SII-Based Speech Prepocessing for Intelligibility Improvement in Noise

Taal, Cees H. ; Jensen, Jesper

Published in:
Proceedings of the International Conference on Spoken Language Processing

Publication date:
2013

Document Version
Early version, also known as pre-print

SII-based Speech Preprocessing for Intelligibility Improvement in Noise

Cees H. Taal[1], *Jesper Jensen*[2,3]

[1]Leiden University Medical Center, Leiden, The Netherlands
[2]Aalborg University, Aalborg, Denmark
[3]Oticon A/S, Smørum, Denmark

c.h.taal@lumc.nl, jje@es.aau.dk, jsj@oticon.dk

Abstract

A linear time-invariant filter is designed in order to improve speech understanding when the speech is played back in a noisy environment. To accomplish this, the speech intelligibility index (SII) is maximized under the constraint that the speech energy is held constant. A nonlinear approximation is used for the SII such that a closed-form solution exists to the constrained optimization problem. The resulting filter is dependent both on the long-term average noise and speech spectrum and the global SNR and, in general, has a high-pass characteristic. In contrast to existing methods, the proposed filter sets certain frequency bands to zero when they do not contribute to intelligibility anymore. Experiments show large intelligibility improvements with the proposed method when used in stationary speech-shaped noise. However, it was also found that the method does not perform well for speech corrupted by a competing speaker. This is due to the fact that the SII is not a reliable intelligibility predictor for fluctuating noise sources. MATLAB code is provided.

Index Terms: Speech intelligibility, speech enhancement, near-end enhancement, speech intelligibility index

1. Introduction

Intelligibility in speech communication systems can be negatively affected by background noise originating from both the far-end and the near-end side of the communication channel. Here we assume that the listener is located at the near-end. In order to eliminate the negative impact of the far-end noise, one would typically apply a (single-channel) noise-reduction algorithm (see [1] for an overview). However, the speech can also be pre-processed before playback in order to become more intelligible in presence of the near-end background noise. The latter approach is the focus in this work where typical examples can be found in the field of telephony and public address systems. The common assumptions here are that a clean version of the far-end signal is available (i.e., the potential noise is assumed to be successfully suppressed) and that we have knowledge of the near-end noise statistics [2]. One obvious solution to the near-end listening enhancement problem is to increase the playback level of the speech. However, at a certain point, increasing the playback level may not be possible anymore due to loudspeaker limitations or unpleasant playback levels. Therefore, a common approach is to fix the speech energy (i.e. the global SNR remains unaltered) and redistribute speech energy over time and/or frequency [3, 4, 2].

One effective and simple way to improve speech intelligibility is by changing the spectrum of the speech. For example, speech understanding will increase when high frequencies are amplified at a cost of low frequencies [5, 6, 7, 8, 9]. However, the exact design of these high-pass type of filters are often heuristic in nature and do not use some type of mathematical descriptor of speech intelligibility. This makes it difficult to claim any form of optimality. As a consequence, many different solutions exist, for example: (1) it was suggested to 'whiten' the speech spectrum (independent of noise type) [5, 7], (2) shape the speech spectrum such that SNRs are equal in each frequency band [3, 10] or (3) adjust the speech spectrum such that is shaped as the inverse noise spectrum [3, 11]. In [12] we proposed a linear filter which maximizes the speech intelligibility index (SII) [13]. Experiments showed large intelligibility improvements with this method over the unprocessed noisy speech and better performance than one state-of-the art method [2]. However, [12] did not provide an extensive analysis of algorithm performance as a function of SNR and noise type. It would be of interest to see its behavior in relation with the three previously mentioned different approaches.

In this paper an additional analysis is provided for the proposed method [12] in terms of processed speech spectra, frequency-dependent SNRs and filter gains. Moreover, our method participated in the Hurricane Challenge[1] [14, 15] for which we will also report the listening test results in this paper. These results also include a non-stationary noise condition; we expect this to be a difficult condition for the proposed method since the SII is not reliable with non-stationary noise sources [16]. First the mathematical details of the proposed method will be summarized followed by analysis of our method and experimental results.

