Near-End Listening Enhancement in the Presence of Bandpass Noises

Bastian Sauert[1] *and Peter Vary*

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

Abstract

When using a mobile phone in acoustical background noise, the near-end listener perceives not only the (clean) far-end speech but also the ambient noise and thus experiences an increased listening effort and a possibly reduced speech intelligibility. Near-end listening enhancement processes the received far-end speech signal depending on the near-end background noise to improve intelligibility. However, in mobile phones it is often not possible to increase the audio power.

In a previous contribution, the authors developed a recursive closed-form solution, which maximizes the Speech Intelligibility Index (SII) under the constraint of an unchanged average power of the audio signal.

This solution, however, shows in bandpass noise environments a disadvantageous narrow bandpass characteristic of the processed speech even though the SII is optimal. Therefore, in this contribution, we analyze this algorithm and propose a new spectral weighting rule which prevents the narrow bandpass effect with only a marginal reduction in SII.

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 *far-end* speech signal depending on the *near-end* background noise, which we call near-end listening enhancement (NELE).

Most speech modification algorithms proposed so far are independent of the actual noise environment. Only recently, an optimization using an estimate of the noise context based on a spectro-temporal perceptual distortion measure was presented in [1].

In [2], we derived a NELE algorithm which maximizes the Speech Intelligibility Index (SII) [3] by frequency selective increase of the speech signal power.

In [4], we considered applications where the loudspeaker signal power is constrained to the power of the original signal. A recursive closed-form optimization of the spectral speech signal power allocation was derived which maximizes the SII under this constraint.

Although, this algorithm results in an optimized SII of the output speech, the resulting optimum speech spectrum level can, e. g., in the presence of bandpass noises, lead to frequency weights which exhibit a narrow bandpass characteristic and thus a disadvantageous or even destructive effect on listening experience. These effects, which are not covered by the SII, can be observed in the rating of a speech-based revised Speech Transmission Index ($\mathrm{STI_{sr}}$).

In this contribution, we analyze the algorithm presented in [4] and investigate a novel weighting rule which prevents these disadvantageous weights and thus yields dramatically better $\mathrm{STI_{sr}}$ ratings with only a small reduction in SII.

[1]This work was funded by Intel Mobile Communication Group.

2 System Overview

As in [2, 4–6], near-end listening enhancement is realized by means of a warped filterbank equalizer [7, 8]. This structure performs a time-domain filtering with coefficients calculated in the frequency-domain and allows for an efficient processing with approximately Bark-scaled spectral resolution according to the human auditory system and low signal delay.

The received (clean) far-end speech signal $s^{\mathrm{in}}(k)$ with time index k is transformed to subband signals $S_i^{\mathrm{in}}(\kappa)$ with subsampled time index κ and subband index i by means of a warped discrete Fourier transform analysis filterbank with 34 subbands at a sampling rate of 8 kHz. The short-term power spectral density (PSD) of the speech signal $\hat{\Phi}_{ss,i}^{\mathrm{in}}(\kappa)$ is estimated by a moving average of the normalized power of the subband signals over the last 2 seconds of voice activity [6].

The near-end microphone signal $y(k)$, which is a mixture of the near-end noise signal $r(k)$ and possibly, e. g., during double-talk, a near-end speech signal, is analogously transformed to subband signals $Y_i(\kappa)$. Since the near-end speech signal should not be considered for NELE, the near-end noise PSD $\hat{\Phi}_{rr,i}(\kappa)$ is estimated with the minimum mean-square error based noise PSD tracking algorithm of [9].

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 warped time-domain filter, which is applied to the far-end speech signal $s^{\mathrm{in}}(k)$.

It should be noted that only a rough overview of the system and its parametrization is given here. These aspects are treated in more detail in [5–7].

3 Near-End Listening Enhancement

3.1 Speech Intelligibility Index (SII)

The SII [3] is a standardized objective measure which correlates with the intelligibility of speech under a variety of adverse listening conditions. Note, that a detailed discussion of the calculation rules of the SII is given in [2].

