Near-End Listening Enhancement: The Impact of Far-End Noise Reduction

Markus Niermann, Peter Vary, Peter Jax
Institut für Kommunikationssysteme, RWTH Aachen University, 52074 Aachen, Germany,
Email: {niermann,vary,jax}@iks.rwth-aachen.de

Abstract

The objective of Near-End Listening Enhancement (NELE) is to improve the speech intelligibility of a communication system at the receiving end (near-end) which is in a noisy environment. Common NELE approaches adaptively filter the received speech signal, taking into account the near-end environmental noise characteristics. Usually it is assumed that the received signal is noise-free. In this contribution we consider the situation that the speech from the transmitting side (far-end) is disturbed by far-end noise and that some noise reduction (NR) algorithm is applied there. We study the impact of the far-end NR on the overall performance of the concatenation of NR preprocessing and NELE postprocessing. It is shown that the blind concatenation of NR and NELE may produce severe performance losses and artifacts. Furthermore, first ideas are presented how to improve the interaction by modifications of the processing algorithms.

Introduction

We investigate the performance of a digital speech communication system with acoustic background noise at the transmitting side (far-end) as well as at the receiving side (near-end). At the transmitting side, some adaptive preprocessing algorithm for noise reduction (NR) is applied while at the receiving side some adaptive postprocessing is used to improve the intelligibility in terms of near-end listening enhancement (NELE) [1, 2]. The received signal may partly be masked by the near-end background noise. Since the near-end noise cannot be influenced by signal processing, NELE adaptively processes the received far-end signal exploiting knowledge about the near-end noise. There are several methods for enhancement such as spectral weighting in subbands, e.g., [3, 4, 5, 6, 7, 8], adaptive time domain filtering [9], and dynamic range compression [5]. In the literature, dynamical spectral weights are often determined by maximizing instantaneous intelligibility measures such as the Speech Intelligibility Index (SII) [4, 10, 8] or the mutual information [6, 7]. Several side constraints may be considered, e.g., upper bound for the total speech power, total amplification or power per subband. NELE was originally developed to be applied in mobile phones in the downlink. Further potential applications for NELE are hearing aids and public announcement systems, e.g., at railway stations. In the latter context, also the term "Speech Reinforcement" has been used [11].

Most previous works on NELE assume that the speech signal from the far-end is clean, i.e., unaffected by noise and other degradations. In many situations, however, this assumption is not appropriate. In the case of a telephone call between two mobile phones for example, the far-end speaker as well as the near-end listener might be located in noisy environments, thus the signal from the far-end is also degraded by noise. The authors of [7] propose a combination of beamforming for noise reduction (NR) and NELE by maximization of mutual information and conclude that a common consideration of near-end and far-end noise is necessary. In [12] it is discovered that a blind concatenation of single-channel NR and NELE leads to severe artifacts. An approach is proposed where the influence of NR and NELE is adaptively reduced in order to avoid those problems.

In this contribution, we study the impact of a far-end single-microphone NR on the overall performance of the concatenation of NR preprocessing and NELE postprocessing. The reason for the severe artifacts that result from a blind concatenation is further examined. After a theoretical analysis we present two approaches on how to improve the interaction. One of them comes along without reducing the influence of NR. Essential parts of this paper are already published in a PHD thesis [13].

System Model

The discussed communication system is visualized in Fig. 1. The microphone at the far-end records speech $s(k)$ from the far-end speaker as well as additive background noise $q(k)$, resulting in the noisy signal $y(k)$. In order to reduce the noise, the noisy signal is processed by a NR algorithm. Afterwards, the de-noised signal $\widetilde{y}(k)$ is transmitted to the near-end. A person at the near-end, listening to the loudspeaker signal $u(k)$, also perceives the near-end background noise $n(k)$ which may degrade intelligibility or may even cause unintelligibility. The intelligibility can be improved by NELE postprocessing, considering the characteristics of the near-end noise $n(k)$.

