Auditory Evaluation of Receive-Side Speech Enhancement Algorithms

Jan Reimes, Günter Mauer, H.-W. Gierlich
HEAD acoustics GmbH, 52134 Herzogenrath, Germany, Email: telecom@head-acoustics.de

Introduction

Communication in noisy situations may be extremely stressful for the person located at the near-end side. Since the background noise is originated from the natural environment, it cannot be reduced for the listener. Thus the only possibility to improve this scenario with support of digital signal processing is the insertion of speech enhancement algorithms in the down-link direction of terminals.

Some of these methods are already integrated in modern state-of-the-art mobile devices. Such algorithms target in general on the improvement of listening comfort on the near end. Methods like (artificial) bandwidth extensions (BWE) or additional noise reduction are already quite common. Additionally, more sophisticated enhancement algorithms manipulate the speech signal with respect to the instantaneous local background noise estimation. The focus here is to improve speech intelligibility. Such methods are also known as speech reinforcement, intelligibility or near-end listening enhancement (NELE).

Whenever speech processing is inserted into a conversation, quality aspects must also be regarded. So far there are no suitable instrumental methods for the assessment of quality and intelligibility of acoustically captured speech signals in the presence of near-end noise. Thus, auditory assessments are currently the only method for performance evaluation.

This contribution presents a new proposal for the auditory evaluation for this use-case in order to evaluate also the trade-off between speech quality and intelligibility.

Motivation

All kinds of signal processing in down-link direction cause some challenges for measurement technology. Devices may behave differently (i.e. non-deterministic) in several noise scenarios or react noise-dependent. Especially when inserting additional gain to the loudspeaker signal, loudness rating requirement can be violated.

Similar, tuning a device towards intelligibility may also decrease speech quality requirements, which are often specified as mandatory. Especially in silent or almost noise-free scenarios, such algorithms should not manipulate the speech signal excessively.

On the other hand, intelligibility in extremely noisy scenarios may be much more important than speech quality aspects (hands-free or eCall). Nevertheless, in all SNR conditions or degradation ranges, a satisfactory balance between speech quality and listening effort is desirable from the user's point of view.

Test Corpus

The first stage of the test corpus of the auditory evaluation was the acoustical noise recordings of the near-end listener. For that purpose, a mockup device was mounted at right ear of head and torso simulator (HATS). With standard 8N application force, a typical leakage was realized. The left ear remained uncovered for the binaural recording.

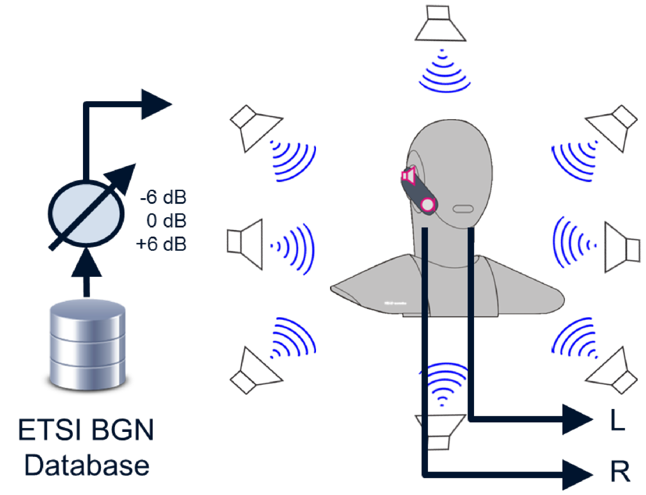


Figure 1: Setup of binaural noise recording procedure

A background noise playback system according to [1] with an 8-speaker-setup according to figure 1 was then used to reproduce a realistic sound field around the HATS. Four standardized noises were evaluated:

- Inside Car Noise - Full-size car 130 km/h
- Public Places Noise - Cafeteria
- Outside Traffic Street Noise - Road
- Public Places Noise - Train station

Each recording was played back with the realistic level. Two additional gains $+6dB$ and $-6dB$ were applied to each scenario to obtain a wider range of noise levels. Finally, silence condition (idle noise $< 30dB(A)$) was also taken into account.

The speech material used for processing consists of eight German fullband sentences according to [2]. Since the the current work was also intended to evaluate the performance of the included speech enhancement algorithms, only offline simulations and no real devices were used. Figure 2 shows the flow chart of the processing chain.

In a first processing step, the original speech material is pre-filtered and down-sampled to narrowband and wideband. Then, encoding and decoding of the widely used

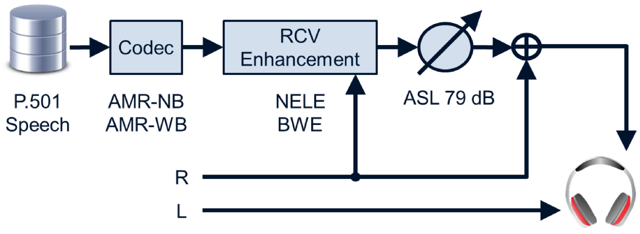


Figure 2: Offline BWE/NELE processing chain

Score	Listening Effort	Speech Quality
5	No effort required	Excellent
4	No appreciable effort required	Good
3	Moderate effort required	Fair
2	Considerable effort required	Poor
1	No meaning understood with any feasible effort	Bad

Table 1: Auditory scales for combined assessment

adaptive multi-rate codec (AMR-NB [3] / AMR-WB [4]) is applied.