2. SII-based linear filter

2.1. Intelligibility Measure

The intelligibility measure which will be used for optimization is based on the standardized SII [13]. We assume that the speech and noise are presented above the threshold in quiet at a comfortable level. Also, effects of masking are excluded from the standard SII procedure (as in [4]). Based on these assumptions the approximated SII measure can be summarized by the following three stages: (1) The long-term average spectra of the speech and noise are estimated within critical bands. (2) A within-band SNR is calculated, clipped between -15 and 15 dB followed by normalization to the range of 0 and 1. (3) A weighted average of the normalized within-band SNRs is calculated to obtain one outcome.

Next, details are given for each stage. Let x and ε denote the time-domain signals of the clean speech and noise, respectively.

[1]Our method is abbreviated as OptimalSII in [14] where the Hurricane Challenge results are reported.

A windowed version of x is denoted by x_m where m denotes the window frame-index. A Hann-window is used with 50% overlap, and 32 ms length. The impulse response of the i^{th} auditory filter is denoted by h_i, where $i \in \{1, \ldots, n\}$ and n is the total number of auditory filters. Subsequently, the energy within one time-frequency (TF) unit is calculated as follows for the clean speech,

$$X_{m,i}^2 = \sum_k |X_m(k)|^2 |H_i(k)|^2, \quad (1)$$

where $X_m(k)$ and $H_i(k)$ denote DFT coefficients of x_m and h_i, respectively, with frequency-bin index k. Signals are sampled at 20 kHz where short-time frames are zero-padded to 64 ms before applying the DFT. In total, 64 auditory filters are used where center frequencies are linearly spaced on an equivalent rectangular bandwidth (ERB) scale between 150 and 8500 Hz [17]. Its squared magnitude responses $|H_i(k)|^2$ are chosen as described in [17]. The average energy within one critical band is based on a long-term sample mean over many short-time frames (e.g., several minutes) and is denoted as follows,

$$\sigma_{X_i}^2 = \frac{1}{M} \sum_m X_{m,i}^2, \quad (2)$$

where M equals the total number of short-time frames and similar definitions hold for $\sigma_{\mathcal{E}_i}^2$, the average noise energy within critical band i. Let the SNR within one critical band be denoted by,

$$\xi_i = \frac{\sigma_{X_i}^2}{\sigma_{\mathcal{E}_i}^2}, \quad (3)$$

which is used to calculate an intermediate measure to determine the audibility of the speech in presence of the noise within one band. This SNR is log-transformed, clipped between -15 and +15 dB and normalized such that its range is between zero and one, i.e.,

$$d(\xi_i) = \max\left(\min\left(10\log_{10}(\xi_i), 15\right), -15\right)/30 + \frac{1}{2}. \quad (4)$$

Subsequently, a weighted average is calculated as follows,

$$SII = \sum_i \gamma_i d(\xi_i), \quad (5)$$

where γ denotes the band-importance function given in the critical-band SII procedure in Table 1 in [13]. In summary, this weighting-function reduces the importance of bands with center frequency below 450 Hz and above 4000 Hz. It is expected that Eq. (5) is a monotonic increasing function of the intelligibility of speech in additive, stationary noise [13].

2.2. Constrained Optimization

The goal is to maximize the speech intelligibility, i.e., maximize Eq. (5), by redistributing the speech energy over the critical bands. Hence, the total energy over all bands remains unchanged. We constrain ourselves to redistribute speech energy using a linear, time-invariant filter. In practice one could estimate the statistics online, e.g., with a noise-tracker [18], and use a time-varying filter. The same mathematical framework can be used for this time-varying case. Let α_i be a real and non-negative scalar applied to each critical band. It follows that $\frac{1}{M}\sum_m (\alpha_i x_{m,i})^2 = \alpha_i^2 \sigma_{X_i}^2$. As a consequence, the constrained problem can be formulated as follows,

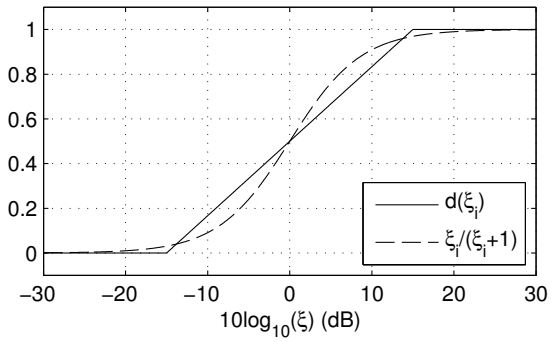

Figure 1: Used error criteria for intelligibility prediction as used by the SII (solid line, Eq. (4)), and proposed approximation (dashed line, Eq. (7)).