Both algorithms, NR and NELE, usually work in the frequency domain. Without limitations of the general analysis and conclusions, we consider frequency domain processing as the result may also be applied to time domain processing or a mixture of both. Thus, a frequency domain communication model is considered in the remainder of this work as shown in Fig. 2, assuming a lossless transmission. The frequency domain correspondence of a time domain signal is expressed using a capital letter with the frequency index i in the subscript. For simplicity the time dependency is omitted. The two algorithms modify the magnitude (respectively the real and imaginary part) of the complex-valued spectral representation of the audio signal by applying real valued gains G_i (NR) and W_i (NELE) while keeping the phase untouched.

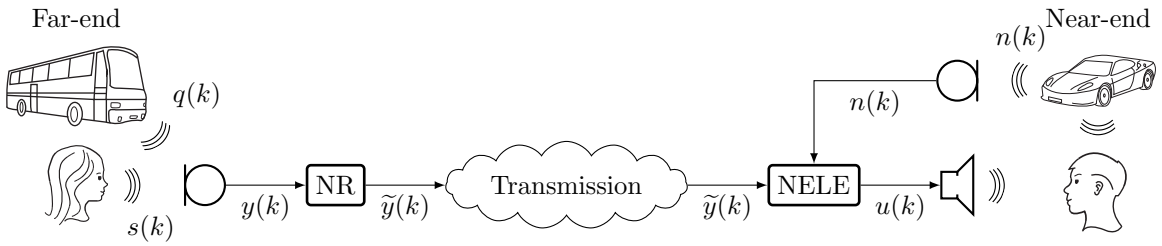

Figure 1: Model of a communication system with background noise at the far-end and at the near-end. There is no coupling between loudspeaker and microphone.

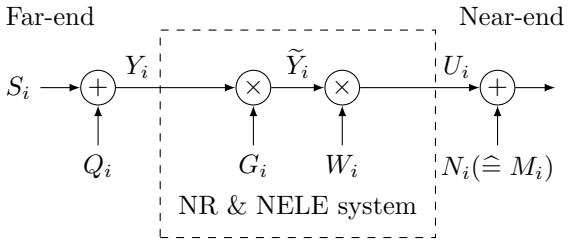

Figure 2: Frequency domain representation of the model in Fig. 1. The index i represents the frequency.

Conventional Near-End Listening Enhancement

In this study, we apply the NELE algorithm *Noise-Masking-Proportional Speech Shaping* (NoiseProp) [13], which uses a similar approach as the algorithm named *SNR Recovery* in [3, 8]. It targets at making speech well audible by raising it over the masking threshold, caused by the near-end noise. By this, the speech color is changed and the result does not sound as natural as the original speech. However, this is clearly accepted by listeners because of a higher intelligibility and lower listening effort. NELE algorithms work with, e.g., 21, subbands that are adapted to the human auditory system or with a corresponding higher number of uniformly spaced subbands. Hence, the frequency domain signals must be obtained by an appropriate subband transformation and the frequency indices i identify the individual subbands.

Previous NELE algorithms do not consider noise at the far-end. This can be seen as a special case of the model in Fig. 2 in which the far-end noise is zero ($Q_i = 0$) and the NR is disabled ($G_i = 1$). In this case the NELE algorithm is based on smoothed subband power estimates of the clean far-end speech S_i and near-end noise N_i by averaging the instantaneous power values over frames or by equivalent recursive averaging. The estimates are denoted by $\overline{|S_i|^2}$ and $\overline{|N_i|^2}$. The NELE NoiseProp algorithm will firstly calculate the near-end masking threshold

$$\overline{|M_i|^2} = \overline{|N_i|^2} + \sum_{\zeta=1}^{i-1} \overline{|N_\zeta|^2} \cdot C_\zeta^{\log_2 \frac{f_{\mathrm{c},i}}{f_{\mathrm{h},\zeta}}} \tag{1}$$

