NEAR-END LISTENING ENHANCEMENT BY NOISE-INVERSE SPEECH SHAPING

Markus Niermann, Peter Jax, and Peter Vary

Institute of Communication Systems
RWTH Aachen University, Germany
{niermann,jax,vary}@iks.rwth-aachen.de

ABSTRACT

In communication systems, clean speech is often reproduced by loudspeakers and disturbed by local acoustical noise. Near-end listening enhancement (NELE) is a technique to enhance the speech intelligibility in environmental noise by adaptively preprocessing the speech based on a noise estimate. Conventional NELE-algorithms adaptively filter the speech by applying spectral gains which are determined by maximizing intelligibility measures. Usually, this leads to speech amplifications at highly disturbed frequencies to overcome masking. In this paper, a new approach is presented which shapes the speech spectrum according to the inverse of the noise power spectrum. It is based on a simple gain rule. Its advantages are a predictable spectral behavior and a fixed computational complexity, since no optimization problem with an unknown number of iterations needs to be solved. Simulations have shown that it copes with a wide range of noise types and provides a similar performance compared to conventional algorithms.

1. INTRODUCTION

The objective of near-end listening enhancement (NELE) is to improve speech intelligibility of communication systems in noisy environments, for example for mobile telephony [1], hands-free communication in cars [2] and public announcement systems [3]. Known contributions optimize an approximation of the *Speech Intelligibility Index* (SII) [4] by redistributing the speech spectral power based on the estimated power spectrum of the noise [5–8]. Zorila et al. [9] and Schepker et al. [10] make use of a dynamic range compressor; [10] additionally distributes speech power uniformly over frequency, controlled by an estimated SII-measure.

Typical objective measures for the evaluation of speech enhancement are the SII [4] and the *Speech Transmission Index* (STI) [11]. The SII is an energy-based measure which predicts intelligibility of speech in noise by taking masking effects into account. The STI is a measure for intelligibility, too, which additionally considers speech distortions and the frequency response of an enhancement system.

In this paper, a new approach is proposed which distributes the speech power proportionally to the inverse noise spectrum. This approach is realized as a simple, low complexity gain rule by applying spectral gains in critical subbands.

2. COMPARISON TO LITERATURE

In this section, the SII-based optimization algorithms *OptSII-recurDist* by Sauert [8] and the proposal of Taal [7] are analyzed and compared to the proposed approach at a sampling rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. In general, SII-based algorithms consider an approximate SII measure to obtain a convex optimization problem. Both considered algorithms work in a subband domain as specified in the SII-standard. At medium signal-to-noise ratios (SNR), Sauert [8] as well as Taal [7] shape the speech power spectrum according to the noise spectral power, i.e., speech power is redistributed over frequency such that the speech overcomes masking by noise in each subbband. In extreme low SNR cases Sauert shapes the speech power spectrum according to the *Band Importance Function* from [4], e.g., it concentrates the power principally in the range 450-4000 Hz. Taal's optimization acts similar, but in contrast to Sauert it selectively eliminates highly disturbed frequency bands and redistributes the saved power to subbands in which it contributes to an increased intelligibility. In high SNR conditions, Sauert does not modify the speech, whereas Taal still redistributes speech power over frequency. Both algorithms support the constraint not to increase the total audio power. Sauert additionally provides a mode in which a maximum amplification or a maximum absolute sound pressure level is specified.

The two above mentioned algorithms filter the speech by applying spectral gains in a critical subband domain, which are obtained by solving an optimization problem. The resulting closed-form expressions are processed iteratively until an optimal solution is found. The number of iterations as well as the weighting behaviour of the algorithm are not known a-priori. In contrast, the proposed algorithm filters the speech by applying spectral gains which do not depend on an optimization problem, but on a simple gain rule. Therefore, the algorithm is well predictable in terms of run time and spectral behaviour. Moreover, its computational complexity is very low. It supports the constraint of equal input and output speech power, which allows fair comparisons between algorithms and is applicable for example in mobile phones and hands-free communication systems.

