NEAR END LISTENING ENHANCEMENT CONSIDERING THERMAL LIMIT OF MOBILE PHONE LOUDSPEAKERS

Bastian Sauert and Peter Vary

Institute of Communication Systems and Data Processing (ind),
RWTH Aachen University, 52056 Aachen, Germany
{sauert,vary}@ind.rwth-aachen.de

Abstract: In mobile telephony, listening enhancement is desired by the near end listener who perceives not only the clean far end speech but also ambient background noise. A typical scenario is mobile telephony in acoustical background noise such as traffic or babble noise. During continuous playback the thermal load of the small loudspeakers is a major limitation.

In this contribution, we adapt two previous approaches to this constraint. Furthermore, the impact of a new speech PSD estimator and state-of-the-art noise PSD tracking algorithms on the system performance is evaluated.

1 Introduction

Mobile telephony is often conducted in the presence of acoustical background noise such as traffic or babble noise. In this situation, the *near end* listener perceives a mixture of the clean *far end* (downlink) speech and the acoustical background noise from the *near end* and thus experiences an increased listening effort and a possibly reduced speech intelligibility. As the noise signal cannot be influenced, a reasonable approach to improve intelligibility is to manipulate the received far end speech signal in dependence of the near end background noise, which we call near end listening enhancement (NELE). This requires an estimate of the instantaneous far end speech power spectral density (PSD) as well as the momentary near end noise PSD, where the latter can be obtained from the mobile phone's microphone signal.

In [11], we derived a NELE algorithm which maximizes the Speech Intelligibility Index (SII) and thus speech intelligibility by frequency selective increase of the speech signal power. Time-domain filtering with the filter coefficients adapted in the frequency domain was performed by means of a frequency warped filter-bank equalizer. This allows for processing with approximately Bark-scaled spectral resolution according to the human auditory system and low signal delay.

In [12], we considered applications where the loudspeaker signal power is constrained to the power of the original signal. A recursive closed-form solution optimization of the spectral speech signal power allocation is derived which maximized the SII under this constraint.

However, for small loudspeakers as used in mobile phones the thermal load during continuous playback is one major limitation, e. g., [10]. Therefore, most mobile phone applications limit the overall power of the loudspeaker signal to a constant maximum power instead of the power of the original signal.

In this contribution, we investigate the approach presented in [12] under this new constraint and compare it with the approach of [11] combined with a frequency independent weight limitation. Furthermore, the impact of a new speech PSD estimator and state-of-the-art noise PSD tracking algorithms on the system performance is evaluated.

2 System Overview

In this contribution, near end listening enhancement is realized by means of a warped filter-bank equalizer (FBE) with a system framework described in this section and depicted in Figure 1. In essence, a time-domain filtering is performed with coefficients calculated in the frequency-domain. In contrast to the discrete Fourier transform (DFT) analysis-synthesis filter-bank, which is widely used for speech enhancement, this structure allows for an efficient processing with approximately Bark-scaled spectral resolution and low signal delay.

The (clean) far end speech signal $s^{\text{in}}(k)$ with time index k from the downlink is transformed to subband signals $S_i^{\text{in}}(\kappa)$ with sub-sampled time index κ and subband index i by means of a warped DFT analysis filter-bank with downsampling. Next, the short-term PSD of the speech signal $\hat{\Phi}_{ss,i}^{\text{in}}(\kappa)$ is estimated as described in Section 2.1.

On the other hand, the near end microphone signal $y(k)$, which is a mixture of the near end noise signal $r(k)$ and possibly an interfering near end speech signal, is analogously transformed to subband signals $Y_i(\kappa)$. Since the interfering near end speech signal should not be considered during NELE, the near end noise PSD $\hat{\Phi}_{rr,i}(\kappa)$ is estimated as shown in Section 2.2.