In short, the SII is based on the equivalent speech spectrum level[2] E_i as well as the equivalent disturbance spectrum level[2] D_i, which accounts for the noise as well as the masking of the speech. Given E_i and D_i, the Speech Intelligibility Index S is calculated as a weighted sum of the band audibility function $A_i(E_i, D_i)$ over all contributing subbands i_{f} to i_{l}:

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

where the band importance function I_i [3] characterizes the relative significance of each 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. Its characteristics are sketched in Figure 1 for a low as well as a high disturbance scenario.

[2]The term "equivalent" is omitted in the following for the sake of clarity.

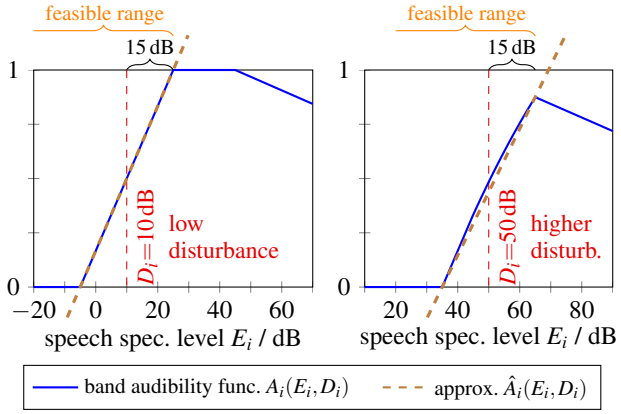

Figure 1: Exemplary plot of band audibility function for low as well as higher disturbance case.

3.2 Concept

Although the SII works on long-term power averages, e. g., over a whole utterance, the time-varying spectral weights of the following algorithms are calculated 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 S under consideration of the current disturbance spectrum level $D_i(\kappa)$:

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

subject to

$$\sum_{i=i_\text{f}}^{i_\text{l}} \Phi_{ss,i} = \sum_{i=i_\text{f}}^{i_\text{l}} \Delta f_i \cdot 10^{E_i/10} \overset{!}{\leq} P_{\max}(\kappa) \coloneqq \sum_{i=i_\text{f}}^{i_\text{l}} \hat{\Phi}_{ss,i}^{\text{in}}(\kappa)\,, \tag{3}$$

where $\underline{E}$ denotes the vector of all contributing speech spectrum level $E_i = 10\log\{\Phi_{ss,i}/\Delta f_i\}$ and Δf_i is the frequency bandwidth of the i-th subband. The constraint limits the short-term power of the loudspeaker signal to a maximum power $P_{\max}(\kappa)$, which is the power of the original signal.

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

$$\hat{\Phi}_{ss,i}^{\text{out}}(\kappa) = W_i^2(\kappa) \cdot \hat{\Phi}_{ss,i}^{\text{in}}(\kappa)\,, \tag{4}$$

which finally lead to the spectral weights

$$W_i(\kappa) = 10^{\left[E_i^{\text{opt}}(\kappa) - E_i^{\text{in}}(\kappa)\right]/20}\,, \tag{5}$$

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

3.3 Recursive Closed-Form Optimization [4]

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

In order to facilitate the envisaged closed-form optimization, the band audibility function is approximated by a linear function $\hat{A}_i(E_i, D_i)$ as depicted in Figure 1.

Then, the equality constrained nonlinear multivariate maximization problem (2) and (3) can be solved using the methods of Lagrange multipliers. As shown in [4], this finally leads to the closed-form solution

$$E_i^{(1)} = 10\log\left\{\frac{\lambda_i}{\sum_{\mu=1}^{i_\text{l}} \lambda_\mu} \cdot \frac{P_{\max}(\kappa)}{\Delta f_i}\right\}. \tag{6}$$

with the gradient λ_i 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 (6) is repeated recursively until all subbands fulfill $E_i^{(\nu)} \leq D_i + 15\,\text{dB}$, leading after $\nu_{\max} \leq i_\text{l} - i_\text{f} + 1$ recursion steps to the final solution $E_i^{\text{opt}} = E_i^{(\nu_{\max})}$.

Please refer to [4] for further details of this algorithm.

3.4 Analysis

In this section, the sketched optimization scheme is studied concerning the characteristics of the resulting spectral weights for different signal-to-noise ratios (SNRs).

