Quality Aspects of Near-End Listening Enhancement Approaches in Telecommunication Applications

Robin Pricken[1], Marcel Wältermann[1], Eva Parotat[1], Michał Sołoducha[2], Alexander Raake[2]
[1] AVM GmbH
[2] TU Ilmenau, Audiovisual Technology Group

Abstract

To increase speech intelligibility in loud environments, various approaches for near-end listening enhancement (NELE) were published in the last years. Different studies dealt with comparisons of these approaches with regard to speech intelligibility. Speech codecs are usually not applied prior to enhancing the speech material in these studies. However, in modern telecommunication applications, which is the major use case for NELE techniques, the audio signal is always encoded at the far end and decoded at the near end, leading to audible degradation. While NELE algorithms mainly aim at increasing the speech intelligibility, little is known about the quality of the processed speech. In this paper we evaluate the listening speech quality from four different NELE algorithms with changing noise conditions and take the typical hearing situation into account. The utilized speech signals were encoded and decoded with different codecs typically used in telecommunication systems before applying the NELE algorithms.

Introduction

Speech intelligibility decreases significantly in loud environments. Typically, each of us is confronted with such situations in public areas, for example on train stations or airports. In order to tackle this problem, various approaches of near-end listening enhancement (NELE) techniques were published in the last years, with the objective of increasing the intelligibility of speech in modern telecommunication scenarios.
Different approaches were compared in large-scale listening tests regarding speech intelligibility [1, 2]. However, ideal unmodified speech signals were used as input to the different NELE algorithms. So, the constraints of modern telecommunication systems were neglected in these studies. Usually, the speech signal is already audibly degraded by codecs which are typically applied to the transmitted signal for bit rate reduction and sometimes also for packet loss concealment. Furthermore, it was not taken into account that the typical hearing situation with a handheld device is monotic with a dichotic noise presentation as proposed by Park et al. [3].
Our work focuses on investigating the quality impact of four different NELE algorithms on speech signals that are processed with the following codecs recommended by the ITU-T: The wideband (WB) codec G.722 and the narrowband (NB) codecs G.711 or G.726. These codecs are widely used in telecommunication systems.
The speech recordings used in this work are interfered by two noise types at different levels: cafeteria and street. The NELE-processed signals were assessed in a listening test based on ITU-T Rec. P.800.
In the following, the different algorithms are introduced. Subsequently, the preparation and design of the listening test is explained and the results of the test are presented. Finally, the results are interpreted and summarized in the last section.

Algorithms

The speech signals were processed with four different algorithms. These algorithms can be split into two groups: noise-adaptive algorithms that process the speech signal in regard to the current noise, and noise-independent algorithms that always process the speech signal in the same way and do not take into account the current noise situation. Three of the considered algorithms are noise-adaptive while one is noise-independent. All of them yield equal energy input and output and their primary goal is to increase speech intelligibility and not speech quality.

AdaptDRC

The algorithm AdaptDRC (adaptive dynamic range compression) is a noise adaptive algorithm [4]. It uses an amplification stage and a DRC stage that are both time- and frequency dependent. The two stages are controlled by a short-term Speech Intelligibility Index (SII) estimation to measure the current speech intelligibility. The amplification aims to increase the power of high-frequency regions, while the DRC aims at amplifying low-level signals that are assumed to be barely audible.

InverseNoise

The basic idea behind the InverseNoise algorithm is to attenuate speech in frequency bands with a high disturbance, where it has no benefit for intelligibility, and redistribute this energy into less disturbed bands. To achieve this, it estimates the power of the noise and speech signal for each sub-band. The energy of the speech signal is redistributed according to a simple gain rule depending on the spectral energy distribution of the speech and noise signal [5].

SelBoost

The algorithm SelBoost (Selective Boost) introduced by Tang and Cooke [6] is noise-dependent, like the two algorithms mentioned before. It is motivated by the fact

that some time-frequency regions have a sufficient SNR for audibility while others are near to the threshold of audibility. Furthermore, some frequency regions are more important for intelligibility than others.
Thus, the algorithm calculates the local SNR in the time and frequency domain and shifts the signal energy from areas with high SNR to areas with low SNR, taking into account high and low importance frequency regions.