The spectral weights $W_i(\kappa)$ are calculated based on both PSD estimates and then limited to prevent damage of the listener's ear and the sound equipment. The limited weights are transformed to coefficients of a time-domain filter, which is applied to the far end speech signal $s^{\text{in}}(k)$.

It should be noted that only a rough overview of the FBE and its parametrization is given here. These aspects are treated in more detail in [7, 13] and [12].

2.1 Speech PSD Estimation

In [11, 12], the short-term PSD of the far end speech signal was estimated as the *recursive* average of the normalized power of the subband signals with a certain time constant, which was adapted to the non-uniform time-frequency resolution of the used warped filter-bank. This time constant needed to be chosen quite high to overcome typical speech pauses without losing the estimate. Besides still failing for long pauses, this also has the disadvantage of a slow adaptation to changes in intensity and spectral envelope of the far end signal.

In this contribution, a still rather simple approach is used as an alternative which basically takes the *moving* average of the normalized power of the subband signals with a look-back over voice activity segments (in sum) of length τ_s. More specifically, the algorithm considers only those

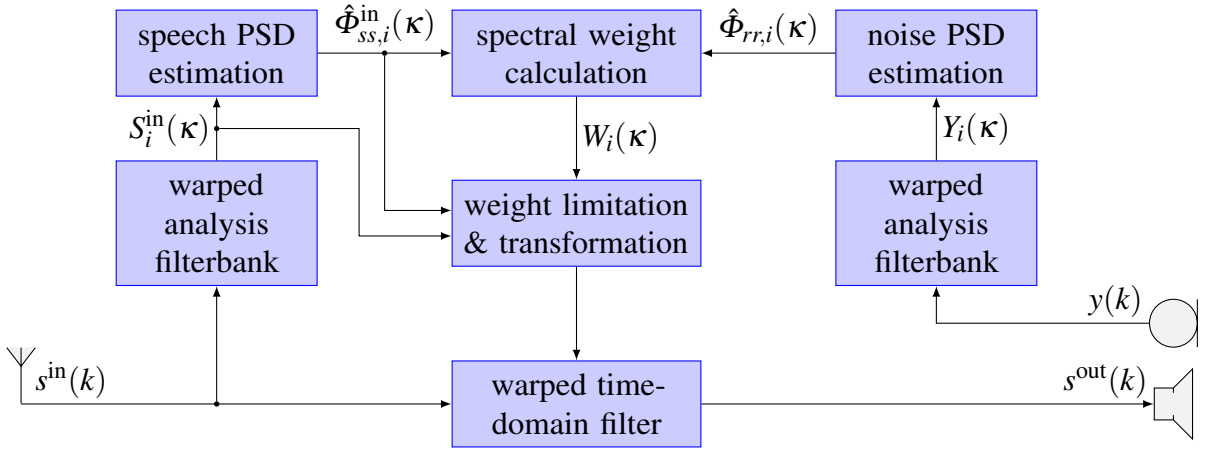

Figure 1: System for NELE with time index k, sub-sampled time index κ, and subband index i.

speech signal segments with voice activity according to the voice activity detector of the G.729 codec [5]. It then calculates the short-term PSD estimate $\hat{\Phi}^{\text{in}}_{ss,i}(\kappa)$ as the arithmetic mean of the squared, normalized magnitudes $|S^{\text{in}}_i(\kappa)|^2$ of the subband signals during the preceding τ_s seconds of these segments.

The duration τ_s determines the memory of the speech PSD estimator. Too small values result in a high variance of the estimate and, thus, a fast and unpleasant fluctuation of the spectral weights. With a too large τ_s the system can only slowly adapt to changes in intensity and spectral envelope of the far end signal. Setting $\tau_s = 2\,\text{s}$ turned out to be a reasonable compromise and is used in the following.