3.4.1 Low SNR: Bandpass Characteristic

For a sufficiently low SNR, $D_i(\kappa) + 15\,\text{dB} \geq E_i^{(1)}$ is fulfilled in all subbands and recursion stops after $\nu_{\max} = 1$ step with $E_i^{\text{opt}}(\kappa) = E_i^{(1)}$. Accordingly, the optimum PSD $\Phi_{ss}^{\text{opt}}(\kappa)$ of the output speech results in

$$\Phi_{ss}^{\text{opt}}(\kappa) = \frac{\lambda_i}{\sum_{\mu=i_\text{f}}^{i_\text{l}} \lambda_\mu} \cdot P_{\max}(\kappa) \tag{7}$$

and it further follows for the spectral weights that

$$W_i(\kappa) = \sqrt{\frac{\lambda_i}{\sum_{\mu=i_\text{f}}^{i_\text{l}} \lambda_\mu} \cdot \frac{\sum_{\mu=i_\text{f}}^{i_\text{l}} \hat{\Phi}_{ss,\mu}^{\text{in}}(\kappa)}{\hat{\Phi}_{ss,i}^{\text{in}}(\kappa)}}\,. \tag{8}$$

It can be shown with [3], that the gradients λ_i and thus the optimum PSD tend to be equally distributed over all subbands below 4.4 kHz. Since the input speech usually has a spectral tilt towards lower frequencies, the spectral weighting shows a bandpass character.

3.4.2 Medium SNR: Transition to Noise-Like Shape

With increasing SNR, in more and more subbands the optimum solution (6) will fall outside the feasible range and is limited to the bound $D_i(\kappa) + 15\,\text{dB}$ during the next recursion step. Accordingly, the optimum solution converges towards

$$E_i^{\text{opt}}(\kappa) = D_i(\kappa) + 15\,\text{dB} \tag{9}$$

and the spectral shape of the output speech roughly follows that of the noise.

3.4.3 High SNR: Transition to Unity Weight

At high SNR, the noise becomes less dominant and the optimum speech spectrum level turns to $E_i^{\text{opt}}(\kappa) = E_i^{\text{in}}(\kappa)$, which results in unity spectral weights

$$W_i(\kappa) = 1\,. \tag{10}$$

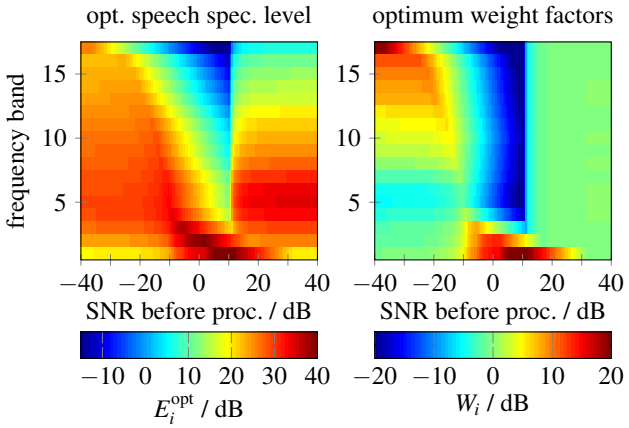

Figure 2: Optimum speech spectrum level and resulting weights after recursive closed-form optimization based on average spectrum levels of TIMIT database and car interior noise.

3.5 Narrow Bandpass Weights

Even though, the recursive closed-form optimization results in an optimized SII of the output speech and improves intelligibility in various noise environments, it leads for special noise scenarios to frequency weights which have a narrow bandpass characteristic and thus a disadvantageous or even destructive effect on listening experience and speech intelligibility (which is, however, not considered by the SII).

As derived in Section 3.4.2, the optimum solution converges for medium SNRs towards $E_i^{\mathrm{opt}}(\kappa) = D_i(\kappa) + 15\,\mathrm{dB}$. If the noise signal has a narrow bandpass disturbance spectrum, this leads to large weight factors in the corresponding frequency bands. This alone is at least annoying for the listener, but due to the tight audio power constraint, all other frequency bands will be attenuated to allow amplification in the noisy bands within the power constraint.

