( y_{1,f}, y_{2,f}, y_{3,f} \) are multiplied by the weights \( a_{1,f}, a_{2,f}, a_{3,f} \), respectively. The extended frame \( y_{3,f} \) and one of its sub-frames \( y_{3,m,f} \) are also shown.
A “summed-reverbed-short frame” number \( k \) that is including all of sub-frames within the short window number \( k \), can be constructed using the summation of current short frame \( y_{k,1,f}(t) \) and the sub-frames of previous extended frames overlapping in frame \( k \):

\[
s_{k,f}(t) = \sum_{m=0}^{M-1} y_{k-m,m+1,f}(t). \tag{6}
\]

For the speech enhancement, a time-varying weight \( \alpha_{n,f} \) according to Figure 2 is multiplied to each short frame and subsequently its convolution with the RIR that constructs the extended frame \( y_{n,f}(t) \). This weight will be consequently multiplied to all \( y_{n,m,f}(t) \) that construct \( y_{n,f}(t) \). By multiplication to these weights, a summed-reverbed-short frame number \( k \) is changed to a “weighted-summed-reverbed-short frame”:

\[
W_{S_{k,f}}(t) = \sum_{m=0}^{M-1} \alpha_{k-m,f} y_{k-m,m+1,f}(t). \tag{7}
\]

In Equation (7), the weights are used to reduce the amount of the overlap-masking in a reverb condition. According to the STOI [19], the temporal envelope of this weighted signal is considered for defining a cost function. A time-frequency unit norm (TF-unit) of a weighted-summed-reverbed-short frame in Equation (7) is calculated by non-coherent summation of the power spectral density values of its discrete Fourier transform (DFT) within a one-third octave band:

\[
W_{S_{k,f}} = \sum_{m=0}^{M-1} \alpha_{k-m,f} \sqrt{\sum_{\text{all bins}} |\text{DFT} \{ y_{k-m,m+1,f}(t) \}|^2}. \tag{8}
\]

The target signal is obtained by processing of the clean speech similar to calculation that is done for the extended frames in Equation (8). But, now only the direct part of RIR \((h_d(t))\) is used for the computation of the target signal:

\[
z_n(t) = x_n(t) * h_d(t). \tag{9}
\]

The TF-unit of this signal in a one-third octave band for frame \( k \) is calculated similar to \( W_{S_{k,f}} \):

\[
E_{k,f} = \sqrt{\sum_{\text{all bins}} |\text{DFT} \{ z_{k,f}(t) \}|^2}. \tag{10}
\]

The cost function for frame \( k \) is now defined as the squared error between the TF-unit of target \((E_{k,f})\) and the TF-unit of weighted frames \((W_{S_{k,f}})\):

\[
CF_{k}(f) = \left( W_{S_{k,f}} - E_{k,f} \right)^2 \tag{11}
\]

This definition of cost function is comparable with the criteria used in the STOI that uses the correlation between temporal envelopes of the clean and degraded speech as an intelligibility score. The temporal envelope in STOI is a vector of TF-units covering a 384 ms time interval. Because increasing the correlation is equivalent to minimizing the squared error, minimizing \( CF_{k}(f) \) for consecutive frames increases the correlation and subsequently the STOI score. Also, by definition of the cost function in the form of Equation (11), the optimization is easier to handle.
An additional factor that is being considered in the cost function is the importance of a frame denoted by $\beta_k$. The importance of a frame is determined by an onset detector. If a frame is detected to be an onset, a higher $\beta_k$ is multiplied by $CF_k(f)$. The role of this additional weight is clarified in the following section that the optimization procedure is explained.

2.3. Optimization unit

The summed-cost-function (SCF) for frame number $k$ ($SCF_k$) is defined by weighted summation of cost functions of that frame ($CF_k(f)$) and upcoming frames ($CF_i(f)$):

$$SCF_k(f) = \sum_{i=k}^{P+k-1} \beta_i CF_i(f),$$

(12)

Each one-third octave band denoted by $f$ is optimized independently. To optimize a SCF, the number of frames influenced by the gain of a frame instead of $M$ is denoted by $P$ in Equation (12). According to the Equation (4) and considering the $T_{60}$, $M$ short frames are overlapped by an extended frame. But, to reduce the computational load, lower number of frames may be considered, because it is reasonable to neglect the effect of the reverberation after $P < M$ frames that is naturally shorter than $T_{60}$. Therefore, this implementation does not require to have a full-length signal available when optimizing a given frame and only a limited look ahead into the future is needed. This $P$ value is imperially set using the energy decay curve (EDC) of RIR when its value is dropping by 25 dB.