2.2 Noise PSD Estimation

For estimation of the short-term PSD of the near end noise signal $\hat{\Phi}_{rr,i}(\kappa)$, the same recursive average estimator as for the short-term speech PSD estimate was used in [11, 12] but with a shorter time constant. However, as this estimator interprets near end speech in double-talk situations as noise, it is not suitable for real-world applications.

Therefore, two noise estimation algorithms from literature are examined in this work:

1. the Minimum Statistics algorithm [8, 9] in the implementation of [2] and
2. a minimum mean-square error (MMSE) based noise PSD tracking algorithm [3] in an implementation provided by the authors [4].

Quite remarkably, both algorithms perform out-of-the-box well with the non-uniform analysis filterbank of the FBE. In general, both algorithms are comparable in terms of average noise PSD estimate for the most quasi-stationary noise signals. However, the MMSE based algorithm tends to track non-stationary as well as speech babble noise better and faster than the Minimum Statistics algorithm. Furthermore, it seems to cope better with interfering speech. Unless stated otherwise, the MMSE based algorithm is used in the following.

3 Near End Listening Enhancement

In this section, NELE algorithms are discussed which try to improve the speech intelligibility using the Speech Intelligibility Index (SII) as optimization criterion.

3.1 Speech Intelligibility Index (SII)

The SII [1] is a standardized objective measure which correlates with the intelligibility of speech under a variety of adverse listening conditions. It is based on the equivalent speech spectrum level[1] E_i as well as the equivalent noise spectrum level[1] N_i in each contributing subband i, both specified in dB. The spectrum level is basically the power average over time in each subband with reference pressure 20 µPa differentiated with respect to the bandwidth of the subband. It can be approximated by the power average over time in each subband divided by its bandwidth [1].

The disturbance spectrum level D_i appropriately accounts for the masking of speech, which also includes within-band masking and out-of-band masking (spread of masking) produced by the noise. For the assumption of significant background noise levels, the disturbance spectrum level D_i only depends on the noise spectrum level N_i.

[1] The equivalent spectrum level is defined as the spectrum level measured at the point corresponding to the center of the listener's head, with the listener absent, under the reference communication situation [1].
In the following, the term "equivalent" is omitted for the sake of clarity.

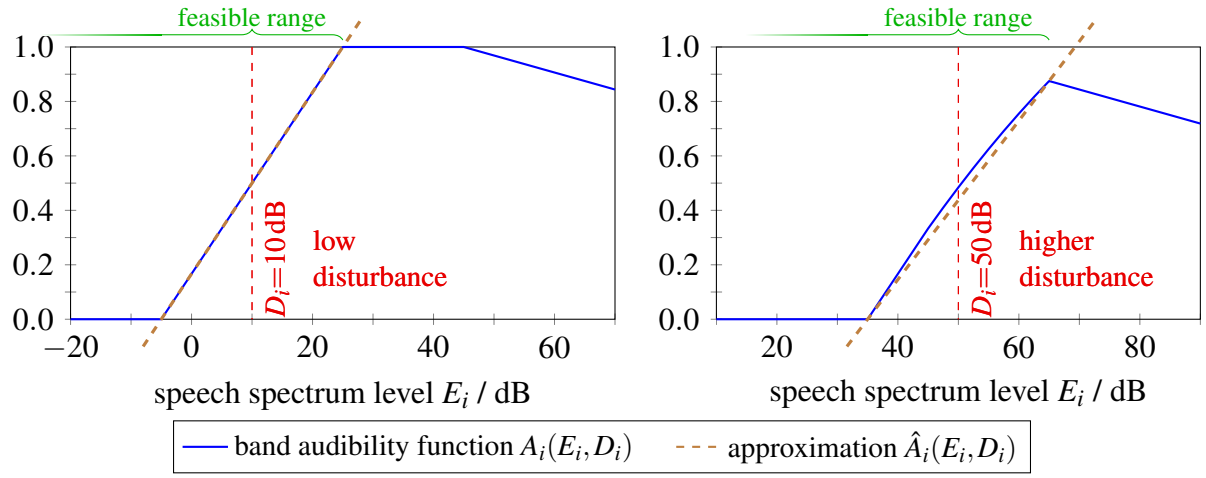

Figure 2: Exemplary plot of band audibility function for low as well as higher disturbance case.