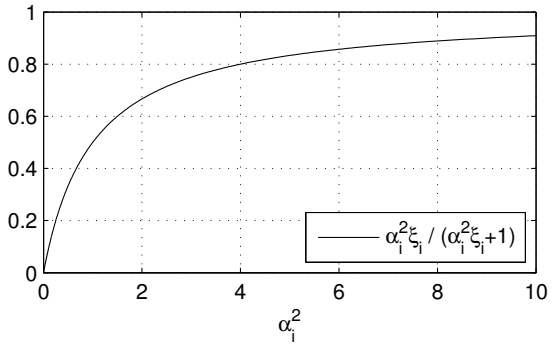

Figure 2: The proposed approximation of the SII, as in Eq. (7), is concave in $\alpha_i^2 x_i^2$. Figure shows results for $\xi = 1$.

$$\begin{aligned} \max \quad & \sum_i \gamma_i d\left(\alpha_i^2 \xi_i\right) \\ \text{s.t.} \quad & \sum_i \alpha_i^2 \sigma_{X_i}^2 = \sum_i \sigma_{X_i}^2 \\ & \alpha_i^2 \sigma_{X_i}^2 \geq 0, \forall i \end{aligned} \quad (6)$$

In order to find a closed-form solution to this constrained optimization problem we propose to approximate Eq. (4) with the following expression which is mathematically tractable,

$$d(\xi_i) \approx \frac{\xi_i}{\xi_i + 1}. \quad (7)$$

Interestingly, $d(\xi_i)$ is the expression for the single-channel Wiener filter [1]. This approximation together with the original intermediate intelligibility, as defined in Eq. (4), is shown in Figure 1. Moreover, the function $d(\alpha_i^2 \xi_i)$ is concave in its argument as illustrated in Figure 2. Hence, the weighted average of these concave functions, as in Eq. (5), is also concave. We obtain convexity by negation and characterize the problem by the following Lagrangian cost-function,

$$J = -\sum_i \gamma_i \frac{\alpha_i^2 \xi_i}{\alpha_i^2 \xi_i + 1} + \nu \left(\sum_i \alpha_i^2 \sigma_{X_i}^2 - r \right) + \sum_i \lambda_i \left(-\alpha_i^2 \sigma_{X_i}^2 \right), \quad (8)$$