In Equation (12), $0 \leq \beta_i \leq 1$ is a coefficient that is determined by the onset detection unit. If frame number $k$ is detected to be an onset, a higher $\beta_i$ value is assigned to its cost function.

To prevent high fluctuations of weights in the optimization, lower and upper bounds for a weight $\alpha_{i,k}$ are set. The lower bound that is the minimum gain is set to -40dB. The output of the onset detection unit is also used here in the optimization routine. If a frame is detected to be an onset, the maximum possible gain (upper bound) is set to 20 dB, otherwise maximally 0 dB is allowed. In the optimization procedure, only positive weights are accepted because the parameter that is going to be controlled is the leaked energy of each frame on the upcoming frames. Since the optimization is targeting energy, the application of negative weights would be meaningless. The constrained optimization problem is summarized as follows:

$$\min SCF_k(f)$$

$$\{\alpha_{i,f}, k - P \leq i \leq k + P - 1\}$$

$$0 \leq \text{lower bound} \leq \alpha_{i,k} \leq \text{upper bound}$$

$$\text{lower bound} = -40\text{dB}$$

$$\text{upper bound} = \begin{cases} 0\text{dB, frame}^"i"\text{is an onset} \\ 20\text{dB, o.w} \end{cases}$$

(13)

To find the minimum of constrained nonlinear multivariable function of Eq. (13), a state-of-the-art method called the Sequential Quadratic Programming (SQP) [20,21] is used. In this numerical optimization method, the Hessian of the Lagrangian function using a quasi-Newton updating method is estimated. The algorithm starts independently for each
frequency band with calculating the coefficient of first frame \((\alpha_{1,f})\). The effect of the first frame is considered up to \(P\) frames after. Therefore, \(SCF_1(f)\) is a function of \(\alpha_{1,f}, \alpha_{2,f}, \ldots, \alpha_{p,f}\). All of these \(P\) weights are determined in the first optimization routine. But only \(\alpha_{1,f}\) is accepted as the final value, because other weights, except \(\alpha_{1,f}\), have an effect on the upcoming SCFs. In the next optimization \(\alpha_{2,f}\) is determined and being fixed and so on. After determination of a weight in each optimization, its value in all CFs is replaced. The routine for the determination of weights and updating the CFs are explained in algorithm 1.

3. Simulations and Measurements

3.1. Stimuli

The Oldenburg Sentence Test (OLSA) [22] corpus with the male speaker is used as the speech material to evaluate the proposed algorithm. The OLSA corpus consists of 120 German sentences for each speaker that a sentence includes name, verb, number, adjective and noun and there are 10 alternatives for each word. The corpus is downsampled to 16 kHz. Speech shaped noise (SSN) and pink noise (PN) are used as the interferers which are convolved with binaural room impulse response (BRIR) and presented at an average level of 65 dB-SPL for left and right ears. The SSN was generated by a summation of all 120 OLSA sentences followed by phase randomization, creating noise with a similar long-term spectrum as the speech corpus. In addition to SNN, also pink noise is used which has an energy distribution similar to environmental noise. It covers the frequency range between 100 Hz to 8 kHz approximately corresponding to the spectral range of the speech material.

3.2. Binaural Room Impulse Responses

In this section the BRIRs used in this paper are described. For the optimization and subjective evaluation, 4 rooms were used that from each of these rooms, 3 recorded BRIRs with a same receiver position are selected. These rooms are described in Table 1 and differ in terms of geometrical dimensions and reverberation time \(T60\). The main BRIR in Table 1 is convolved with a speech source and the resulting left ear signal is used for the optimization to find the weights, but is also used for rendering and evaluating the pre-processed speech signal. The second BRIR in this table is used to evaluate the robustness of the algorithm for a different position of the listener compared to the weights obtained for the main BRIR. Thus, the algorithm uses a different IR for the preprocessing than the IR that was originally obtained in the optimization. Finally, the third BRIR is used for convolving with a noise signal to create a binaural noise for both main and robustness evaluation scenarios. The first room (R1) with a relatively short \(T60\) of 0.6 s is selected from a set of BRIR measurements made at our university in Oldenburg. The second room (R2) is a music hall that is selected from the BRAS database [23] with a \(T60\) equal to 1.1 s.
Algorithm 1. Example text of a theorem. Determination of weights $a_{1,f}, a_{2,f}, \ldots, a_{k,f}, \ldots, a_{N,f}$ in the one-third octave (f).