The basic idea of this proposal is to attenuate the speech in frequency bands with high disturbance, where it barely

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

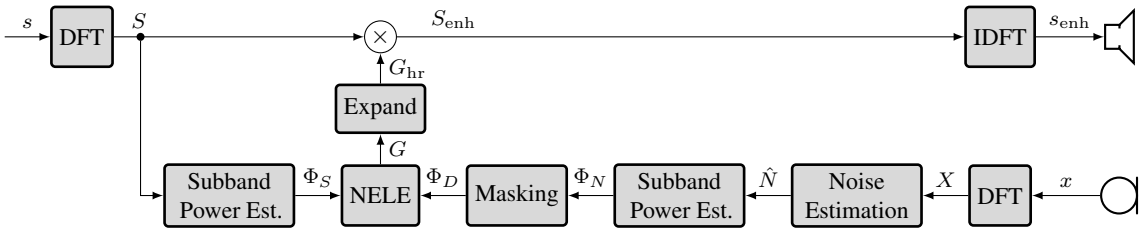

Fig. 1: Overview of the system framework.

contributes to the intelligibility, and to redistribute the power into less-disturbed subbands. In contrast to Sauert and Taal, regarding low SNR conditions, the speech spectrum is shaped according to the *inverse* of the noise power spectrum. In high SNR cases the speech passes the enhancement system without being filtered to maintain the original tone color. In medium SNR ranges it would be beneficial to shape the speech spectrum according to the noise power spectrum as in [8]. However, it turned out that the intelligibility is not significantly impaired if the same processing as in the low SNR case is performed. Therefore, we interpolate between the inverse noise shape and the original speech shape to account for the range between highly disturbed and undisturbed environments. The interpolation is controlled by a simplified SII-measure.

3. FRAMEWORK

An overview of the framework is given in Fig. 1. The input signals are the clean speech from the far-end $s(k)$ and the near-end microphone signal $x(k)$, composed of near-end noise and potentially speech. The enhanced output speech signal $s_{\mathrm{enh}}(k)$ will be acoustically reproduced by the near-end loudspeaker.

At first $s(k)$ is segmented into overlapping frames (index λ, frame length M_F, frame advance M_{adv}), windowed by a time-domain window function $w(k)$ and transferred to the frequency domain (frequency index μ) by means of a discrete Fourier transform (DFT) of length M_F:

$$S(\lambda,\mu) = \sum_{k=0}^{M_F-1} s\left(k+(\lambda-1)M_{\mathrm{adv}}\right)\cdot w(k)\cdot e^{-j2\pi\frac{\mu k}{M_F}}. \quad (1)$$

$x(k)$ is transformed to the frequency domain $X(\lambda,\mu)$ analogously. An estimation of the near-end noise power spectrum $|\hat{N}|^2$ is obtained from X using the Speech Presence Probability (SPP) algorithm [12].

Next, the frequency domain signals are transformed to $M_{\mathrm{SB}} = 21$ critical subbands (index i) by combining several frequency bins. To achieve this, frequency domain windows $W_i(\mu)$ are designed with an overall flat frequency characteristic and a linear slope in the transition between adjacent windows. The windows are designed according to [4, Table 1], where the center, low and high frequencies $f_{\mathrm{c},i}$, $f_{\mathrm{l},i}$ and $f_{\mathrm{h},i}$ are defined for each band. The bandwidths can be deduced from the window according to:

$$\Delta f_i = \frac{f_{\mathrm{s}}}{M_F}\cdot\sum_{\mu=0}^{M_F} W_i(\mu), \quad i=1,...,M_{\mathrm{SB}}. \quad (2)$$

According to [4] the subband powers are normalized to Δf_i. For complexity reasons, this step is included into the frequency window:

$$\begin{aligned} W_i'(\mu) &= W_i(\mu)\cdot\frac{1}{\Delta f_i} \\ &= \frac{W_i(\mu)}{\sum_{\zeta=0}^{M_F} W_i(\zeta)}\cdot\frac{M_F}{f_{\mathrm{s}}}. \end{aligned} \quad (3)$$

Under these preconditions the frequency-normalized subband powers for far-end speech and near-end noise are calculated:

$$\phi_S(\lambda,i) = \frac{1}{c_{\mathrm{w}}}\cdot\sum_{\mu=0}^{M_F}|S(\lambda,\mu)|^2\cdot W_i'(\mu), \quad i=1,...,M_{\mathrm{SB}}, \quad (4)$$

$$\phi_N(\lambda,i) = \frac{1}{c_{\mathrm{w}}}\cdot\sum_{\mu=0}^{M_F}|\hat{N}(\lambda,\mu)|^2\cdot W_i'(\mu), \quad i=1,...,M_{\mathrm{SB}}, \quad (5)$$