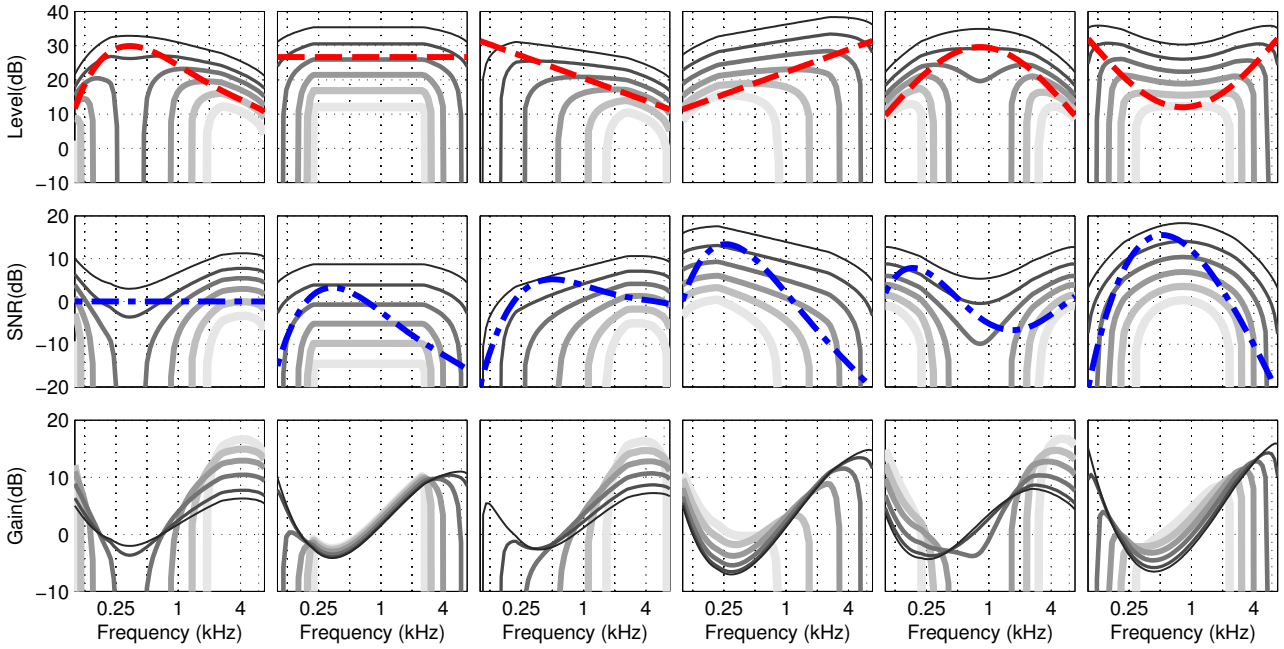

Figure 3: Auditory band spectra (top row), SNRs (middle row) and filter gains (bottom row) for the proposed processing method for six artificially generated noise types (from left to right: speech-shaped, white, low-pass, high-pass, bandpass and bandstop). Red-dashed lines denote noise spectra and blue dot-dashed lines denote unprocessed frequency-dependent SNRs. Used speech spectrum is the same as the speech-shaped noise spectrum depicted by the red line in the top-left plot. Six different global SNRs are used (-20, -15, -10, -5, 0 and 5 dB), where the thickest and thinnest line denote -20 and 5 dB SNR, respectively.

where $r = \sum_i \sigma_{X_i}^2$ and ν and λ_i are Lagrangian multipliers related to the energy constraint and inequality constraints in Eq. (6), respectively. Since our problem is convex and differentiable, any point that satisfies the following Karush-Kuhn-Tucker (KKT) conditions,

$$\begin{aligned} \sum_i \alpha_i^2 \sigma_{X_i}^2 &= r \\ \alpha_i^2 \sigma_{X_i}^2 &\geq 0 \quad ,\forall i \\ \lambda_i &\geq 0 \quad ,\forall i \\ \lambda_i \alpha_i^2 \sigma_{X_i}^2 &= 0 \quad ,\forall i \\ \frac{-2\gamma_i \xi_i \alpha_i}{(\xi_i {\alpha_i}^2+1)^2} + 2\nu\alpha_i\sigma_{X_i}^2 - 2\lambda_i\alpha_i &= 0 \quad ,\forall i \end{aligned} \qquad (9)$$

is optimal [19, Ch. 5.5.3, p. 243]. Solving gives,

$$\alpha_i^2 \sigma_{X_i}^2 = \max\left(\frac{\sigma_{\mathcal{E}_i}\sqrt{\gamma_i}}{\sqrt{\nu}} - \sigma_{\mathcal{E}_i}^2, 0\right), \forall i, \qquad (10)$$

where ν is chosen such that the energy constraint is satisfied,

$$\frac{1}{\sqrt{\nu}} = \frac{r + \sum_{i\in\mathcal{M}} \sigma_{\mathcal{E}_i}^2}{\sum_{i\in\mathcal{M}} \sqrt{\gamma_i}\sigma_{\mathcal{E}_i}}, \qquad (11)$$

and where $\mathcal{M} = \{i \in \{1,\ldots,n\} : \alpha_i^2 > 0\}$ denotes the set of critical band indices for which the optimal α_i^2 is positive. Since the set $\mathcal{M}$ depends on α_i^2, the Lagrange multiplier ν is also dependent on α_i^2. In order to cope with this recursive dependency, the optimal value of ν may be found by evaluating Eq. (10) for a range of ν-values or, e.g., using a bi-section method [19, Ch. 4.2.5, p. 146] such that the energy constraint is satisfied.