**Step 1:** $k=1$

**Step 2:** The weight of frame number $k$ is determined. All the upcoming CFs that are influenced by $a_{k,f}$, construct the $SCF_k(f)$ according to Equation (12). For construction of $SCF_k(f)$, previous determined weights in CFs are used and the new weight $a_{k,f}$ is determined:

$$SCF_k(f) = \text{fun}(a_{1,f}, a_{2,f}, \ldots, a_{k-1,f}, a_{k,f}, a_{k+1,f}, \ldots, a_{P+k-1,f})$$

$$\approx \text{fun}(a_{k-p+1,f}, a_{k-1,f}, a_{k+1,f}, \ldots, a_{P+k-1,f})$$

optimization

$$\rightarrow \begin{array}{c} a_{k,f}, a_{k+1,f}, \ldots, a_{P+k-1,f} \\ \end{array}$$

For example, the $SCF_1(f)$ is a function of $a_{1,f}, a_{2,f}, \ldots, a_{P,f}$. All of these weights are determined in the first optimization. But, only $a_{1,f}$ is accepted as the final value:

$$SCF_1 = \text{fun}(a_{1,f}, a_{2,f}, \ldots, a_{P,f})$$

optimization

$$\rightarrow \begin{array}{c} a_{1,f}, a_{2,f}, \ldots, a_{P,f} \\ \end{array}$$

**Step 3:** All CFs that are influenced by $a_{k,f}$ are replaced by its numerical value.

$$CF_k(f) = \text{fun}(a_{k+p-1,f}, a_{k+1,f}, \ldots, a_{k,f})$$

$$CF_{k+1}(f) = \text{fun}(a_{k+p,f}, a_{k+1,f})$$

$$M$$

$$CF_{k+p-1}(f) = \text{fun}(a_{k,f}, a_{k+1,f}, \ldots, a_{k+p-1,f})$$

For example, for $a_{1,f}$:

$$CF_1(f) = \text{fun}(a_{1,f})$$

$$CF_2(f) = \text{fun}(a_{1,f}, a_{2,f})$$

$$M$$

$$CF_p(f) = \text{fun}(a_{1,f}, a_{2,f}, \ldots, a_{P,f})$$

**Step 4:** $k=k+1$

If $k<N$ go to step 2
else finish

The recorded BRIRs in R2 have a relatively long distance between the source and receiver and therefore the direct-to-diffuse ratio is low. The third room (R3) selected again from the BRAS database is a seminar room with a T60 equal to 1.5 s. This room is critical in terms of T60 and speech intelligibility because it has a long reverberant tail that creates a considerable amount of time smearing of the source speech signal. Room (4) represents a church selected from Air database [24]. The T60 is very long (about 5 sec) because of the large dimensions of the room and low degree of damping. Because of relatively small source-receiver distances (3 m), however, the selected BRIRs in the church has a high direct-to-reverb diffuse ratio.
Table 1. The BRIRs used for the optimization and subjective evaluation. The room names in the database, reverberation times, the length of P frames and selected BRIRs are mentioned.

| Room  | Name in database | T60 (s) | length of P frames (s) | Main BRIR | BRIR for Robustness Evaluation | BRIR of Noise |
|-------|------------------|---------|------------------------|-----------|-------------------------------|---------------|
| R1    | VarEcoic         | 0.60    | 0.25                   | kas_none_r00_az000 | kas_none_r00_az030            | kas_none_r00_az030 |
| R2    | Music Hall       | 1.1     | 0.45                   | CR3_BRIR_LS7_MP6_HATO0 | CR3_BRIR_LS7_MP6_HATO0       | CR3_BRIR_LS7_MP6_HATO0 |
| R3    | Seminar Room     | 1.5     | 0.6                    | CR2_BRIR_LS7_MP6_HATO0 | CR2_BRIR_LS7_MP6_HATO0       | CR2_BRIR_LS7_MP6_HATO0 |
| R4    | Church           | 5.0     | 1                      | air_binaural_ula_carolina_1_1_3_90 | air_binaural_ula_carolina_1_1_3_135_3 | air_binaural_ula_carolina_1_1_3_135_3 |

The distance from speech and noise sources to the listener positions of the main and robustness evaluation scenarios is almost held constant within a room to avoid differences in the direct-to-diffuse ratio. A collection of all used room-acoustical scenarios, reverberation times, selected BRIRs and the length of P frames are shown in Table 1. P is a number of future frames that is used to construct the summed-cost-function (SCF) in Equation (12) for the optimization purpose. As previously explained, the length of P frames is determined according to 25 dB drop of EDC. For a larger T60, more future frames are needed for the optimization.