modSSDRC

The modSSDRC (modified SSDRC) algorithm is a self-implemented version of the SSDRC (spectral shaping dynamic range compression) algorithm by Zorila et al. [7]. In contrast to the other three algorithms, it is noise-independent. It consists of two stages: spectral shaping – similar to the Lombard effect – and dynamic range compression in the time domain to increase audibility of less sonorant parts.
The first two parts of the spectral shaping stage sharpens the formants and flattens the spectral tilt, depending on the probability of voicing. Furthermore, the last part is a pre-emphasis filter which boosts higher frequency regions that are more important for intelligibility and decreases lower frequency regions that are less important. This is performed independently from the voicing index. The filter considers only the magnitude instead of taking the phase of the signal into account, which is why a linear phase was added. The output of the SS-stage is then used as the input of the DRC-stage.
The modifications of the original SSDRC algorithm also comprise different parameters of the DRC-stage, an adopted filter design in the SS-stage and a packet-wise processing of the speech signal.

Test Preparation

Multiple processing stages are applied to the original speech and noise recordings, to synthesize the final test sample. The different steps are explained in the following two paragraphs and are depicted in Figure 1.

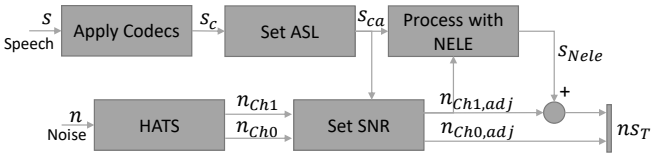

Figure 1: Block diagram of the audio signal preparation process

Preprocessing

In our study, we used audio recordings from the "German Lombard Speech Data Base" [8] as speech references s. There are two speakers, one female and one male. The records are phonetically balanced and have an average length of 9 seconds. For each speaker, 40 different sentences are available. The active speech level (ASL) was set to -26 dBov (s_{ca}) for all speech signals, after one of the ITU-T codecs G.711, G.722 or G.726 had been applied, using software tools from ITU-T Rec. G.191 [9].

Two different noises were used to mask the speech, street and cafeteria noise, where the street noise is more stationary than the cafeteria noise. The noise signals were recorded with artificial ears of a head and torso simulator (HATS) to simulate a real-life listening situation. The right ear (n_{Ch1}) of the HATS was covered with a handset and the other ear (n_{Ch0}) was left open during recording. The setup was in accordance with ETSI TS 103 224 [10].

Mixing of speech and noise

Prior to processing the speech signals with the three noise adaptive algorithms, the signal-to-noise ratio (SNR) has to be set. For this purpose the noise level was modified to achieve four different target SNRs $SNR_t \in \{0dB, 5dB, 10dB, \infty\}$. The last case is a special case: No noise is added to the processed speech signal later. However, the noise-adaptive algorithms need an input also in this case, so the authors chose a speech-shaped noise (SSN) with an SNR of 35 dB.
To set the correct level of the noise signals, the SNR_o of the original noise signals is calculated for each noise channel $i \in \{0, 1\}$ using the ASL of s_{ca} and the root mean square (RMS) of the noise as shown in Equation (1):

$$SNR_o(i) = ASL(s_{ca}) - 20 \cdot log_{10}\left(RMS(n_{Ch(i)})\right) \quad (1)$$

With the SNR_c for each channel and the desired SNR_t, the noise level is adjusted as shown in Equation (2). The level difference $\alpha \in$ {street: 2.2 dB, cafeteria: 1.5 dB} between the two channels, which is caused by the recording setup with the handheld device, has to be taken into account, too.

$$n_{Ch(i),adj} = \frac{n_{Ch(i)}}{10^{(SNR_t+\alpha-SNR_o(i))/20}} \quad (2)$$

As the next step, the different NELE algorithms can process the signals with s_{ca} as the speech and $n_{Ch1-adj}$ as the noise reference. The processed speech signal s_{Nele} is added to $n_{Ch1-adj}$ as shown in Equation (3). Both channels then form the final test sample ns_T, with a monotic speech presentation and a dichotic noise presentation.

$$\begin{aligned} ns_T(i=1) &= n_{Ch(i),adj} + s_{Nele} \\ ns_T(i=0) &= n_{Ch(i),adj} \end{aligned} \quad (3)$$

Test design

In the actual test, 210 different test samples were presented in individual random order to each participant. Each sample is characterized by a unique combination of the column values, as can be seen in Table 1.