3. Filter Analysis

The behavior of the proposed method is investigated for six artificially generated stationary noise types including speech shaped noise (SSN), white noise, noise with a low-pass and high-pass characteristic and speech with a band-pass and band-reject type of spectrum. The six noise spectra are shown in the the top row of plots in Figure 3 by the red-dashed lines. The long-term speech spectrum is the same as the spectrum of SNN as depicted in the top-left plot. Results are analyzed for six different global SNRs (-20, -15, -10, -5, 0 and 5 dB). Each global SNR is indicated by a line differing in thickness and darkness, where the thickest and lightest line shows results for -20 dB SNR and the thinnest and darkest line is related to 5 dB SNR. The top-row plots show the modified speech spectra, the middle-row plots the frequency dependent long-term SNRs and the bottom-row plots the magnitude response of the applied filter. Note, that all plots show critical-band spectra as in Eq. (1).

Regarding the processed spectra as observed in the top-row plots in Figure 3 it is clear that the shape of the spectrum does not only depend on the noise type, but also on the global SNR. This is an important difference with many other approaches where the shape of the speech spectrum is typically independent of global SNR, e.g., [5, 6, 20, 10]. It can be observed that for higher global SNRs, the processed speech spectrum tends to shape like the noise. Shaping the speech as the noise was also proposed in, e.g., [3, 20], however, this is not optimal for lower global SNRs according to our cost function, Eq. (7). This approach would also result in an equal SNR as a function of frequency, which is clearly not the case when observing the middle-row plots in Figure 3. Moreover, making the speech spectrum constant as in [5], seems only appropriate in the case

of white noise for the highest global SNRs (5 dB SNR).

Another property of the proposed method which is revealed from the Figures, is the fact that certain frequency bands are set to zero when lowering the global SNRs. This makes sense, since the SII assumes that all bands with SNR below -15 dB do not contribute to intelligibility anymore and can therefore be discarded. Indeed, observing the distributions of the SNRs over frequency in the second row in Figure 3 do not show any SNRs below -15 dB. This explains why our method will typically improve over other state-of-the-art methods mainly for lower global SNRs [12].

In the bottom-row plots the filter gains are shown, i.e., α_i. It is clear that most of the filters have a high-pass characteristic, however, typically low frequencies are also preserved. A real high-pass filter is only observed for certain noise types at very low SNRs, e.g., low-pass noise at a global SNR of -10 dB and lower. These filter shapes are in line with the results observed in [7] where a high-pass filter and a format-equalization filter also have a positive effect on speech intelligibility. Only for the high-pass filter noise, the filter shows a low-pass characteristic for lower global SNRs. The fact that this type of filter was never proposed in literature is probably due to the fact that a high-pass shape of the noise is very unnatural and typically not used for evaluation.

4. Experimental Evaluation

The listening experiment is performed within the Hurricane Challenge which included 20 different algorithms [15, 14]. A brief summary of the experimental setup and results only for our method will be given. In total, 127 native English speaking subjects listened to sentences from the Harvard corpus [21]. The Harvard corpus contains sentences such as "the salt breeze came across from the sea", spoken by a male British English talker. Two noise types were used at 3 different SNRs including stationary SSN and a highly non-stationary competing female speaker (CS). The SNRs for SSN and CS were [-9, -4, 1] and [-21, -14, -7] dB, respectively. In total, each listener evaluated 9 sentences for each of the 20 algorithms (including the proposed method), which gives a total of 180 sentences per participant. SNRs and noise types are balanced over all listeners (see [14] for more details).

In addition to the listening test, SII scores were also obtained for the same set of conditions as used in the listening test. The long-term average speech and noise spectra were estimated based on 180 Harvard sentences [21]. The SII-scores were calculated as in Eq. (5).

5. Results

Listening test results are shown in the top row plots of Figure 4. The bottom two plots show the predictions of the SII.

Both listening test results and SII predictions show an improvement in speech intelligibility for SSN for all three SNRs. The largest benefit was measured with the lowest SNR for SSN, where intelligibility improved from 17.3% to 50.6% words correct. For the highest SNR the improvement is smaller, which is probably caused by ceiling effects, i.e., the fact that speech intelligibility is upper bounded by 100%. Since SII-scores are not close to 1 yet, this ceiling effect is not observed with the objective scores. These scores are in line with the results from [12], where large improvements were found with a female speaker rather than a male speaker.