Finally, the Speech Intelligibility Index S is calculated as weighted sum of the band audibility function $A_i(E_i,D_i)$

$$S = \sum_{i=1}^{i_{\max}} I_i \cdot A_i(E_i,D_i) \tag{1}$$

with the number of subbands $i_{\max}$. The band importance function I_i [1, Table 1] characterizes the relative significance of the subband to speech intelligibility.

The band audibility function $A_i(E_i,D_i)$ specifies the effective proportion of the speech dynamic range within the subband that contributes to speech intelligibility and its characteristics are sketched in Figure 2 for a low as well as a high disturbance scenario.

3.2 Concept

Even though the SII works on long-term power averages, e. g., over a whole utterance, the presented algorithms calculate time-varying spectral weights based on the short-term PSDs using the unchanged SII calculation rules as criterion. The basic idea of these algorithms is to first determine an optimum speech spectrum level $E_i^{\text{opt}}(\kappa)$ which maximizes the SII under consideration of the current disturbance spectrum level $D_i(\kappa)$:

$$\underline{E}^{\text{opt}}(\kappa) = \arg\max_{\underline{E}} \sum_{i=1}^{i_{\max}} I_i \cdot A_i\big(E_i, D_i(\kappa)\big)\,, \tag{2}$$

where $\underline{E}$ denotes the vector of all contributing E_i, subject to

$$\sum_{i=1}^{i_{\max}} \Delta f_i \cdot 10^{E_i/10} \overset{!}{\leq} P_{\max}\,. \tag{3}$$

This constraint limits the short-term power of the loudspeaker signal to a constant maximum power $P_{\max}$, which could be derived from the specification of the loudspeaker during design of the mobile phone. As mentioned in the introduction, this is suitable for most mobile phone applications as the thermal load is one major limitation for small loudspeakers.

Next, the spectral weights are calculated which are necessary to achieve this optimum speech spectrum level at the ear of the listener. Assuming sufficiently stationary spectral weights $W_i(\kappa)$,

the short-term PSD estimate of the enhanced speech signal $S_i^{\text{out}}(\kappa) = W_i(\kappa) \cdot S_i^{\text{in}}(\kappa)$ can be expressed as $\hat{\Phi}_{ss,i}^{\text{out}}(\kappa) = W_i^2(\kappa) \cdot \hat{\Phi}_{ss,i}^{\text{in}}(\kappa)$, which leads to the spectral weights

$$W_i(\kappa) = 10^{\left[E_i^{\text{opt}}(\kappa) - E_i^{\text{in}}(\kappa)\right]/20}, \tag{4}$$

where $E_i^{\text{in}}(\kappa)$ denotes the current input speech spectrum level calculated from $\hat{\Phi}_{ss,i}^{\text{in}}(\kappa)$.

3.3 Extended Approach of [11]: Limited Unconstrained Optimization

In this section, the SII should first be maximized with the only constraint that the speech signal should not be attenuated in a noise-free environment. The power constraint (3) is afterwards accomplished by a frequency independent weight limitation.

As can be seen in Figure 2 and as derived in [11], the optimum speech spectrum level

$$E_i^{\text{opt}}(\kappa) = \max\left\{D_i(\kappa) + 15\,\text{dB}, E_i^{\text{in}}(\kappa)\right\} \tag{5}$$

fulfills the above requirements. With (4), this leads to the spectral weights

$$W_i(\kappa) = \max\left\{10^{\left[D_i(\kappa) + 15\,\text{dB} - E_i^{\text{in}}(\kappa)\right]/20}, 1\right\}, \tag{6}$$