40 different sentences are available for each speaker, as mentioned in a previous section, so it is not possible to use a different sentence for each combination. Thus, every combination was processed with every sentence. Then one random sentence is picked for each individual test and combination.
All participants read the same instructions, which explain the general test procedure and the listening situation: A telephone call in loud environment where the

NELE–Alg.	Noise	SNR	Codec	Speaker
None	No noise	∞	G.711	male
AdaptDRC	Street	10 dB	G.722	female
inverseNoise	Cafeteria	5 dB	G.726	-
SelBoost	-	0 dB	-	-
modSSDRC	-	-	-	-

Table 1: Overview of the test sample characteristics

speech signal is presented on one ear, and noise on both ears. In addition, pictures showing typical environmental noise scenarios underlaid with the noise were presented before the actual test.
The participants were told to rate each sample on the continuous rating scale shown in Figure 2, according to the question: 'How do you rate the overall quality of the telephone call listened to?'. Due to the wide range of different characteristics, the authors chose this scale to achieve a finer granularity in the ratings.
For playback of the different test samples Beyerdynamic headphones DT290 were used. The sound pressure level was set to 73 dB and the headphones were diffuse-field equalized.

Figure 2: Continuous rating scale according to Bodden and Jekosch [11], German version. English translations (left to right): 'extremely bad', 'bad', 'poor', 'fair', 'good', 'excellent', 'ideal'.

Listeners

In total 33 persons, 15 female and 18 male, participated in the listening test. The average age was 24.67 years with 18 years as the youngest and 46 years as the oldest participant. They got paid for participation, and all of them reported normal hearing.

Results

Figure 3 shows the results of the listening test. The given values from 0 to 6 are nonlinear normalized after Köster et al. [12] to the mean opinion score (MOS). The plotted error bars show 95% confidence intervals. Mean values were calculated over the two noise types, cafeteria and street, as well as the two speakers, male and female. The different characteristics of SNRs, NELE algorithms, and codecs are shown individually.
As can be seen from Figure 3, the overall quality decreases with decreasing SNR. Also, as can be expected, the WB codec is better rated than the NB codecs. From the reference curve, where no NELE is applied, shown in black and listed as 'None', it can be seen that the different NB codecs G.711 and G.726 do not have a significant impact if noise is added to the signal, although G.726 is rated lower than G.711 in the noise-free cases. Furthermore, G.722 with 0 dB SNR is nearly equally rated to the two NB codecs with 5 dB SNR.
The SelBoost algorithm, highlighted in green, does not perform as well as the other algorithms in terms of MOS, especially in the case without noise. The overall rating gets better with lower SNRs, but in most of the cases it is lower than the reference.
AdaptDRC, shown in blue, yields the highest MOS ratings in most cases. Without noise, its is approximately equal to the reference. Furthermore, in most cases the samples with G.726 are equal or, with lower SNR, even higher rated than G.711.
The algorithm inverseNoise is, apart from the condition with G.726, as good as the reference without noise. It is the only approach that did not benefit from the WB codec with lower SNRs. The quality decreases nearly linearly from 5 dB to 0 dB over the different codecs, while the other algorithms, as well as the reference, shows a peak for the WB codec for all SNR levels.
The only noise-independent algorithm modSSDRC, plotted in red, was not rated as good as the other algorithms in the absence of noise or an SNR of 10 dB, apart from SelBoost. It performs nearly as good as AdaptDRC at lower SNRs of 5 dB and 0 dB. Like most of the other algorithms, modSSDRC benefits from the WB codec. Here. the increase in quality is higher than with the NB codecs.

In summary, the quality of a speech signal can be increased by the use of NELE approaches in noisy conditions – a fact that is not necessarily obvious, as the main goal of NELE is increasing speech intelligibility and not overall quality. The AdaptDRC algorithm performed as one of the best regarding the overall speech quality. The only noise-independent algorithm, modSSDRC, got decent ratings with lower SNRs as well.
In this comparison it has to kept in mind, that the algorithms were developed to increase intelligibility and it has been proven in respective studies that this objective has been fulfilled. From this perspective, a decrease in quality is acceptable to some extent and also anticipated.

Conclusion

This study compared different NELE algorithms and their impact on the overall speech quality in a formal listening test. We set up a processing sequence to achieve a realistic hearing situation with different codecs and under various noise conditions. It is shown that speech quality can be increased by NELE algorithms under certain circumstances.

Acknowledgment

The authors would like to thank Jan Rennies–Hochmuth from Fraunhofer IDMT and Markus Niermann from RWTH Aachen for processing speech samples with their algorithms, as well as Stefan Bleiholder from HEAD acoustics GmbH for providing the noise recordings.