The difference between SII-predictions and listening experiment results for the competing speaker is somewhat remarkable. Although it is known that the SII is not reliable for fluctuating noise sources [16], it actually turns out that improving the SII for these non-stationary cases may even *decrease* intelligibility. Overall the decrease in intelligibility is the largest for the lowest SNR conditions where scores dropped from 24.8% to 10.2%. To overcome this, the proposed method may be modified by optimizing for the extended SII as proposed in [16]. Here the SII is calculated in short-time frames rather than on long-term average spectra which improves prediction results for non-stationary noise. The mathematical framework presented in this paper still holds when deriving an optimal filter for such a time-varying SII.

In addition to use a time-varying filter, results may be further improvement by combining the proposed method with other types of processing, e.g., dynamic range compression [22].

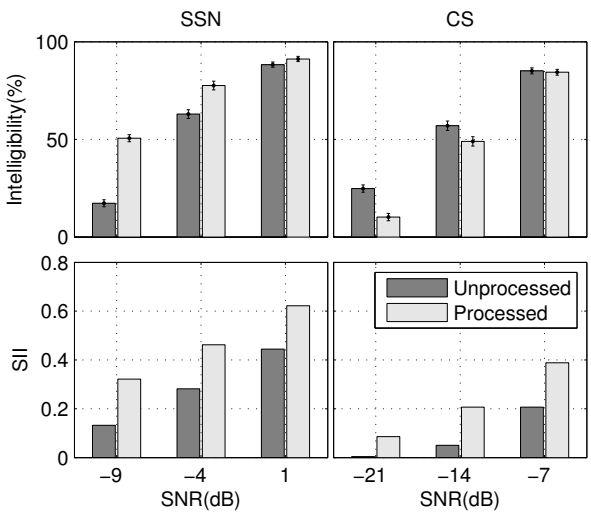

Figure 4: Listening test results (top) and SII predictions (bottom) for speech shaped noise (SSN) (left) and a competing speaker (CS) (right).

6. Conclusions

A linear time-invariant filter was proposed to optimize the intelligibility of speech in noise for the near-end listener without affecting the global SNR. This was accomplished by redistributing the speech energy over frequency such that an approximation of the speech intelligibility index (SII) was maximized. The resulting filter is dependent both on noise spectrum, speech spectrum and global SNR and, in general, has a high-pass characteristic. In contrast to existing methods, the proposed filter sets certain frequency bands to zero when the per-band SNRs are so low that they do not contribute to intelligibility anymore. Intelligibility test results and SII predictions show a large intelligibility improvement for speech mixed with speech-shaped noise. However, despite improved SII predictions it was also found that the method does not increase intelligibility when speech is corrupted by a competing speaker. This is caused by the fact that the SII is not a reliable intelligibility predictor for fluctuating noise sources. MATLAB code is provided at http://www.ceestaal.nl/.

7. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC, Boca Raton, FL, 2007.

[2] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachbericht-Sprachkommunikation*, 2010.

[3] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power constraining," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[4] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2010.

[5] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Am.*, vol. 43, no. 1, pp. 81–86, 1968.

[6] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 277 – 282, 1976.

[7] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 280–285, 2010.

[8] M. Skowronski and J. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[9] P. Chanda and S. Park, "Speech intelligibility enhancement using tunable equalization filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 613–616.

[10] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression," in *Proc. EUSIPCO*, 2012, pp. 2075–2079.

[11] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4061 – 4064.

[12] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225 –228, 2013.

[13] ANSI, "Methods for calculation of the speech intelligibility index," *S3.5-1997, (American National Standards Institute, New York)*, 1997.

[14] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge," in *Proc. Interspeech*, 2013.

[15] M. Cooke, C. Mayo, B. Sauert, Y. Stylianou, C. Valentini-Botinhao, and Y. Tang, "The hurricane challenge," 2012. [Online]. Available: http://www.listening-talker.org/hurricane/

[16] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[17] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Appl. Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[18] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4266–4269.

[19] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Pr, 2004.

[20] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, 2010, pp. 1636–1639.

[21] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[22] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3236–3245, 2009.