The "narrow bandpass effect" becomes very apparent for the car interior noise of the NOISEX-92 database, as this noise signal accumulates almost all its energy in the three frequency bands below 0.4 kHz, where the speech of the TIMIT database only has comparatively few energy in these bands. In Figure 2, the optimum speech spectrum level and the resulting weight factors are exemplary plotted for the average disturbance spectrum level of the car interior noise. At SNRs of -5 to 10 dB, the lowest three frequency bands are amplified by up to 20 dB and all other bands are attenuated by 15 to 20 dB. To make things even worse, the input signal may have more noisy content in these bands than useful speech information, due to a high pitch of the far-end speaker or an unfavorable transfer characteristic of the communication system chain.

Similar problems would arise for other mono-frequent or bandpass noise sources with a peak at higher frequencies, like alarm signals or brake squeal of trains. Therefore, any solution which only copes with this special problem of car noise at low frequencies would not be sufficient.

3.6 Novel Adaptively Parametrized One-Step Closed-Form Optimization

The recursive closed-form optimization experiences some difficulties at medium SNRs and bandpass noises as explained above. The performance in low as well as high SNR situations, on the other hand, is not effected by bandpass noises as the resulting weights are almost independent of the noise.

This motivates the new approach to replace the SII optimized spectral weights of Section 3.3 for medium SNRs by cross-fading between (8) at low SNR and (10) at high SNR using the weighting rule

$$W_i(\kappa) = \sqrt{\frac{\lambda_i^{1-\gamma(\kappa)} \cdot \left[\hat{\Phi}^{\mathrm{in}}_{ss,i}(\kappa)\right]^{\gamma(\kappa)}}{\sum_{\mu=i_{\mathrm{l}}}^{i_{\mathrm{l}}} \lambda_\mu^{1-\gamma(\kappa)} \cdot \left[\hat{\Phi}^{\mathrm{in}}_{ss,\mu}(\kappa)\right]^{\gamma(\kappa)}} \cdot \frac{\sum_{\mu=i_{\mathrm{f}}}^{i_{\mathrm{l}}} \hat{\Phi}^{\mathrm{in}}_{ss,\mu}(\kappa)}{\hat{\Phi}^{\mathrm{in}}_{ss,i}(\kappa)}}, \quad (11)$$

where $\gamma(\kappa)$ is a (time-adaptive) parameter with $0 \le \gamma(\kappa) \le 1$.

In heuristic experiments it turned out, that the transition between low and high SNR is well indicated by the signal-to-disturbance difference (SDD) of the first closed-form solution (6) in the "best" subbands. Therefore, the parameter $\gamma(\kappa)$ is chosen as follows:

1. Calculate the speech spectrum level $E_i^{(1)}(\kappa)$ of the first closed-form solution (6).
2. Calculate the SDD $\psi_i(\kappa) = E_i^{(1)}(\kappa) - D_i(\kappa)$ of the first closed-form solution.
3. Calculate the average SDD $\overline{\psi}(\kappa)$ as the arithmetic mean (in dB) of the $\overline{\psi}_{\mathrm{contr}}$ largest SDDs $\psi_i(\kappa)$.
4. Calculate the parameter $\gamma(\kappa)$ as

$$\gamma(\kappa) = \min\left\{\max\left\{\frac{\overline{\psi}(\kappa) - \overline{\psi}_{\mathrm{b}}}{\overline{\psi}_{\mathrm{e}} - \overline{\psi}_{\mathrm{b}}}, 0\right\}, 1\right\}, \quad (12)$$

where $\overline{\psi}_{\mathrm{b}}$ and $\overline{\psi}_{\mathrm{e}}$ denote the begin resp. end of the transition range.

Simulations with various noise signals show, that the settings $\overline{\psi}_{\mathrm{contr}} = 2$, $\overline{\psi}_{\mathrm{b}} = 15\,\mathrm{dB}$, and $\overline{\psi}_{\mathrm{e}} = 35\,\mathrm{dB}$ provide a good compromise between $\mathrm{STI}_{\mathrm{sr}}$ and SII rating over the whole transition range and for all evaluated noise signals.