using a masking spreading function as in [14, 8]. The function

$$C_i = 10^{-8} \cdot (\overline{|N_i|^2} \cdot \Delta f_i)^{0.6}. \tag{2}$$

indicates a slope per octave. Obligatory parameters are the bandwidths Δf_i as well as the lower and upper frequency thresholds per subband $f_{\mathrm{l},i}$, $f_{\mathrm{h},i}$ and the center frequency $f_{\mathrm{c},i}$. Secondly, the weights

$$W_i = \begin{cases} 1 & \text{if } \overline{|s(k)|^2}/\overline{|n(k)|^2} > C \\ \sqrt{\frac{\overline{|M_i|^2}}{\overline{|S_i|^2}} \cdot A_{\mathrm{SMR}}} & \text{else} \end{cases} \tag{3}$$

are determined based on the masking threshold $\overline{|M_i|^2}$ and the far-end speech power $\overline{|S_i|^2}$ (for the frequency index i), using the parameter A_{SMR} as a target signal-to-masking ratio. This rule ensures that the speech is raised above the masking threshold. The ratio A_{SMR} defines the desired distance between speech and masking.

NELE is switched off if the total speech to near-end noise ratio is greater than the empirical threshold C.

Conventional Noise Reduction

Typical NR algorithms are implemented in frequency bands with a much higher frequency resolution than NELE. Usually a Discrete Fourier Transformation (DFT) is employed. However, in a first approach, for the following analysis it is assumed that the NR and NELE work in the same frequency resolution and the same frequency index i is used. The case of differing frequency resolutions is discussed later.

In contrast to the NELE algorithm, whose gain rule is based on strongly smoothed power values, NR gain rules should be based on power estimates that react fast to changing signals. In order to analyze the principal effects of NR and NELE, we assume exemplary a Wiener NR rule according to

$$G_i = \frac{\hat{\sigma}_S^2(i)}{\hat{\sigma}_Y^2(i)} \tag{4}$$

where $\hat{\sigma}_S^2(i)$ is a short-term estimate of the speech and $\hat{\sigma}_Y^2(i)$ is a short-term estimate of the noisy signal powers in frequency band i. Furthermore, it is assumed that state-of-the-art methods are used to reduce musical tones which are due to statistical estimation errors.

NELE and Far-End Noise

This section applies the NELE weighting rules to the signal at the far-end which is disturbed by noise. In the case of a noise-free far-end signal as considered in Sec. , the NELE gain rule depends only on the far-end speech

power $\overline{|S_i|^2}$ and the masking threshold $\overline{|M_i|^2}$, caused by the near-end noise. The masking threshold is deduced from the near-end noise $\overline{|N_i|^2}$. In this study, NELE is applied to the de-noised far-end speech $|\widetilde{Y}_i|$. The NELE gain W_i consequently depends only on $\overline{|\widetilde{Y}_i|^2}$ and $\overline{|M_i|^2}$.

The NoiseProp gain rule from (3) is reformulated as

$$W_i = \sqrt{\frac{\overline{|M_i|^2}}{\overline{|\widetilde{Y}_i|^2}} \cdot A_{\mathrm{SMR}}} \tag{5}$$

considering in the following for simplicity only the active NELE case, i.e., $\overline{|s(k)|^2}/\overline{|n(k)|^2} \leq C$.

NELE only

First, we consider a case without NR but with NELE. The absence of a NR is modelled by setting $G_i = 1$. As a consequence, it holds

$$\overline{|\widetilde{Y}_i|^2} = \overline{|Y_i|^2}. \tag{6}$$

Inserting this into the NELE gain rule yields

$$W_i = \sqrt{\frac{\overline{|M_i|^2}}{\overline{|\widetilde{Y}_i|^2}} \cdot A_{\mathrm{SMR}}} \tag{7a}$$

$$= \sqrt{\frac{\overline{|M_i|^2}}{\overline{|Y_i|^2}} \cdot A_{\mathrm{SMR}}}. \tag{7b}$$

Due to the lack of any NR, not only the far-end speech signal is elevated above the masking threshold and played back at the near-end, but also the far-end noise signal.