For an unbiased power estimator, the influence of the time-domain window $w(k)$ on power levels is compensated using a constant c_{w}:

$$c_{\mathrm{w}} = \frac{1}{M_F}\cdot\sum_{k=0}^{M_F-1} w^2(k). \quad (6)$$

To avoid fast changing weighting gains, the speech subband powers are recursively smoothed over time using a smoothing constant β. A binary Voice Activity Detector (VAD) $v(\lambda)\in\{0,1\}$ is implemented according to [13] to ensure that only far-end frames with voice activity will be considered.

$$\Phi_S(\lambda,i) = \begin{cases} (1-\beta)\cdot\phi_S(\lambda,i)+\beta\cdot\Phi_S(\lambda-1,i), & v(\lambda)=1 \\ \Phi_S(\lambda-1,i), & v(\lambda)=0. \end{cases} \quad (7)$$

The noise powers are smoothed accordingly:

$$\Phi_N(\lambda,i) = (1-\beta)\cdot\phi_N(\lambda,i)+\beta\cdot\Phi_N(\lambda-1,\mu). \quad (8)$$

In the following, only smoothed power estimates will be used. To account for masking effects, a disturbance power Φ_D is calculated as in [4, 8] under the assumption that self-speech masking and internal noise floors can be neglected. The calculation rule respects inter-band masking from lower to higher bands:

$$\Phi_D(\lambda,i) = \Phi_N(\lambda,i)+\sum_{\zeta=1}^{i-1}\Phi_N(\lambda,\zeta)\cdot C_\zeta(\lambda)^{\log_2\frac{f_{\mathrm{c},i}}{f_{\mathrm{h},\zeta}}}, \quad (9)$$

with a slope per octave of

$$C_i(\lambda) = 10^{-8} \cdot (\Phi_N(\lambda, i) \cdot \Delta f_i)^{0.6}. \quad (10)$$

Based on the power estimates Φ_S and Φ_D, the gain rule in Sec. 4 determines weighting gains $G(\lambda, i)$ for each frequency band. They are expanded to the high-resolution frequency domain,

$$G_{\mathrm{hr}}(\lambda, \mu) = \sum_{i=1}^{M_{\mathrm{SB}}} G(\lambda, i) \cdot W_i(\mu), \quad (11)$$

multiplied to the speech, $S_{\mathrm{enh}}(\lambda, \mu) = S(\lambda, \mu) \cdot G_{\mathrm{hr}}(\lambda, \mu)$, and transformed to the time-domain $s_{\mathrm{enh}}(k)$ by means of Inverse Fast Fourier Transform (IFFT) and overlap-add.

4. NEW GAIN RULE

In this section, the new rule for the gain $G(\lambda, i)$ is developed. For reasons of simplicity, λ and i are omitted if possible.

In low disturbance cases, speech is well understandable without listening enhancement and should not be influenced by the system. The processed speech power spectrum $G^2 \cdot \Phi_S$ will equal the original speech power spectrum:

$$G^2 \cdot \Phi_S = \Phi_S, \quad (12)$$

with unity gains. In high disturbance cases, the speech power spectrum is shaped proportionally to the inverse of the disturbance power:

$$G^2 \cdot \Phi_S \sim \frac{1}{\Phi_D}. \quad (13)$$

To avoid excessive amplifications in frequency bands with relatively low disturbance power, the disturbance power spectrum is modified by applying a minimum threshold:

$$\Phi'_D(\lambda, i) = \max\left(\Phi_D(\lambda, i), \frac{\gamma}{M_{\mathrm{SB}}} \cdot \sum_{i=1}^{M_{\mathrm{SB}}} \Phi_D(\lambda, i)\right). \quad (14)$$