which are then limited to satisfy the maximum overall power $P_{\max}$:

$$W_i'(\kappa) = \begin{cases} W_i(\kappa) & \text{if } \sum_{i=1}^{i_{\max}} W_i^2(\kappa) \cdot \hat{\Phi}_{ss,i}^{\text{in}}(\kappa) \le P_{\max} \\ \sqrt{\dfrac{P_{\max}}{\sum_{i=1}^{i_{\max}} W_i^2(\kappa) \cdot \hat{\Phi}_{ss,i}^{\text{in}}(\kappa)}} \cdot W_i(\kappa) & \text{otherwise.} \end{cases} \tag{7}$$

3.4 Approach of [12]: Constrained Optimization

If $E_i = D_i + 15\,\text{dB}$ fulfills the constraint (3), the maximum SII can be reached [12]. If not, all power must be used to maximize the SII. In this case, the solution lies within the feasible range $E_i^{\text{opt}} \le D_i + 15\,\text{dB}$ and the inequality constraint (3) becomes an equality constraint

In order to facilitate the closed-form optimization, the band audibility function is approximated by a linear function $\hat{A}_i(E_i, D_i)$ [12] as depicted in Figure 2. In the most relevant range $D_i - 15\,\text{dB} \le E_i \le D_i + 15\,\text{dB}$, the approximation $\hat{A}_i(E_i, D_i)$ is identical to $A_i(E_i, D_i)$ in the low disturbance case and slightly underestimates $A_i(E_i, D_i)$ in the higher disturbance case.

The equality constrained nonlinear multivariate maximization problem (2) and (3) can be solved using the methods of Lagrange multipliers. This finally leads to the closed-form solution [12]

$$E_i^{(1)} = 10\log\left\{\frac{\gamma_i}{\sum_{\lambda=1}^{i_{\max}} \gamma_\lambda} \cdot \frac{P_{\max}}{\Delta f_i}\right\}. \tag{8}$$

where γ_i is the gradient of the linear approximation $\hat{A}_i(E_i, D_i)$. This solution might, however, fall outside the feasible range. Therefore, further steps $\nu = 2, 3, \ldots$ are necessary, where the preceding solution $E_i^{(\nu-1)}$ is limited to $D_i + 15\,\text{dB}$ and the closed-form solution (8) is repeated recursively until all subbands fulfill $E_i^{(\nu)} \le D_i + 15\,\text{dB}$, leading after $\nu_{\max} \le i_{\max}$ recursion steps to the final solution $E_i^{\text{opt}} = E_i^{(\nu_{\max})}$.

4 Instrumental Evaluation

4.1 Simulation Environment

The performance of the presented algorithms is evaluated in terms of the SII using the so-called critical band procedure [1] for every speech file of the TIMIT database, in total 5.4 hours, disturbed by speech babble (`babble`), white noise (`white`), or car interior noise (`volvo`) from the NOISEX-92 database at a sampling rate of $f_s = 8\,\text{kHz}$. Afterwards, the average SII over all speech files is taken. Good communication systems have an SII of 0.75 or better while the SII of poor communication systems is below 0.45 [1].

Prior to processing, each speech file is scaled to match an overall active speech level [6] corresponding to a sound pressure level of 62.35 dB as specified in [1] for normal voice effort. The desired input signal-to-noise ratios (SNRs) ranging from -40 dB to 30 dB in steps of 2.5 dB are achieved by adjusting the overall active speech level [6] of the noise file in relation to a sound pressure level of 62.35 dB.

The algorithms are evaluated with a maximum output audio power $P_{\text{max}} = 94\,\text{dBSPL}$, a value chosen in accordance with [10]. For the comparison of the different speech and noise PSD estimators, the constrained optimization algorithm as of Section 3.4 was used.