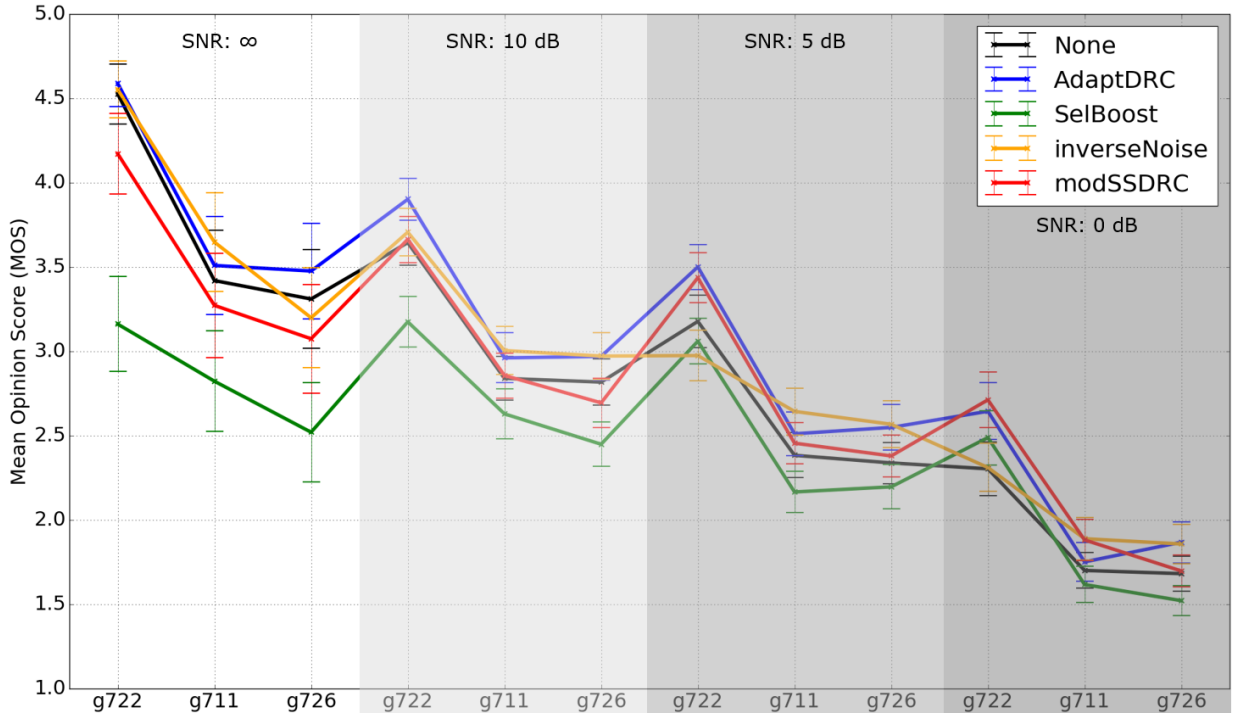

Figure 3: MOS values with changing codecs, SNRs and algorithms. Error bars represent 95% confidence interval.

References

[1] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The hurricane challenge," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3552–3556, 2013.

[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, no. 4, pp. 572–585, 2013.

[3] H. Park, J. Y. Yoon, J. H. Kim, and E. Oh, "Improving perceptual quality of speech in a noisy environment by enhancing temporal envelope and pitch," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 489–492, 2010.

[4] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 138, no. 4, pp. 3577–3581, 2015.

[5] M. Niermann, P. Jax, and P. Vary, "Near-end listening enhancement by noise-inverse speech shaping," *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 2390–2394, 2016.

[6] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," *Interspeech*, vol. 0, no. 3, pp. 1636–1639, 2010.

[7] T. C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," *Interspeech*, 2012.

[8] M. Sołoducha, A. Raake, F. Kettler, and P. Voigt, "Lombard speech database for German language," *https://zenodo.org/record/48713*, 2016.

[9] ITU-T, "Recommendation G.191," 2010.

[10] ETSI, "A sound field reproduction method for terminal testing including a background noise database," *Etsi*, vol. TS 103 224, 2014.

[11] M. Bodden and U. Jekosch, "Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität," *Project report, Institute of Communication Acoustics, Ruhr-University Bochum*, 1996.

[12] F. Köster, D. Guse, M. Wältermann, and S. Möller, "Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech," *DAGA*, pp. 150–153, 2015.