Improving speech intelligibility in background noise by SII-dependent amplification and compression*

Henning Schepker[1], Jan Rennies[2]. Simon Doclo[1,2]

[1] *University of Oldenburg, Department of Medical Physics and Acoustics,*
Signal Processing Group, Oldenburg, Germany,
{henning.schepker,simon.doclo}@uni-oldenburg.de

[2] *Fraunhofer IDMT Project Group Hearing-, Speech-, and Audio Technology, Oldenburg, Germany,*
jan.rennies@idmt.fraunhofer.de

Introduction

In many speech communication applications it is of great interest to achieve a high intelligibility to ensure good communication. However, in these applications speech is often disturbed by additive noise and/or reverberation. Therefore, it is desirable to develop algorithms that are able to maintain a high intelligibility in such disturbed scenarios. While amplifying the speech to achieve good signal-to-noise ratios (SNR) is an easy approach, it is often not applicable due to technical limitations of the amplification system or unpleasantly high sound levels. Consequently, algorithms that increase speech intelligibility while maintaining equal powers are preferable.

Several algorithms have been proposed in the past that use either frequency-dependent amplification, dynamic range compression, transient amplification, or modulations filtering techniques.

The first attempt to investigate the effect of different signal processing strategies on speech intelligibility was made by Licklider and Pollack [1]. While they did not consider any additive noise or reverberation they could demonstrate that speech intelligibility in quiet is not necessarily affected by strategies such as high-pass or low-pass filtering and clipping.

Niederjohn and Grotelueschen [2] proposed a preprocessing algorithm that uses high-pass filtering followed by static rapid amplitude compression. They observed an increase in speech intelligibility for preprocessed speech in white noise over the unprocessed speech at the same SNR. Zorila et al. [4] adopted the idea of dynamic range compression as a mean to increase the speech intelligibility. They used a static input-output characteristic for their dynamic range compression and used several frequency-dependent amplification steps prior to compression.

Recently, Sauert and Vary proposed an algorithm that uses time- and frequency-dependent amplification of the speech signal aiming to maximize the SII [3]. However, this approach suffers from spectral adaption to the background noise. Therefore in a recent approach they considered an SNR-dependent transition between SII-weighted and unity-weighting of the speech signal [6].

In this contribution we describe an algorithm and its evaluation using objective measures and formal listening tests that combines time- and frequency dependent amplification and time- and frequency-dependent dynamic range compression. In the following we will first describe the considered scenario and our proposed algorithm. Then the evaluation using objective measures and a listening test is described. A correlation analyses between objective and subjective data is carried out and the last section concludes this contribution.

Scenario

Consider the acoustic scenario depicted in Fig. 1. The unprocessed clean speech signal $s[k]$ at discrete time k is modified using the weighting function $W\{\cdot\}$ and played back via a loudspeaker. A microphone picks up the disturbed signal $y[k]$ which consists of the convolutive mixture of the modified speech signal $\tilde{s}[k]$ and the room impulse response $h[k]$ and the additive noise disturbance $r\,[k]$, i.e.

$$y[k] = \tilde{s}[k] * h[k] + r[k], \tag{1}$$

where $*$ denotes convolution. An estimate $\hat{r}[k]$ of the noise signal $r[k]$ can be obtained by using e.g. adaptive filtering techniques to model the room impulse response $h[k]$. Using the estimated noise $\hat{r}[k]$, the estimated impulse response $\hat{h}[k]$, and the known clean speech signal the processed speech signal $\tilde{s}[k]$ is then computed as:

$$\tilde{s}[k] = W\{s[k],\hat{r}[k],\hat{h}[k]\}s[k] \tag{2}$$

In the following we assume that a perfect noise estimate is available, i.e. $\hat{r}[k] = r[k]$, and no reverberation is presented, i.e. $h[k] = \delta[k]$. Furthermore, we aim at finding a weighting function $W\{\cdot\}$ that enhances the intelligibility of $\tilde{s}[k] + r[k]$ over $s[k] + r[k]$ under an equal power constraint.