The threshold is chosen to be a fraction of the average disturbance subband power with $0 < \gamma < 1$. Based on this, the processed speech power spectrum in the high disturbance case is set to

$$G^2 \cdot \Phi_S = c \cdot \frac{1}{\Phi'_D}. \quad (15)$$

The variable c will be chosen later such that the total speech power remains unchanged. To cover also intermediate scenarios between undisturbed environments and high disturbances, an interpolation between Eq. 12 and 15 is performed in the logarithmic amplitude domain:

$$G^2 \cdot \Phi_S = c \cdot \left(\frac{1}{\Phi'_D}\right)^{\alpha} \cdot (\Phi_S)^{1-\alpha}, \quad (16)$$

using a control parameter $\alpha \in [0, 1]$. Finally, the constant c ensures that the total input and output speech power are equal:

$$\sum_{i=1}^{M_{\mathrm{SB}}} G^2 \cdot \Phi_S \cdot \Delta f = \sum_{i=1}^{M_{\mathrm{SB}}} \Phi_S \cdot \Delta f. \quad (17)$$

$$\Leftrightarrow c = \frac{\sum_{i=1}^{M_{\mathrm{SB}}} \Phi_S \cdot \Delta f_i}{\sum_{i=1}^{M_{\mathrm{SB}}} (\Phi'_D)^{-\alpha} \cdot (\Phi_S)^{1-\alpha} \cdot \Delta f_i} \quad (18)$$

4.1. Choice of the control parameter

The control parameter $\alpha(\lambda)$ regulates the aggressiveness of the enhancement system. For $\alpha = 0$ the speech remains unprocessed. For $\alpha = 0.5$ the resulting speech spectrum is the geometric mean of the original speech spectrum and the inverse disturbance spectrum, i.e., the speech is partly adapted to the noise. It is not recommended to completely adapt the speech to the noise ($\alpha = 1$) since the original spectral envelope would be lost and the speech would sound unnatural. Informal tests have shown that α should be limited to $[0, \alpha_{\max}]$ with $\alpha_{\max} = 0.5$.

The influence of α on objective measures is visualized in Fig. 2. In low SNR conditions, the best intelligibilities in terms of STI are achieved by choosing high interpolation values. In contrast, low values of α maximize the STI in high SNR cases.

The interpolation parameter is controlled by a measure of current intelligibility. At low intelligibilities high values of α are chosen and vice versa. We use a simplified short-time SII

$$\hat{I}(\lambda) = \sum_{i=1}^{M_{\mathrm{SB}}} A_i \cdot \max\left[0, \min\left[1, \frac{10\,\mathrm{lg}\,\frac{\Phi_S(\lambda,i)}{\Phi_D(\lambda,i)} + 15\,\mathrm{dB}}{30\,\mathrm{dB}}\right]\right] \quad (19)$$

which is a mean of the subband intelligibilities, weighted by the *Band Importance Function* A_i [4]. A_i allocates less importance to subbands with center frequencies below 450 Hz or above 4000 Hz. This implicitly accounts for the fact that the impairment of speech intelligibility by noise depends not only on the SNR, but also on the noise type. SII thresholds for good and bad intelligibilities are $I_{\mathrm{high}} = 0.85$ and $I_{\mathrm{low}} = 0.45$, respectively. α is chosen with a linear dependency to $\hat{I}$ such that the full enhancement is achieved below I_{low} and no filtering is performed above I_{high}:

$$\alpha(\lambda) = \alpha_{\max} \cdot \max\left[0, \min\left[1, \frac{I_{\mathrm{high}} - \hat{I}(\lambda)}{I_{\mathrm{high}} - I_{\mathrm{low}}}\right]\right]. \quad (20)$$

The dashed curve in Fig. 2 affirms that Eq. 20 leads to a reasonable choice of α in terms of STI.

5. EVALUATION

In this section, the performance of the new algorithm is evaluated and compared to state-of-the-art algorithms by means of simulations, using the parameters in Table 1. For the simulation, speech files are taken randomly from the TIMIT database [14] and normalized to 62.35 dB SPL as in [4]. Five different noise types (cf. Table 2) are used as environmental noise. The noise level is adjusted in steps such that the overall SNR at the listener is between -30 dB and 30 dB. The proposed algorithm is compared to the published Matlab implementation of Taal [7] and *OptSIIrecurDist* by Sauert [8]. The first two

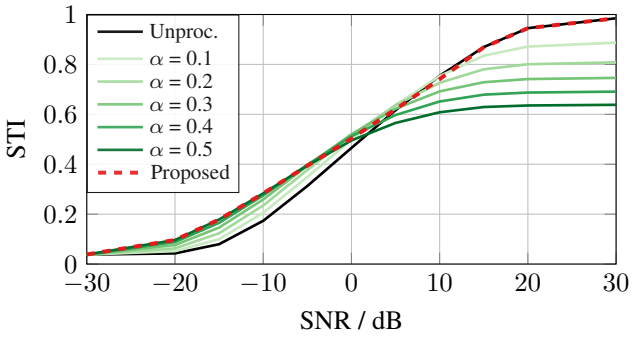

Fig. 2: STI as a function of SNR for different choices of α.