4.2 Results

Figure 3 shows that the limited unconstrained optimization approach of Section 3.3 as well as the constrained optimization approach of Section 3.4 perform identical if the power constraint is not active. This is to be expected as both algorithms are basically unconstrained in that case and lead to the same spectral weights. If the constraint is active, the solution which is optimized for this case has a better average SII of about 0.05 or reaches the same SII at an up to 4 dB lower SNR.

Concerning the speech PSD estimators, the recursive and the moving average estimators basically lead to the same performance as can be seen in Figure 4. Since the moving average estimator has better adaptation properties, it should be preferred.

For white and car interior noise all evaluated noise PSD estimators perform almost identical as shown in Figure 5. In constrast, for speech babble and mid-range SNRs, the MMSE based algorithm results in much better average SIIs than the Minimum Statistics algorithm. The recursive average estimator yields results similar to the MMSE based algorithm, but can not cope with double-talk situations, as explained above.

5 Conclusions

In this contribution, two SII based near end listening enhancement algorithms are compared under the side condition, that the short-term power of the loudspeaker signal is limited to a constant maximum power: a limited unconstrained optimization approach, partly presented in [11], and a constrained optimization approach, presented in [12]. The side condition accounts for the fact, that the thermal load during continuous playback is one major limitation for small loudspeakers.

The instrumental evaluation by means of the average SII shows the lead of the constrained optimization if the constraint is active and an identical performance after processing with both algorithms otherwise.

Furthermore, it is found, that both presented speech PSD estimators basically lead to the same

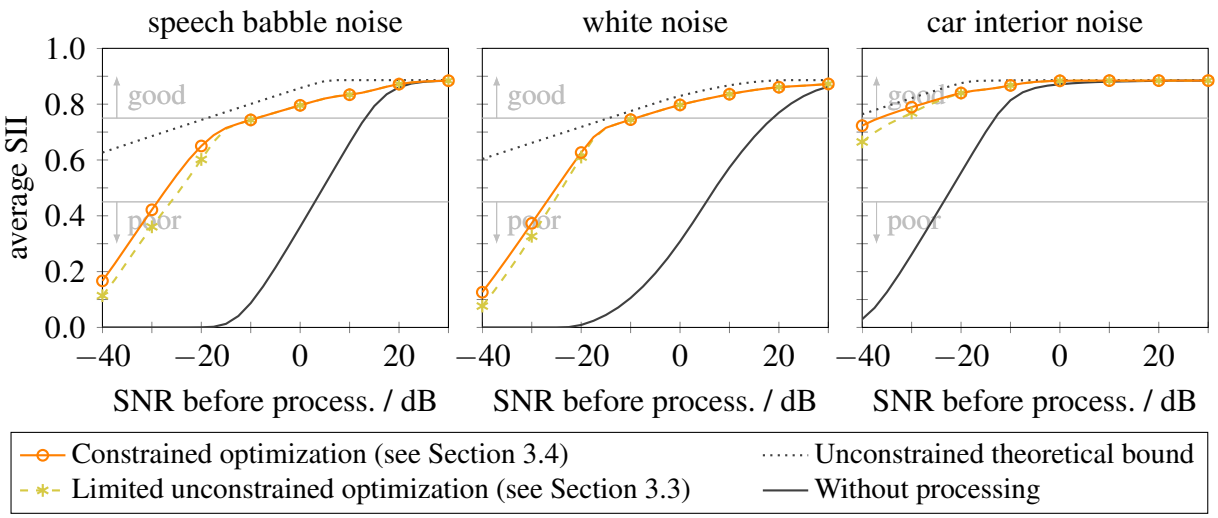

Figure 3: Comparison of constrained optimization (see Section 3.4) and limited unconstrained optimization (see Section 3.3).