Algorithm

In this section the proposed algorithm as schematically depicted in Figure 2 will be described. A more detailed description can be found in [5]. The proposed *DynComp* algorithm combines two time- and frequency-dependent stages. Namely, a time- and frequency-dependent amplification and a time- and frequency-dependent dynamic range compression. Both stages are controlled by an

*This work was carried out when the first author was with Fraunhofer IDMT, Project Group Hearing-, Speech-, and Audio Technology, Oldenburg, Germany

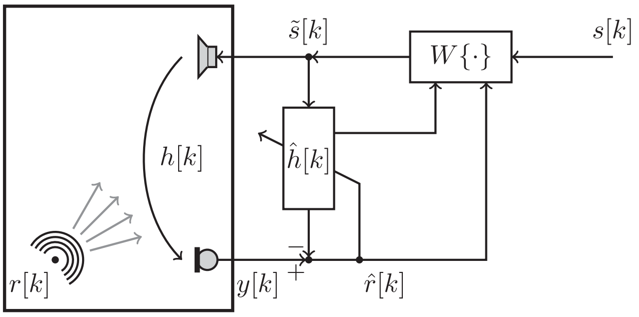

Figure 1: Considered acoustical scenario.

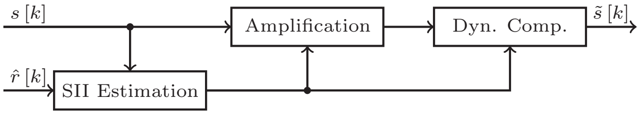

Figure 2: Schematic flow-graph of the proposed algorithm

SII-estimation and can therefore be considered as noise-adaptive. The time- and frequency-dependent amplification stage aims at improving the speech intelligibility by enhancing high-frequency content of the speech signal assuming that typical noises encountered in application scenarios have spectral content mainly in low- and mid-frequency regions. The dynamic range compression stage aims at boosting low-level contents of the speech signal to audibility and compressing frequency-bands with high SNRs that are assumed to be well audible.
In the remainder our algorithm is compared to the algorithm of Sauert and Vary [6] which only considers a time- and frequency-dependent amplification. Their goal is to achieve an SII-weighting shape of the speech spectrum in conditions of low SNRs and no change of the speech spectrum when the SNR is sufficiently high. Since we used a different filterbank than proposed in [6] and therefore a different set of parameters, we will denote this algorithm by *ModSau*.

Objective Measures

The proposed algorithm has been evaluated using objective measurements that had shown high correlations to speech intelligibility measured in formal listening tests in previous studies. Ten randomly selected German sentences from the Oldenburg Sentence test (OLSA) were used that were degraded by additive stationary speech-shaped noise (SSN) at SNRs ranging from -30 dB to +30 dB. SSN was created by randomly superimposing sentences from the sentence test, therefore yielding a longterm spectrum that is equivalent to the average spectrum of the speech material. The following measures were used to quantify the effect of the proposed *DynComp* algorithm:

- **STI:** [7] The Speech Transmission Index is based on the observation that a fully modulated signal experiences a reduction in modulation depth due to additive noise and/or reverberation present in a given transmission channel. This reduction in modulation highly correlates with measures of speech intelligibility.

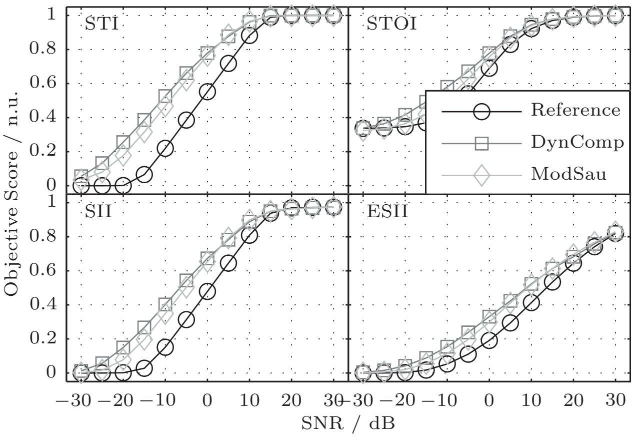

Figure 3: Results for the evaluation using different objective measures and speech shaped noise.