3.3. Signal Processing Details

The corpus and noises are downsampled to 16 kHz. The length of the analysis and synthesis frame is 30 ms with 50% overlap. A square-root Hann window is used in the signal framing in both the analysis and synthesis to avoid audible artifacts because of the cyclic convolution. To synthesis the signal, the overlap-add (OLA) method is used. The frequency resolution used in separation of an extended frame in Equation (3) to the one-third octave bands is limited by the length of P frames according to Table 1. According to length of P frames and the sampling rate of 16 kHz, it could be at least 4096 for the shortest room impulse response (T60 = 0.6s) and 16384 for the longest room impulse response (T60 = 5s). The bins are grouped into 17 one-third octave band. Similar to the STOI [19], the lowest center frequency is set to 150 Hz and the highest one-third octave band has a center-frequency approximately equal to 6 kHz. A frequency resolution used in Equation (8) for analysis of weighted sub-frames in Equation (5) is determined by the window length and sampling rate equal to 512 bins. The RMS values of the processed signal is adjusted to that of the unprocessed signal to keep levels equal between the output of the NELE algorithms and the unprocessed signal.

3.4. Effect of the Algorithm on Signal

The cochleagram of a clean speech from the OLSA corpus (first raw) and two preprocessed speech signals, one of them pre-processed with the OE algorithm [16] (second raw) and another one preprocessed by the proposed algorithm (third raw) are depicted in Figure 3(a). The weights are calculated for room R4. It can be seen that the preprocessing of OE and the proposed algorithm causes a high-pass filter effect on the speech, which is
Figure 3. (a) The cochleagram of clean speech and two preprocessed speech signals are shown. The first-row panel shows the unprocessed speech sentence. Second-row panel shows a pre-processed speech sentence using OE algorithm and the third-row panel shows the preprocessed speech using the proposed algorithm obtained for room R4. (b) The cochleagram of three reverberated signals in room R4 are shown. The first-row panel shows the unprocessed clean sentence. In second-row panel, a reverberated unprocessed speech sentence is depicted. Third and fourth row panels show a reverberated pre-processed speech sentence using OE algorithm and the proposed algorithm respectively.

caused by the fact that both enhancement algorithms reduced the amplitude of speech portions with a high and constant energy over time or that are exposed to a longer T60. It can be seen that this attenuation is stronger for the proposed method in comparison to the OE. Because of the importance of onsets for intelligibility, and in accordance with Equation (12), onsets are more strongly weighted in the proposed method. The other steady portions, on the other hand are allowed to be attenuated more in order to minimize the defined cost function.

The signals of Figure 3(a) are now convolved with the left ear of BRIR in room R4 and their cochleagrams are plotted in. The cochleagram of reverberated unprocessed, reverberated preprocessed by OE and reverberated preprocessed by the proposed algorithm are shown the first, second and third panels of Figure 3(b), respectively. Beside the onset enhancement and high-pass filtering effect of the proposed approach, the effect of the frame attenuations can be compared with the two other reverberated signals. The effect can be seen in the longer silent gap of the clean speech around second 1. It can be seen that in the third panel belonging to the proposed method, there is a low amount energy leakage from previous frames into the silent gap in comparison to the unprocessed and OE-preprocessed reverberated signals. A similar effect can be seen in second 1.6. This effect can potentially contribute to higher speech intelligibility due to the reduced overlap masking of preceding speech segments.
3.4. Objective Evaluation of the Algorithm Using two intelligibility Models

For the objective evaluation of the proposed algorithm, the left ear of above-mentioned BRIIRs were used for the weight computations of the OE and proposed algorithm. The BRIIRs were then convolved with OLSA speech material that was either unprocessed, OE-preprocessed, or preprocessed with proposed method. Two intelligibility models, the STOI [19] and the multi-resolution generalized power-spectrum model (mr-GPSM) [24] were used. For both intelligibility models, the intelligibility scores were averaged across 120 sentences. For the STOI, the unprocessed and preprocessed reverbered signals without additive noise are compared with clean speech. The reason for the selection of STOI is the fact that its metric is very similar to our defined cost function using the optimization method. It is expected that by minimizing the cost function in Equation (11) based on the square error, the STOI score based on correlation to be also improved. For evaluation with mr-GPSM, the SSN and PN without convolving with BRIIRs are added to the reverbered speech materials in three SNR values of -15, -10 and 0 dB. This is done to keep the influence of background noise identical across conditions. For each reverbered sentence, different samples of the full noise token are added. To have a good averaging across noise samples, 5 additional runs for each set of reverbered speech materials are performed.