Parameter	Settings
Sampling frequency f_s	16 kHz
FFT length M_F	512 (with zero padding)
Frame length M	320 $\widehat{=}$ 20 ms
Frame advance M_{adv}	160 $\widehat{=}$ 10 ms
Time-domain window $w(k)$	$\sqrt{\text{Hann}}$
Number subbands M_{SB}	21
Smoothing parameter β	0.9802 $\widehat{=}$ 1 s
Disturbance limitation γ	0.1

Table 1: Simulation settings.

		SII | STI			
Noise Type	**SNR**	**Unproc.**	**Proposed**	**Sauert**	**Taal**
White [16]	0 dB	0.36 | 0.35	0.45 | 0.38	0.46 | 0.39	0.46 | 0.38
Babble [16]	0 dB	0.39 | 0.47	0.51 | 0.50	0.51 | 0.53	0.51 | 0.52
Volvo [16]	-10 dB	0.37 | 0.43	0.48 | 0.48	0.49 | 0.50	0.50 | 0.49
Traffic [17]	0 dB	0.29 | 0.34	0.40 | 0.41	0.41 | 0.42	0.42 | 0.43
Bandpass	-10 dB	0.64 | 0.64	0.67 | 0.59	0.64 | 0.64	0.69 | 0.41

Table 2: Instrumental Measures SII and STI as a function of the noise type and the algorithm. Frequency range of bandpass (BP) noise: 800-1100 Hz.

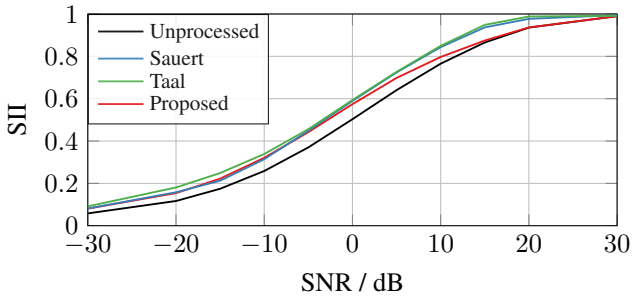

Fig. 3: Comparison of NELE algorithms in terms of SII.

seconds of each processed speech file are not considered for the evaluation in order to remove possible transient effects. In total 225 s of speech are evaluated per noise type, SNR setting and algorithm. Despite the constraint of not increasing the speech power, the algorithms might produce slight deviations in the output power due to slight inaccuracies in the estimation of the input speech power. For the fairness of comparison we normalize the enhanced speech to the same power a-posteriori.

Two special noise types are discussed first to clarify the behavior of the gain rule: White noise and bandpass noise. The proposed algorithm is based on a weighted geometric average of the original speech spectrum and the inverse noise (Eq. 16). In the case of bandpass noise, speech power from the disturbed frequency range is redistributed to undisturbed frequency parts. In the presence of white noise, the speech power spectrum becomes flat. Since speech usually exhibits a decreasing spectral slope to high frequencies, this corresponds to a highpass which is known to increase intelligibility [15].

Table 2 shows the results of the enhancement algorithms for different noise types. With regard to white, babble, in-car (Volvo) and traffic noise, the proposal enhances the intelligibility significantly compared to the unprocessed case and is competitive with Sauert and Taal. Only for bandpass noise, the STI decreases, but the SII and informal listening tests point out improvements.

Figures 3 and 4 visualize the performance of the algorithms as a function of SNR, averaged over the noise types from Table 2. We do not expect significant SII-improvements compared to Sauert and Taal, who employ the SII as optimization criterion. However, in the relevant SNR range between -10 dB and 5 dB the proposed algorithm performs similarly well as Taal and Sauert and achieves a significant improvement in comparison to the unprocessed case. In high SNR conditions (> 10 dB) the results are inferior to conventional algorithms, but there is no need for intelligibility improvement due to the fact that speech is already well understandable in these conditions. In very bad situations ($\text{SNR} < -5$ dB) speech is hardly understandable, even with NELE. In terms of STI, the proposed algorithm competes with Taal and Sauert for $\text{SNR} < 10$ dB. Taal's intelligibility saturates at $\text{STI} = 0.64$ because the spectral gains do not converge to 1 for high SNRs in the presented range. To sum up, the new approach produces similar results as conventional algorithms, but the complexity is significantly lower since no real-time optimization is required.