- **SII:** [8] The Speech Intelligibility Index can be considered, in a first approximation, as an intelligibility weighted SNR. In addition, several more effects are accounted for, such as the upward spread of masking and spectral smearing at high intensities.
- **ESII:** [9] The Extended SII is an extension of the original SII, that allows for time-dependent prediction of speech intelligibility. Therefore within the ESII, SII scores are calculated in short time windows. The overall score is calculated as the mean of all short-term predictions.
- **STOI:** [10] The Short Time Objective Intelligibility Measure employs the correlation between the clean speech signal and the disturbed speech signal as predictive measure of speech intelligibility.

The SNR-dependent results are shown in Figure 3. In all objective measures both algorithms yield an improvement compared to the unprocessed *Reference*. Although differences are small, there is a tendency that the proposed *DynComp* algorithm outperforms *ModSau* in most conditions. Note that the SII and ESII yield different values although there output values should be equivalent in case of stationary noises [9]. This is due to the use of the real speech signals instead of stationary noises derived from the speech signals as speech signal inputs to the objective measures.

Subjective Measures

Speech intelligibility testing was carried out with eight normal-hearing subjects, i.e. pure-tone thresholds of not larger than 20 dB HL. The mean age of the subjects was 25,9 years. Two different noises were used, a stationary car noise and a more instationary cafeteria noise. Speech material was taken from the Oldenburg sentence test [12]. During presentation the level of the speech signal was kept at 60 dB SPL and the noise signal was scaled to yield the desired SNR. In a preliminary study with four of the eight subjects for each noise type three different SNRs were determined using an adaptive procedure that yielded thresholds of approximately 20%, 50% and 80% word intelligibility in the unprocessed

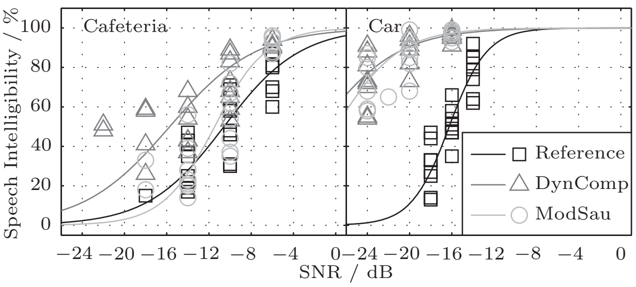

Figure 4: Results for listening test for the cafeteria noise (left) and car noise (right). Symbols indicate individual results and lines show psychometric functions obtained by parametric averaging of individual psychometric functions.

Reference condition. The following SNRs were obtained for the different noises:

- Car noise: −18 dB, −16 dB, −14 dB
- Cafeteria noise: −14 dB, −10 dB, −6 dB.

When speech intelligibility values over a wide range of SNRs are obtained, psychometric functions can be used to describe the intelligibility of SNRs not measured explicitly. A psychometric function can be fitted by varying the parameters SNR_{50} and s_{50}, that correspond to the SNR of 50% speech intelligibility and the slope of the function at that point, respectively., in e.g. [11]

$$P\,(SNR) = \frac{1}{1 + e^{4 \cdot s_{50} \cdot (SNR_{50} - SNR)}}. \tag{3}$$

From informal listening tests we expected the algorithms to increase speech intelligibility, especially for the stationary car noise. Thus, a good fit of the psychometric function is not guaranteed. Therefore, we used an semi-adaptive procedure to chose the SNRs based on listeners individual responses. This resulted in SNRs ranging from −24 dB to −14 dB and −22 dB to −6 dB for the car noise and cafeteria noise, respectively. A detailed description of this procedure can be found in [5]. Figure 4 shows the results for individual subjects as well as the psychometric functions computed by parametric averaging of the individual SNR_{50} and s_{50} values. For the car noise (right figure) both algorithm increase the speech intelligibility by up to 70% for the *DynComp* algorithm at an SNR of -24 dB. For this noise both algorithms achieve comparable results for the whole range of considered SNRs. For the instationary cafeteria noise at the highest SNR of -6 dB both algorithms show comparable results, while for the lower SNRs only the proposed *DynComp* algorithm improves speech intelligibility over the unprocessed *Reference.*