Figure 5 shows the STOI score of the main scenario in each room for reverbered unprocessed, reverbered OE-enhanced and the reverbered enhanced signal with the proposed algorithm. The OE does not show much difference compared to the unprocessed case but, the improvement for the proposed approach is considerable. The STOI improvement for rooms R1, R2 and R3 is in about 0.05 but, lower improvement is seen for room R4 with very high reverberation time.

Figure 6(a) shows the predictions of mr-GPSM for SSN. The SNR-mr-GPSM is the intelligibility score of this intelligibility model which is a summation of envelope-SNR and DC-SNR [24]. The speech reception threshold (SRT) is an increasing monotonic function of SNR-mr-GPSM. A comparison between three signals for each panel shows that both OE and the proposed approach enhance the speech intelligibility. But, the improvement for the proposed algorithm is considerable in comparison to the OE method.

**Figure 5.** The STOI scores for the main scenario in each room for reverbered unprocessed, reverbered OE-enhanced and the reverbered enhanced signal with the proposed algorithm. The proposed algorithm shows improvement for the STOI score in comparison to the reverbered unprocessed and OE-enhanced signals.
Figure 6. (a) The predictions of mr-GPSM in the presence of SSN for three SNR values. The intelligibility scores (SNR-mr-GPSM) show up to 5 dB for OE and up to 15 dB improvement in comparison to the unprocessed signal in low SNR = -15 dB (first-raw panel). The improvement is seen for other SNRs specifically in rooms R2 and R3. The model shows lower improvement for preprocessed signals in rooms R1 and R4.; (b) The predictions of mr-GPSM in the presence of PN for three SNR values. The intelligibility scores (SNR-mr-GPSM) show less than 1 dB improvement for OE and in about 2 dB improvement using the proposed approach.

Specifically for low SNR = -15 dB, this improvement is in about 7 SNR-mr-GPSM for rooms R2 and R3. Generally, the model shows less improvement for rooms R1 and R4. The intelligibility score of R1, because of lower T60, is higher than others and therefore this score may reach near to its maximum possible value such that more improvement is not possible specifically for high SNRs near 0 dB (third-row panel in Figure 6(a)). For room R4 with a long reverberation tail, the model, because of high amount of time smearing, does not show much improved scores specifically for more noisy conditions with SNR = -10 and -15 dB. The same evaluation using mr-GPSM is shown in Figure 6(b), now using PN. The model shows lower scores for the PN in comparison to SSN and also lower improvements caused by the preprocessing algorithm. Generally, the model predictions show less than 1 dB improvement for OE and about 2 dB improvement using the proposed method. In spite of lower values of improvements, the curves of Figure 6(b) show a consistent increase of speech intelligibility using the OE and the proposed approaches.

The intelligibility prediction models are also applied to the robustness evaluation scenarios. In Figure 7, the STOI scores similar to Figure 5 shows an improvement of intelligibility for the proposed method in comparison to the unprocessed and OE-enhanced signals.
Figure 7. The STOI scores for robustness evaluation scenarios in each room for reverber unprocessed, reverbered OE-enhanced and the reverbered enhanced signal with the proposed algorithm. The proposed algorithm similar to the main scenarios in Figure 5 shows improvement for the STOI score in comparison to the reverber unprocessed and OE-enhanced signals.

In Figure 8, the SNR-mr-GPSM scores for evaluation of robustness of the proposed algorithm is compared with the unprocessed and OE-enhanced signals. The improvements are in the range obtained for the main scenarios in Figure 6.