4 Experimental Evaluation

4.1 Simulation Environment

The performance of the presented algorithms is evaluated for every speech file of the TIMIT database, 5.4 hours in total, disturbed by speech babble, white noise, or car interior noise (`volvo`) from the NOISEX-92 database at a sampling rate of $f_{\mathrm{s}} = 8\,\mathrm{kHz}$. Prior to processing, each speech file is scaled to match an overall active speech level [10] corresponding to a sound pressure level of 62.35 dB as specified in [3] for normal voice effort. The desired input SNRs ranging from -40 dB to 40 dB in steps of 2.5 dB are achieved by adjusting the overall level of the noise file in relation to a sound pressure level of 62.35 dB.

As first instrumental measure the Speech Intelligibility Index (SII) is calculated using the so-called critical band procedure [3] for every speech file and afterwards averaged. Good communication systems have an SII of 0.75 or better while the SII of poor communication systems is below 0.45.

The algorithms are further compared regarding an adaptation of the speech-based STI using the envelope regression method [11] to the revised STI [12], which we call speech-based revised Speech Transmission Index ($\mathrm{STI}_{\mathrm{sr}}$). The $\mathrm{STI}_{\mathrm{sr}}$ ratings range from unintelligible (≤ 0.3) to excellent (≥ 0.75).

4.2 Results

Figure 3 shows the performance of the presented NELE algorithms under the constraint that the short-term power of the output signal is less or equal than the short-term power of the input signal.

The recursive closed-form optimization yields a SII of a good communication system at a 3 dB to 7 dB lower input