To test for significant effects an analysis of variance (ANOVA) was carried out for each noise condition. Since the SNRs measured differed across subjects and algorithms we chose those three SNRs that most measured values were available for. To estimate missing values at these SNRs the individual psychometric function were evaluated at these points. The resulting SNRs in the statistical analysis were −24 dB, −20 dB and −16 dB for car noise and −14 dB, −10 dB and −6 dB for the cafeteria noise.

Influence of Noise

To investigate the influence of the different noises a two-way ANOVA was carried out with factors of algorithm and noise. For both noises only the data corresponding to the SNR of 50% in the unprocessed *Reference* was used. According to Shapiro-Wilk test normality could be assumed for all data involved in this analysis. Results showed a significant influence of both factors and their interaction on speech intelligibility (Algorithm: $F\,(2;14) = 108.61$, $p < 0.001$, Noise: $F\,(1;7) = 23.25$, $p < 0.001$, Algorithm×Noise: $F\,(2;14) = 55.65$, $p < 0.001$). Post-Hoc tests were carried out using t-tests for dependent variables and Bonferroni correction for 9 comparisons, indicating a significant difference across noises for both algorithms, a significant improvement of both algorithms over the *Reference* condition for the car noise and a significant improvement of *DynComp* over the *Reference* and *ModSau* for the cafeteria noise.

Influence of Algorithms

To test the influence of processing in different SNRs, both noises were considered separately. A Shapiro-Wilk test showed that not all condition followed a normal distribution, thus an aligned rank transformation according to [13] was carried out prior to ANOVA. The resulting three data sets for the main factors SNR, algorithm and their interaction were then analysed separately. Results for the car noise showed a significant influence of both factors and their interaction (SNR: $F\,(2;14) = 149.34$, $p < 0.001$, Algorithm: $F\,(2;14) = 98.15$, $p < 0.001$, SNR×Algorithm: $F\,(4;28) = 37.21$, $p < 0.001$). Post-hoc analysis for the factor of algorithm showed that both *ModSau* and *DynComp* increase the speech intelligibility significantly over the over unprocessed *Reference.* For the cafeteria noise also both main effects as well as their interaction was significant (SNR: $F\,(2;14) = 108.64$, $p < 0.001$, Algorithms: $F\,(2;14) = 65.08$, $p < 0.001$, SNR×Algorithms: $F\,(4;28) = 6.75$, $p < 0.001$). A post-hoc analysis for the main effect of algorithm was carried out showing that *DynComp* significantly increased speech intelligibility over the *Reference* ($p < 0.001$) and *ModSau* ($p < 0.001$).

Prediction of Subjective Measures

The predictive ability of the objective measures used in the objective testing was investigated using correlation analyses of the objective and subjective data. Therefore, all sentences used in the subjective testing were also evaluated using the four different objective measures of speech intelligibility. To account for possible non-linear relationships between model predictions and listening test scores logistic function can be used to transform model values into speech intelligibility values. We used (4) as proposed by [14] and optimized the parameters m and b using the unprocessed *Reference*, while parameters a and c were given due to the boundary conditions of $P\,(0) = 0$ and $P\,(1) = 1$.

$$P\,(SII) = \frac{m}{a + e^{-b \cdot SII}} + c \tag{4}$$

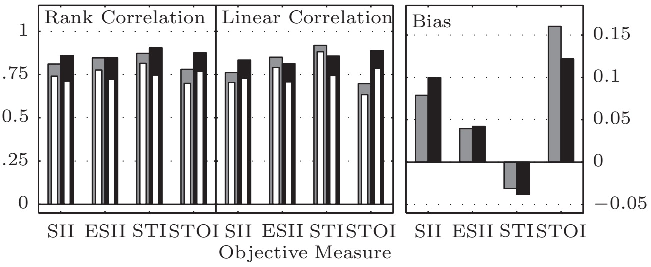

Figure 5: Results for the correlation analyses of the objective and subjective data. Gray bars indicate results for the car noise and black bars indicate results for the cafeteria noise. White inner bars show the correlation results based on the individual results while outer bars show results based on averaged values.