Figure 8. (a) The predictions of mr-GPSM for robustness evaluation scenarios in the presence of SSN. The SNR- mr-GPSM show up to 3 dB enhancement for OE and up to 10 dB improvement in comparison to the unprocessed signal in low SNR = -15 dB; (b) The predictions of mr-GPSM in the presence of PN for three SNR values. The intelligibility scores (SNR-mr-GPSM) show in about 2 dB improvement for OE and in about 4 dB improvement using the proposed approach.
In Figure 8(a) the maximum improvement predicted by the mr-GPSM in the presence of SSN are for the OE-preprocessed signal is 3 dB and for the proposed approach it is about 9 dB. For the PN, a similar tendency can be seen in the three panels of Figure 8(b). The data show an overall improvement of 3 to 4 dB for the proposed method depending on the scenario. This improvement is sometimes better than the main scenario for SSN and PN. This underlines that the proposed algorithm is very robust against changes in listener position and that detailed knowledge of the IRs is not essential. The intelligibility score is more dependent on the listening scenario and was sometimes better than the main scenario for which the weights were calculated because of more binaural advantage caused by more azimuthal separation of the target and noise sources.

3.5. Subjective Evaluation

Figure 9 shows the 50% speech reception thresholds (SRT\%) for 8 subjects obtained with the OLSA matrix test for all four room-acoustical scenarios and SSN for the unprocessed, OE and the proposed approach. In the left panel, median values are shown together with the 25% and 75% quantiles and outliers across eight subjects. The right panel shows the mean values and the standard error across the subject’s mean values and the standard error at the most right-hand side is calculated across all subjects and rooms.

Considering all rooms, it can be seen in Figure 9(b) that the intelligibility is enhanced up to 3.5 dB for OE and 5 dB for the proposed approach compared to the unprocessed speech. In the rooms R2 and R3, the proposed approach has a slightly larger effect on intelligibility which may be caused by the midrange values of T60. Data obtained in rooms R2 and R3 for SSN show an improvement in about 1.5 dB and 1 dB for the proposed method in comparison to the OE. The proposed approach shows only a small improvement in Rooms R1 and R4, which is due to the low and very high reverberation times that make more improvements difficult. Room 4, which is a church with a reverberation time of T60 = 5 s, shows no large difference between the proposed and OE approach. For this room, in spite of the higher reverberation time, the SRTs are lower in comparison to that of rooms R2 and R3. This is mainly because of high direct-to-reverberant ratio in the BRIRs of the

Figure 9. (a) Boxplots of speech reception thresholds at 50% speech intelligibility (SRT\%) for 8 subjects, measured in the presence of SSN are illustrated. The unprocessed and the preprocessed signals for the main scenarios using the OE and proposed method are compared. (b) The mean SRT\% and the standard errors for the same data in (a) are plotted.
Figure 10. (a) Boxplots of speech reception thresholds at SRT₅₀ for 8 subjects in the presence of PN are illustrated. The unprocessed and the preprocessed signals for the main scenarios using the OE and proposed method are compared; (b) The mean- SRT₅₀ and the standard errors for the same data in (a) are plotted.

church and also the larger azimuth angle difference between source and the noise for BRIRs in the church.

In Figure 10, the same plots are depicted, but this time for the PN interferer. Both the OE and proposed approach shows fairly good improvement in comparison to the unprocessed signal. But altogether a low improvement of about 1 dB is seen for the proposed approach in comparison to the OE. Only for room 2, there is more than 1 dB improvement.

The results of SRT₅₀ measurements for the robustness evaluation scenarios are shown in Figures 11 and 12 for SSN and PN, respectively. These figures show the SRTs optimized on the main scenario and applied to the robustness evaluation positions. For the SSN in Figure 13, a comparison between the three signals shows that for SSN there is an improvement up to 3.5 dB for the OE and 5.5 dB for the proposed method relative to the unprocessed signal. A similar tendency can be seen in Figure 14 for PN. The data in Figure 14 show an overall improvement of 1.5 to 3 dB for OE and 2 to 3.5 dB for the proposed approach depending on the room. In general, comparing the thresholds of the robustness evaluation scenario with that of main scenarios, it can be seen that, similar to the predictions of the intelligibility models, both the OE and proposed methods are very robust against changes in position and a detailed knowledge of the IR is not necessary.
Figure 11. (a) Boxplots of speech reception thresholds at SRT50 for 8 subjects to evaluate the robustness of preprocessing methods, measured in the presence of SSN are illustrated. The unprocessed and the preprocessed signals for the robustness evaluation scenarios using the OE and proposed method are compared.; (b) The mean-SRT50 and the standard errors for the same data in (a) are plotted.