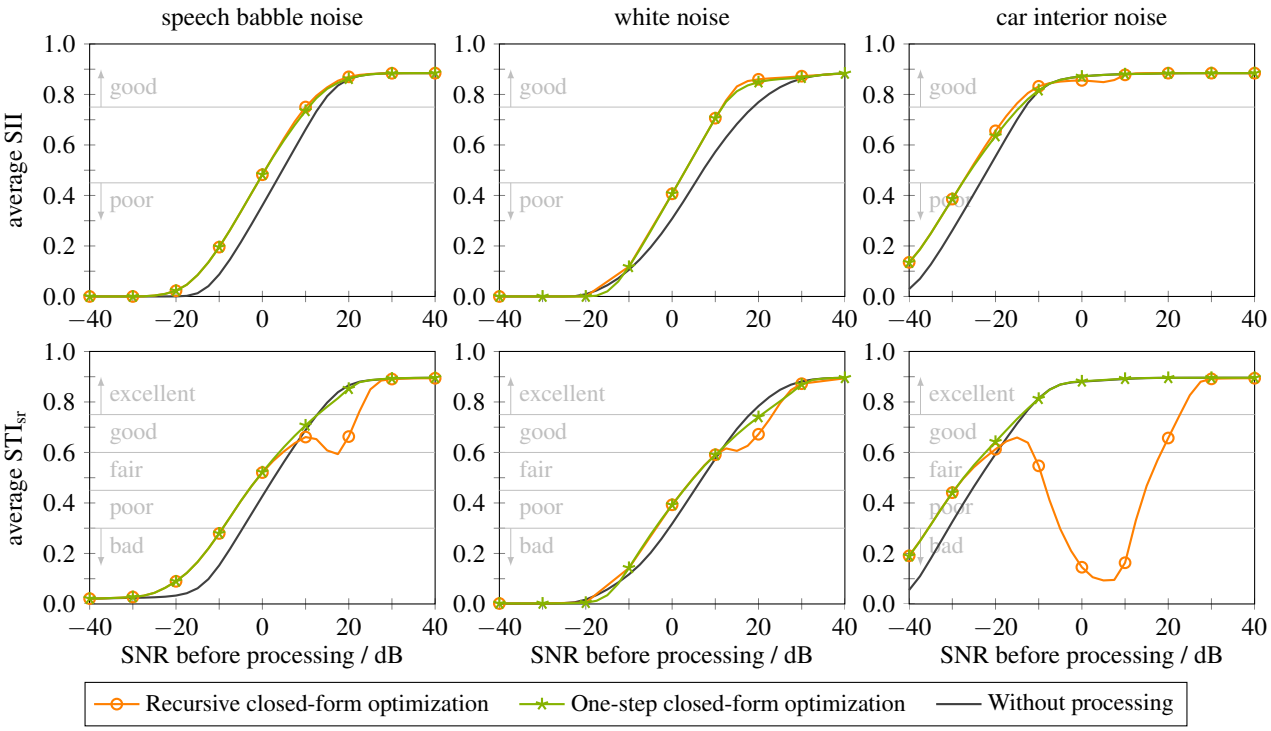

Figure 3: Comparison of presented algorithms with no additional audio power spent.

SNR compared to a system without processing. As explained above, the STI$_{sr}$ rating is deteriorated for all evaluated noise types at a medium to mid-high SNR range. Especially for the car interior noise, the average STI$_{sr}$ rating has a deep notch and reaches its minimum of below 0.1 at a SNR before processing of +5 dB.

The adaptively parametrized one-step closed-form optimization eliminates the notches in STI$_{sr}$ completely. Even though for medium SNRs the replaced spectral weights are sub-optimal w. r. t. the SII, the average SII rating is still very comparable.

Sound samples and further information can be found at `http://www.ind.rwth-aachen.de/~bib/sauert12/`.

5 Conclusions

In this contribution, a new adaptively parametrized one-step closed-form SII optimization algorithm for NELE is derived, with the constraint that the output signal has the same average audio power as the input signal. The novel weighting rule yields good results even with narrow bandpass noise signals, where the previously proposed recursive closed-form optimization [4] turned out to fail.

The instrumental evaluation by means of the average SII has shown a comparable performance after processing with both algorithms, which is noticeably better than without processing. Concerning the average STI$_{sr}$ rating, the previous algorithm shows deep notches for mid-high SNR ranges, which are eliminated with the proposed algorithm.

References

[1] 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 *Proc. of ICASSP*, (Kyoto, Japan, Mar. 25–30, 2012), Mar. 2012, pp. 4061–4064.

[2] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index", in *Proc. of EUSIPCO*, (Glasgow, Scotland, Aug. 24–28, 2009), vol. 17, Aug. 2009, pp. 1844–1848.

[3] ANSI S3.5-1997, *Methods for the calculation of the speech intelligibility index*, ANSI, 1997.

[4] B. Sauert and P. 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, Oct. 2010, ISBN: 978-3-8007-3300-2.

[5] B. Sauert, H. W. Löllmann, and P. Vary, "Near end listening enhancement by means of warped low delay filterbanks", in *Proc. of ITG-Fachtagung Sprachkommunikation*, (Aachen, Germany, Oct. 8–10, 2008), vol. 8, Oct. 2008, ISBN: 978-3-8007-3120-6.

[6] B. Sauert and P. Vary, "Near end listening enhancement considering thermal limit of mobile phone loudspeakers", in *Proc. of ESSV*, (Aachen, Germany, Sep. 28–30, 2011), vol. 61, Sep. 2011, pp. 333–340, ISBN: 978-3-94271-037-4.

[7] H. W. Löllmann and P. Vary, "Uniform and warped low delay filter-banks for speech enhancement", *Speech Communication*, vol. 49, pp. 574–587, 7-8 Jul. 2007, ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.04.009.

[8] P. Vary, "An adaptive filterbank equalizer for speech enhancement", *Signal Processing*, vol. 86, no. 6, pp. 1206–1214, Jun. 2006. DOI: 10.1016/j.sigpro.2005.06.020.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity", in *Proc. of ICASSP*, (Dallas, Texas, USA, Mar. 14–19, 2010), Mar. 2010, pp. 4266–4269. DOI: 10.1109/ICASSP.2010.5495680.

[10] ITU-T Recommendation P.56, *Objective measurement of active speech level*, version 03/93, ITU, Mar. 1993.

[11] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations", *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004, ISSN: 0001-4966. DOI: 10.1121/1.1804628.

[12] IEC 60268-16:2003, *Sound system equipment – Part 16: objective rating of speech intelligibility by speech transmission index*, IEC, May 2003.