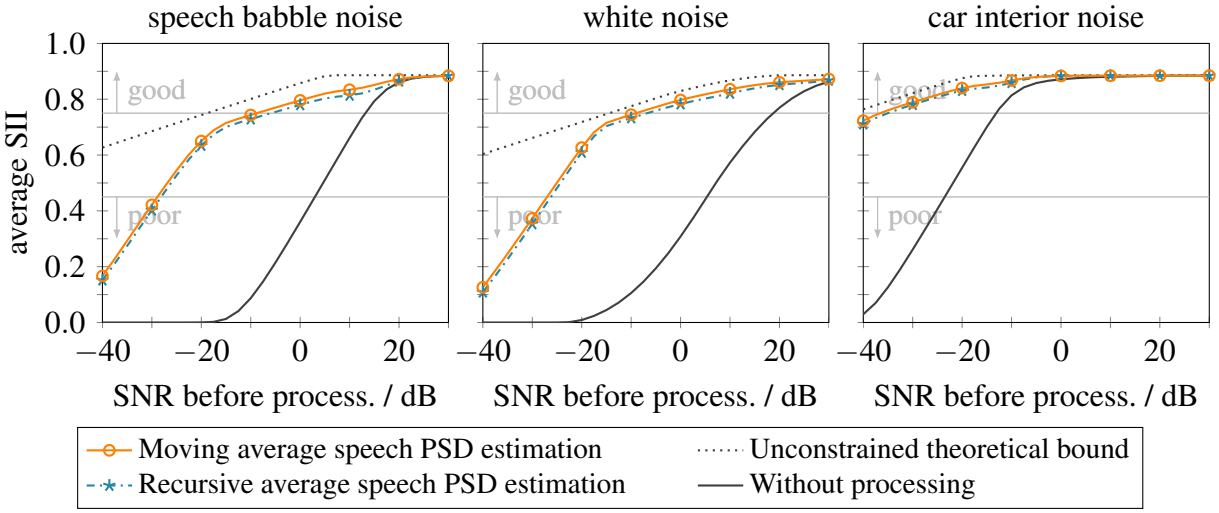

Figure 4: Comparison of speech PSD estimation algorithms (see Section 2.1).

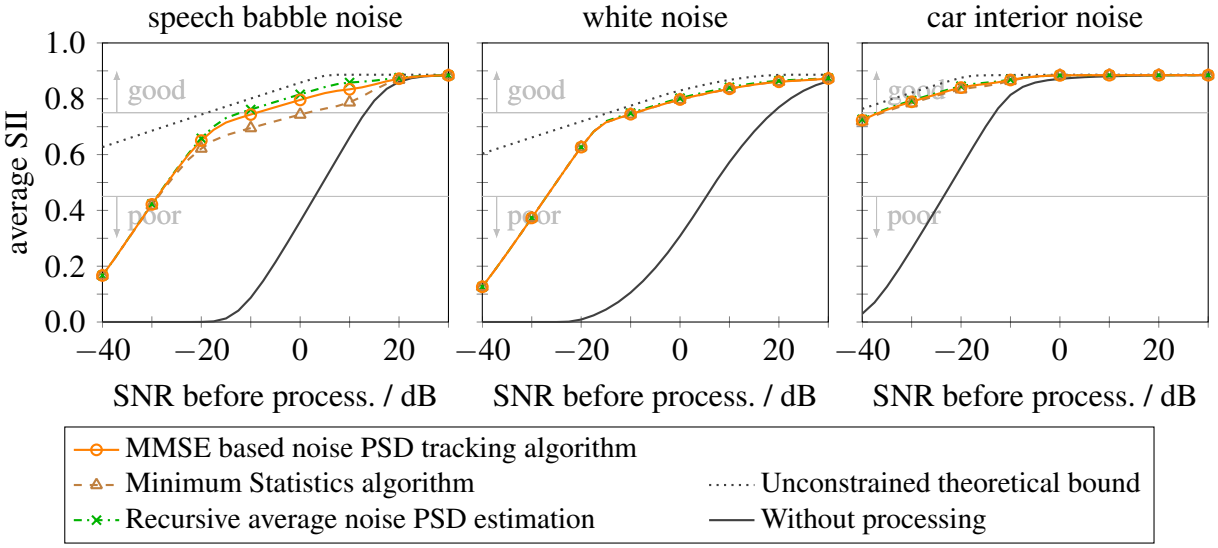

Figure 5: Comparison of noise PSD estimation algorithms (see Section 2.2).

performance. Concerning the noise PSD trackers, the MMSE based algorithm results in much better average SIIs than the Minimum Statistics algorithm for speech babble and mid-range SNRs. For white and car interior noise both perform almost identical.