NR & NELE

NR is activated and the NR gain takes values $G_i \in [0, 1]$. The NR output is described as

$$\overline{|\widetilde{Y}_i|^2} = \overline{|Y_i \cdot G_i|^2} = \overline{|Y_i|^2} \cdot \widetilde{G}_i^2. \tag{8}$$

where $\widetilde{G}_i$ describes the *effective* gain with respect to the estimate $\overline{|Y_i|^2}$. The NELE gain rule is still the same as in (5). It depends only on information which is known to the NELE system, i.e., $\overline{|M_i|^2}$ and $\overline{|\widetilde{Y}_i|^2}$. The combined NR&NELE system (cf. Fig. 2) is characterized by an overall gain

$$K_i = G_i \cdot W_i = G_i \cdot \sqrt{\frac{\overline{|M_i|^2}}{\overline{|\widetilde{Y}_i|^2}} \cdot A_{\mathrm{SMR}}} \tag{9}$$

that consists of G_i and W_i. In order to describe the performance of the concatenated NR&NELE system as a function of the noisy input $|Y_i|$, the medium-term power $\overline{|\widetilde{Y}_i|^2}$ is expressed by using (8):

$$K_i = G_i \cdot \sqrt{\frac{\overline{|M_i|^2}}{\overline{|Y_i|^2} \cdot \widetilde{G}_i^2} \cdot A_{\mathrm{SMR}}} \tag{10a}$$

$$= \frac{G_i}{\widetilde{G}_i} \cdot \sqrt{\frac{\overline{|M_i|^2}}{\overline{|Y_i|^2}} \cdot A_{\mathrm{SMR}}}. \tag{10b}$$

The concatenated system is described by:

- the original NR gain G_i, based on *fast* changing short-term estimates,
- the NELE parameters $\overline{|M_i|^2}$ and $\overline{|Y_i|^2}$, based on *slowly* changing medium-term estimates,
- and the effective medium-term gain $\widetilde{G}_i$.

Interpretation and Performance

The factors $\overline{|M_i|^2}$, $\overline{|Y_i|^2}$, and A_{SMR} correspond to NELE postprocessing applied to the noisy signal without NR. The effective gain $\widetilde{G}_i$ describes the effect of the NR on the NELE algorithm actually applied to the preprocessed de-noised signal $\widetilde{y}(k)$. The alternative representation of the overall gain K_i according to (10b) suggests the following interpretations:

1. The NELE part of the concatenated system looks like operating on the noisy far-end signal Y_i as in (7b).
2. The NR part of the concatenated system looks like using the modified spectral weights $G_i/\widetilde{G}_i$.

The degree of modification of the original weights G_i is determined by the extent to which NELE averages the input signals. This relates to the degree of smoothing as well as to time synchronization in case of block averaging. We discuss two cases.

a) NR with short-term averaging, NELE without averaging.
Because of $\widetilde{G}_i = G_i$, the original NR gain G_i is cancelled out completely. The overall system performs like a NELE system applied to the noisy signal $y(k)$. Depending on the level and the spectral distribution of the far-end noise $q(k)$, it may be masked more or less by the near-end noise.

b) NR with short-term averaging, NELE with medium-term averaging.
Due to the stronger averaging, $\widetilde{G}_i$ – according to (8) – follows somewhat slower the original gain G_i. $\widetilde{G}_i$ compensates the average, but is sometimes smaller and sometimes larger than G_i. If $\widetilde{G}_i$ is smaller than G_i, the modified gain $G_i/\widetilde{G}_i$ becomes larger than one. Because of the statistical behavior, this increases significantly the musical tone effects and distortions.

These effects have been confirmed by listening experiments for the two cases.

a) The missing NR is not too critical as the perception of the far-end noise is reduced due to masking by the near-end noise. However, NR is effectively deactivated even if the near-end noise is weak.

b) The performance losses in terms of increased musical tones and artifacts may be severe.

In practical applications, only case b) is relevant as the type and the time constants of the NR algorithm are not

known at the receiving end and as it is not desirable to switch off the NR.