Figure 5 shows the results of the correlations analyses, i.e. Spearman's rank correlation (left), Pearson's linear correlation coefficient (middle) and the bias (right), i.e. the linear deviation from a perfect match. Note that in contrast to the objective evaluation, for the ESII a speech-shaped noise as a speech signal was generated as proposed in [9]. From the results it can be concluded that all models perform nearly equivalently in predicting the order of the results from the listening test. Also, in nearly all conditions a linear correlation $r > 0.75$ is achieved. Only STOI yields a lower predictive ability for the car noise and also exhibits the largest bias.

Conclusion

In this contribution a new algorithm for the processing of speech signals prior to their presentation was proposed. Objective and subjective evaluation was carried out showing that the proposed *DynComp* algorithm outperforms our implementation of the state-of-the-art algorithm by [6]. Gains in intelligibility of up to 70% for the entire range of considered SNRs could be observed. Furthermore, the predictive ability of several objective measures was investigated showing that all models are in general capable of predicting speech intelligibility when their respective bias is taken into account.

References

[1] Licklider, J. C. R. und Pollack, I. (1948). *"Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech"*. J. Acoust. Soc. Am., 20(1):42-51.

[2] Niederjohn, R. und Grotelueschen, J. (1976). *"The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression"*. IEEE Trans. Acoustics, Speech and Signal Processing, 24(4):277-282.

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

[4] Zorila, T.-C., Kandia, V. und Stylianou, Y. (2012b). *"Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression"*. In: Proceedings of Interspeech 2012 (Portland, USA).

[5] Schepker, H. (2012). *"Entwicklung und Evaluation von Vorverarbeitungsalgorithmen zur Verbesserung der Sprachverständlichkeit im Störgeräusch"*, Master Thesis, Carl-von-Ossietzky University Oldenburg, Oldenburg, Germany

[6] Sauert, B. and Vary, P. (2012). *"Near-end listening enhancement in the presence of bandpass noises"*. In: Proc. of ITG-Fachtagung Sprachkommunikation. (Braunschweig, Germany, Sept. 26-288, 2012).

[7] IEC, (2003). *"Sound System Equipment - Part 16: Objective rating of speech intelligibility by speech transmission index"*. International Standard IEC 60268-16 (International Electrotechnical Commission), Geneva, Switzerland.

[8] ANSI, (1997). *"Methods for Calculation of the speech intelligibility index"*. American National Standard ANSI S3.5-1997 (American National Standards Institute, Inc.), New York, USA.

[9] Rhebergen, K. S. and Versfeld, N. J. (2005). *"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., 117(4):2181-2192.

[10] Taal, C., Hendriks, R., Heusdens, R. and Jensen, J. (2011). *"An algorithm for intelligibility prediction of time frequency weighted noisy speech"*. IEEE Trans. Audio, Speech, and Language Processing, 19(7):2125-2136.

[11] Brand, T. and Kollmeier, B. (2002). *Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests.* J. Acoust. Soc. Am., 111(6):2801-2810.

[12] Wagener, K., Brand, T. and Kollmeier, B. (1999). *Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests"*. Zeitschrift für Audiologie/Audiological Acoustics, 38:86-95. 111(6):2801-2810.

[13] Wobbrock, J. O., Findlater, L., Gergle, D. und Higgins, J. J. (2011). *"The aligned rank transform for nonparametric factorial analyses using only anova procedures"*. In: Proceedings of the 2011 annual conference on Human factors in computing systems, CHI Â´11, S. 143-146, New York, NY, USA. ACM.

[14] Beutelmann, R. and Brand, T. (2006). *"Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners"*, J. Acoust. Soc. Am., 120(1):331-342.