References

[1] ANSI S3.5-1997. *Methods for the Calculation of the Speech Intelligibility Index*. American National Standards Institute, 1997.

[2] Mike Brookes. *VOICEBOX. Speech Processing Toolbox for MATLAB*. Department of Electrical & Electronic Engineering, Imperial College. Feb. 2, 2011. URL: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html` (visited on 02/03/2011).

[3] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "MMSE based noise PSD tracking with low complexity". In: *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. (Dallas, Texas, USA, Mar. 14–19, 2010). IEEE. Mar. 2010, pp. 4266–4269. ISBN: 978-1-4244-4296-6. DOI: 10.1109/ICASSP.2010.5495680.

[4] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. *MMSE based noise PSD tracking algorithm*. Version V1. Signal & Information Processing Lab, Delft University of Technology. Apr. 15, 2010. URL: `http://siplab.tudelft.nl/content/mmse-based-noise-psd-tracking-algorithm` (visited on 02/03/2011).

[5] ITU-T Recommendation G.729. *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*. Version 01/2007. International Telecommunication Union, Jan. 2007.

[6] ITU-T Recommendation P.56. *Objective Measurement of Active Speech Level*. Version 03/93. International Telecommunication Union, Mar. 1993.

[7] Heinrich W. Löllmann and Peter Vary. "Uniform and Warped Low Delay Filter-Banks for Speech Enhancement". In: *Speech Communication* 49 (7-8 2007): *Special Issue on Speech Enhancement*, pp. 574–587. ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.04.009.

[8] Rainer Martin. "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics". In: *IEEE Transactions on Speech and Audio Processing* 9.5 (July 2001), pp. 504–512. ISSN: 1063-6676. DOI: 10.1109/89.928915.

[9] Rainer Martin. "Bias compensation methods for minimum statistics noise power spectral density estimation". In: *Signal Processing* 86.6 (June 2006): *Special Issue on Applied Speech and Audio Processing (dedicated to Prof. Hänsler)*. Ed. by H. Puder and G. Schmidt, pp. 1215–1229. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2005.07.037.

[10] NXP Semiconductors. *RA 8x12x2 Receiver*. Specification. Version G. Order No. 2403 260 00031. NXP Semiconductors Netherlands B.V., Mar. 30, 2010.

[11] Bastian Sauert and Peter Vary. "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index". In: *Proc. of European Signal Processing Conf. (EUSIPCO)*. (Glasgow, Scotland, Aug. 24–28, 2009). Vol. 17. European Association for Signal Processing (EURASIP). New York, NY: Hindawi Publ., Aug. 2009, pp. 1844–1848.

[12] Bastian Sauert and Peter Vary. "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement". In: *Proc. of ITG-Fachtagung Sprachkommunikation*. (Bochum, Germany, Oct. 6–8, 2010). Vol. 9. Berlin [u.a.]: VDE-Verlag, Oct. 2010. ISBN: 978-3-8007-3300-2.

[13] Peter Vary. "An Adaptive Filterbank Equalizer for Speech Enhancement". In: *Signal Processing* 86.6 (June 2006): *Special Issue on Applied Speech and Audio Processing (dedicated to Prof. Hänsler)*. Ed. by H. Puder and G. Schmidt, pp. 1206–1214. DOI: 10.1016/j.sigpro.2005.06.020.