Improving the Interaction of NR and NELE: First Ideas

In seeking solutions to improve the case b) performance, we propose two trial approaches (cmp. also [13]):

1. **Reduction of the NR performance**
   The fluctuations of the quotient $G_i/\widetilde{G}_i$ in (10b) can be reduced by increasing the spectral floor of the NR algorithm which avoids that G_i and thus $\widetilde{G}_i$ fall below a certain limit. This reduces the amount of NR, but this would require some control of the NR algorithm from the receiving end.

2. **Joint NR and NELE processing at the receiving end**
   The fluctuations of the effective gain $\widetilde{G}_i$ in the NELE gain rule (10b) are identified to be the source of the problem. The reason is that NELE uses *medium-term averaging* of $\overline{|Y_i|^2}$ where $|\widetilde{Y}_i| = G_i \cdot |Y_i|$ is determined by the *short-term averaging* NR gain rule G_i. This effect can be eliminated by swapping the order of NR and NELE, i.e., applying NELE before NR. Thus, NR has to be moved from the transmitting end to the receiving end. Hence, NELE does not apply medium-term averaging of short-term averaged quantities. By running NELE on the noisy far-end signal $y(k)$, W_i looks like in (7b) and the term $\widetilde{G}_i$ does not occur. The NR is placed behind NELE. This is allowed, as the multiplication of the noisy signal with W_i does not change the SNR of frequency subband i. Therefore, the original NR gains G_i, derived from the noisy Y_i *before* NELE, can be applied to $\widetilde{Y}_i$ *after* NELE, resulting in the overall gain rule

$$K_i = W_i \cdot G_i = G_i \cdot \sqrt{\frac{\overline{|M_i|^2}}{\overline{|Y_i|^2}} \cdot A_{\mathrm{SMR}}}. \qquad (11)$$

Informal listening tests confirmed that both counter measures reduce the musical tones and the artifacts, where clearly better results are obtained with joint NR and NELE processing at the receiving end.

Conclusions

In digital speech communication systems which are used in noisy environments at the transmitting and at the receiving end, NR is applied at the transmitting side in terms of preprocessing and NELE is applied at the receiving side as postprocessing. In this paper it is shown that in a blind concatenation of NR and NELE, the NELE postprocessing counteracts the NR preprocessing such that performance losses in terms of increased musical tones and artifacts are observed. In the extreme case the effect of NR is completely cancelled out. Two first proposals for improving the interaction of NR and NELE are described by either reducing the NR performance or by moving NR to the receiving end for joint NELE (first) and NR (second) processing. For simplicity, the investigations were carried out under the assumption that NR and NELE use identical spectral resolution. The principal described effects are also observable if different frequency resolutions are applied.

References

[1] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, 2015.

[2] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, pp. 572–585, May 2013.

[3] B. Sauert and P. Vary, "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toulouse, France), pp. 493–496, IEEE, May 2006.

[4] B. Sauert and P. Vary, "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, (New York, NY), pp. 1844–1848, Hindawi Publ., Aug. 2009.

[5] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, pp. 2075–2079, Aug. 2012.

[6] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, 2015.

[7] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 654–658, IEEE, 2016.

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

[9] M. Niermann, C. Thierfeld, P. Jax, and P. Vary, "Time domain approach for listening enhancement in noisy environments," in *Proc. of ITG-Fachtagung Sprachkommunikation*, pp. 282–286, VDE Verlag GmbH, Oct. 2016.

[10] B. Sauert and P. Vary, "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, pp. 1919–1923, EURASIP, Aug. 2010.

[11] J. B. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 54–66, 2014.

[12] M. Niermann, P. Jax, and P. Vary, "Joint near-end listening enhancement and far-end noise reduction," in *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5970–5974, IEEE, 2017.

[13] M. Niermann, *Digital Enhancement of Speech Perception in Noisy Environments*. Dissertation, IND, RWTH Aachen, Mar. 2019.

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