Figure 12. (a) Boxplots of speech reception thresholds at SRT50 for 8 subjects to evaluate the robustness of preprocessing methods in the presence of PN are illustrated. The unprocessed and the preprocessed signals for the main scenarios using the OE and proposed method are compared.; (b) The mean- SRT50 and the standard errors for the same data in (a) are plotted.

4. Discussion

In this study a new reverb-based NELE approach based on the optimization of a cost function was proposed that reduces the time-smearing effect of reverberation on speech and similar to the OE amplifies the onsets and has high-pass filter characteristics. Its main advantage to the OE is considering future frames in finding filtering weights and for this reason, more reduction of the overlap masking of reverberation tail is seen. The amount of overlap masking is considered in the defined cost function and is used to control the weights applied on the original speech signal speech segments that would make the upcoming frames inaudible. Higher importance for the onset segments in the cost function is assigned to avoid attenuation of onsets which would decrease intelligibility. Both the
model predictions and listening-test results showed improvement in SRTs. It has been demonstrated by the model prediction that the proposed algorithm is better able to compensate for the detrimental effects of reverberation than the OE method. The subjective evaluation showed that depending on the scenario there is an improvement of 0.5 dB up to 2 dB in comparison to the OE. The mr-GPSM model predicts a larger improvement for the proposed method over the OE method than what is actually observed in the listening tests. One reason could be the stronger artifacts created by the proposed algorithm compared to the OE method. A possible modification of the method could be using a quality-assessment criteria in the optimization procedure. The algorithm evaluation using both the intelligibility models and the listening test showed that improvements in speech intelligibility did not depend on having an exact match between the positions of the source and the listener used for obtaining the optimal weights. Similar to the OE approach, this underlines the robustness of this algorithm for errors in the estimation of room impulse responses. For the proposed algorithm only a course spectro-temporal representation of the room impulse response is used and the exact magnitudes and phases of the transfer function are not needed. Therefore, the robustness problem that exists in the inverse-filtering approaches is avoided in the proposed method.

Another important point about the proposed approach is the fixed parameters being used in construction of the cost function and optimization. The fixed set of parameters are used for all of scenarios. The performance of the algorithm is dependent on the parameters that are set. The first parameter is the number of the future frames (P) that are considered in the summed-cost-function (SCF) of Equation (12) which is based on the reverberation time. To reduce the computational load, it could set much less than the number of frames covered by the T60. Informal listening tests showed that beyond a specific number of future frames, the signal is not much more improved. Imperially, P was determined according to the EDC of the RIR, until the point it drops 25 dB. Other parameters that are fixed empirically are the weights (β) assigned to onsets in the defined SCF again in Equation (12) and the lower and upper bound for the gains in Equation (13). For the future work these parameters could be selected according to an intelligibility model.

A disadvantage of the proposed algorithm is its high computational load for the determination of weights although, faster algorithms may be found in the future. The proposed method is in the category of reverberation-base NELE algorithms. According to the literature, algorithms that use priori knowledge of the maskers and RIRs do not perform better than noise-independent algorithms. The ASE and SSDRC approaches that are not using the characteristics of the playback environment outperformed other methods in the NELE challenge [25]. It is surprising that until now enhancement algorithms with the goal of enhancing the noise and reverberation effect on the speech has not performed well. A promising approach could be a combination of three categories of NELE algorithms including rule-based, noise-dependent and reverberation-dependent to benefit from the advantages of separate methods. For example in Adaptive Compressive Onset-Enhancement (ACO) method [26], sequential and independent combination of a modified version of the AdaptDRC [6] and the OE [16] is used to enhance the speech in a reverb and noisy room with the knowledge of statics of additive noise and RIR respectively.

**Author Contributions:** Conceptualization, A.F. and S.v.; methodology, A.F. and S.v.; software, A.F.; validation, A.F. and S.v.; formal analysis, A.F.; investigation, A.F.; resources, A.F. and S.v.; data curation, A.F.; writing—original draft preparation, A.F.; writing—review and editing, S.v.; visualization, A.F. and S.v.; supervision, S.v.; project administration, S.v.; funding acquisition, S.v. All authors have read and agreed to the published version of the manuscript.
**Funding:** This research was funded by Deutsche Forschungsgemeinschaft as part of the SFB-1330 HAPPA, project C2, grant number 5620221459.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Carl von Ossietzky University of Oldenburg.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Not Applicable.

**Acknowledgments:** The authors would like to thank Christoph Kirsch and Thomas Biberger for providing recorded BRIRs and the mr-GPSM speech intelligibility model.