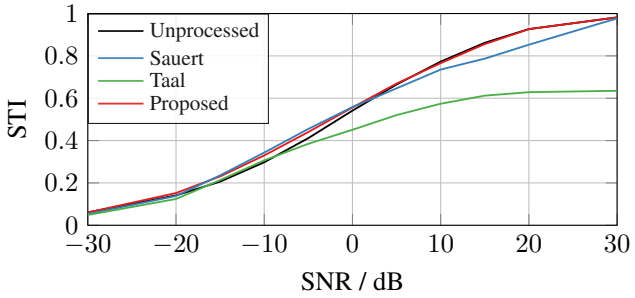

Fig. 4: Comparison of NELE algorithms in terms of STI.

6. CONCLUSIONS

In this contribution, a new NELE algorithm is presented which distributes the speech spectral power similar to the inverse disturbance spectral power. Conventional algorithms are based on solving a realtime optimization problem iteratively with varying runtimes. Sauert, for example, needs an unknown number of recursive steps to find an optimal solution in the admissible range (usually less than 3 steps though) [8, p.59] and also Taal's optimization results in a recursive dependency which must be solved by evaluating expressions with an unknown number of different parameter values. In contrast, the proposed algorithm calculates the spectral gains by evaluating only one simple equation. Therefore, the computational complexity is predetermined. Moreover, the new gain rule leads to a predictable spectral behaviour. It copes with a wide range of different noise types and provides similar evaluation results as state-of-the-art techniques.

Although the processed speech sounds unnatural due to a high tone color, informal listening indicates that presented with noise, the speech modifications do not disturb the listener. Instead, they support the comprehension. The best objective results are achieved for $\alpha_{\max} = 0.5$, but subjective tests exhibit higher intelligibilities for a more aggressive setting ($\alpha_{\max} = 0.7$). This reveals slight discrepancies between objective measures and subjective impressions.

REFERENCES

[1] B. Sauert, F. Heese, and P. Vary, "Real-Time Near-End Listening Enhancement for Mobile Phones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. May 2014, IEEE, Show and Tell Demonstration.

[2] M. Niermann, F. Heese, and P. Vary, "Intelligibility Enhancement For Hands-Free Mobile Communication," in *Proceedings of German Annual Conference on Acoustics (DAGA)*. 2015, pp. 384–387, DEGA.

[3] R. Hendriks, J. Crespo, J. Jensen, and C. Taal, "Speech Reinforcement in Noisy Reverberant Conditions under an Approximation of the Short-Time SII," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, Apr. 2015, pp. 4400–4404.

[4] ANSI S3.5-1997, *Methods for the Calculation of the Speech Intelligibility Index*, ANSI, 1997.

[5] 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)*. Aug. 2010, pp. 1919–1923, EURASIP.

[6] B. Sauert and P. Vary, "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement," in *ITG-Fachtagung Sprachkommunikation*, Berlin, Germany, Oct. 2010, VDE Verlag GmbH.

[7] C. H. Taal, J. Jensen, and A. Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement.," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[8] B. Sauert and P. Vary, *Near-End Listening Enhancement: Theory and Application*, PhD thesis, IND, RWTH Aachen University, Aachen, May 2014.

[9] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-In-Noise Intelligibility Improvement based on Power Recovery and Dynamic Range Compression," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug. 2012, pp. 2075–2079.

[10] H. F. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression.," in *INTERSPEECH*, 2013, pp. 3577–3581.

[11] R. L. Goldsworthy and J. E. Greenberg, "Analysis of Speech-Based Speech Transmission Index Methods with Implications for Nonlinear Operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[12] T. Gerkmann and R. C. Hendriks, "Noise Power Estimation based on the Probability of Speech Presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011*. 2011, pp. 145–148, IEEE.

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

[14] J. S. Garofolo and L. D. Consortium, *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.

[15] R. J. Niederjohn and J. H. Grotelueschen, "The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 277–282, 1976.

[16] A. Varga and H. J. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] ETSI EG 202 396-1, *Speech and multimedia Trans. Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation techniques and background noise database*, ETSI, Mar. 2009.