**References**

1. Sauert, B.; Vary, P. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006; pp. 14-19.
2. Zorilă, T.-C.; Stylianou, Y.; Ishihara, T.; Akamine, M. Near and far field speech-in-noise intelligibility improvements based on a time–frequency energy reallocation approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2016, 24, 1808-1818.
3. Zorila, T.-C.; Kandia, V.; Stylianou, Y. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, 2012.
4. 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* 1986, 80, 1599-1607.
5. Chermaz, C.; King, S. A sound engineering approach to near end listening enhancement. In Proceedings of the Proceedings of Interspeech, 2020.
6. Schepker, H.; Rennies, J.; Doclo, S. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. *The Journal of the Acoustical Society of America* 2015, 138, 2692-2706.
7. Schepker, H.; Hülsmeier, D.; Rennies, J.; Doclo, S. Model-based integration of reverberation for noise-adaptive near-end listening enhancement. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, 2015.
8. Chermaz, C.; Valentini-Botinhao, C.; Schepker, H.F.; King, S. Evaluating Near End Listening Enhancement Algorithms in Realistic Environments. In Proceedings of the INTERSPEECH, 2019; pp. 1373-1377.
9. Taal, C.H.; Hendriks, R.C.; Heusdens, R. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Computer Speech & Language* 2014, 28, 858-872.
10. Li, H.; Fu, S.-W.; Tsao, Y.; Yamagishi, J. iMetricGAN: Intelligibility Enhancement for Speech-in-Noise using Generative Adversarial Network-based Metric Learning, *arXiv preprint arXiv:2004.00932 2020.*
11. Schädler, M. Optimization and evaluation of an intelligibilityimproving signal processing approach (IISPA) for the Hurricane Challenge 2.0 with FADE. In Proceedings of the Proceedings of Interspeech, 2020.
12. Mertins, A.; Mei, T.; Kallinger, M. Room impulse response shortening/reshaping with infinity-and $ p $-norm optimization. *IEEE Transactions on Audio, Speech, and Language Processing* 2009, 18, 249-259.
13. Kusumoto, A.; Arai, T.; Kinoshita, K.; Hodoshima, N.; Vaughan, N. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication* 2005, 45, 101-113.

14. Arai, T.; Hodoshima, N.; Yasu, K. Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners. *IEEE transactions on audio, speech, and language processing* 2010, 18, 1775-1780.

15. Petkov, P.N.; Stylianou, Y. Adaptive gain control for enhanced speech intelligibility under reverberation. *IEEE Signal Processing Letters* 2016, 23, 1434-1438.

16. Grosse, J.; van de Par, S. A speech preprocessing method based on overlap-masking reduction to increase intelligibility in reverberant environments. *Journal of the Audio Engineering Society* 2017, 65, 31-41.

17. Arai, T.; Kinoshita, K.; Hodoshima, N.; Kusumoto, A.; Kitamura, T. Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. *Acoustical science and technology* 2002, 23, 229-232.

18. Hodoshima, N.; Arai, T.; Kusumoto, A.; Kinoshita, K. Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments. *The Journal of the Acoustical Society of America* 2006, 119, 4055-4064.

19. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing* 2011, 19, 2125-2136.

20. Nocedal, J.; Wright, S. *Numerical optimization*; Springer Science & Business Media: 2006.

21. Fletcher, R. *Practical methods of optimization*; John Wiley & Sons: 2013.

22. Wagener, K.; Brand, T.; Kollmeier, B. Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test. *Zeitschrift Fur Audiologie* 1999, 38, 86-95.

23. Aspöck, L.; Vorländer, M.; Brinkmann, F.; Ackermann, D.; Weinzierl, S. Benchmark for Room Acoustical Simulation (BRAS). 2020.

24. Biberger, T.; Ewert, S.D. The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking. *The Journal of the Acoustical Society of America* 2017, 142, 1098-1111.

25. Rennies, J.; Schepker, H.; Valenti-Botinhao, C.; Cooke, M. Intelligibility-enhancing speech modifications–the hurricane challenge 2.0. *Proc. Interspeech, Shanghai, China* 2020.

26. Bederna, F.; Schepker, H.; Rollwage, C.; Doclo, S.; Pusch, A.; Bitzer, J.; Rennies, J. Adaptive compressive onset-enhancement for improved speech intelligibility in noise and reverberation. In Proceedings of the Proceedings of Interspeech, 2020.