If applicable, the right ear signal is used as an additional input for the speech signal enhancement (here: NELE). After this step, the active speech level is normalized to 79 dB SPL according to [5]. Especially common NELE algorithms utilize the maximum possible and allowed speech level. Here only the impact of sound manipulation should be regarded, thus all possibly occuring level differences are equalized. The resulting signals are assumed to be the output of a mobile phone without further degradations, i.e. neglecting non-linear speaker distortion or any arbitrary transfer function.

Overall, nine NELE algorithms (eight for WB, one for NB), two BWE methods, two combinations of both and processing with AMR-NB and AMR-WB only were included per background noise/gain set.

Finally, the signal of the right artificial ear is mixed with the processed speech. By combining this signal with the left ear signal of the unprocessed background noise, a binaural stimulus is created for the listening test.

Auditory Testing

The test corpus presented in the previous section was created with speech material which has already been used for speech quality evaluations, but is not intended for intelligibility tests in any way. Thus listening effort is utilized here as a alternative way to assess intelligibility. The main difference here is that opinion scores (listening effort) are compared against intelligibility indices which are assessed by auditory test like e.g. rhyme tests. As already discussed in [6], both dimensions are correlated and provide a reasonable estimator for each other if the test design is chosen properly.

Attributes for the assessment of listening effort and speech quality are already defined in ITU-T P.800 [7] and are used for the current work. However, classical auditory tests according to [7] usually collect only one mean-opinion score (MOS) per sample. A new proposal for a combined assessment test design similar to ITU-T P.835 [8] is introduced in this work.

Within the auditory test, both attributes according to table 1 are prompted during the presentation of the sample. Similar to [8], test subjects listen twice to each sample and provide a rating after each presentation, first for listening effort, second for speech quality. Test subjects were instructed to concentrate on the specified attribute for each playback.

One presented sample of 8.0s duration included two sentences of one talker. Thus, four samples per condition are obtained. In overall, 197 conditions with 788 different samples were auditorily evaluated with the proposed test design. 56 test subjects participated in the evaluation. Each participant listened to one sample per condition, which lead to 56 votes per condition or 14 votes per sample (for listening effort and speech quality).

Auditory Results - NELE

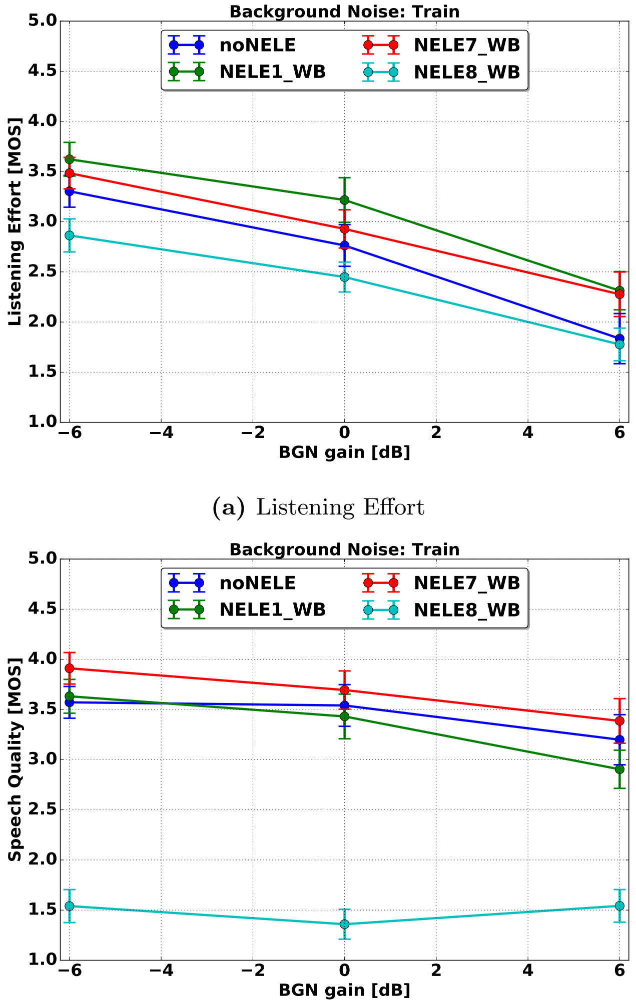
Figure 3 and 4 illustrate some selected results from the auditory evaluation. For the sake of convenience, not all NELE algorithms are shown in the graphs.

NELE8 is a simple algorithm [9] and thus results in poor speech quality, which is substantiated by the curves shown in figure 3b and 4b. These results indicate that large signal degradations may also impact listening effort and/or intelligibility. More sophisticated algorithms like NELE1 or NELE7 show a clear improvement of the listening effort for the loud and complex scenario train station (see figure 3a) while speech quality is only slightly decreased (see figure 3b). In this situation, NELE7 performs even better with reagrd to speech quality than the AMR-WB-only processing.

This behavior seems to be explainable by psychoacoustical masking effects of the background noise. In other noise scenarios with lower level and masking, i.e. car noise, test participants are able again to judge AMR-WB as best possible quality here (see figure 4b). Concerning the listening effort, only a slight preference for the NELE algorithms is observable. This may be explained by the reduced possibilities of the algorithms to increase high frequency components, since even the level-increased car noise has not much energy in the important higher frequency bands.

Auditory Results - BWE

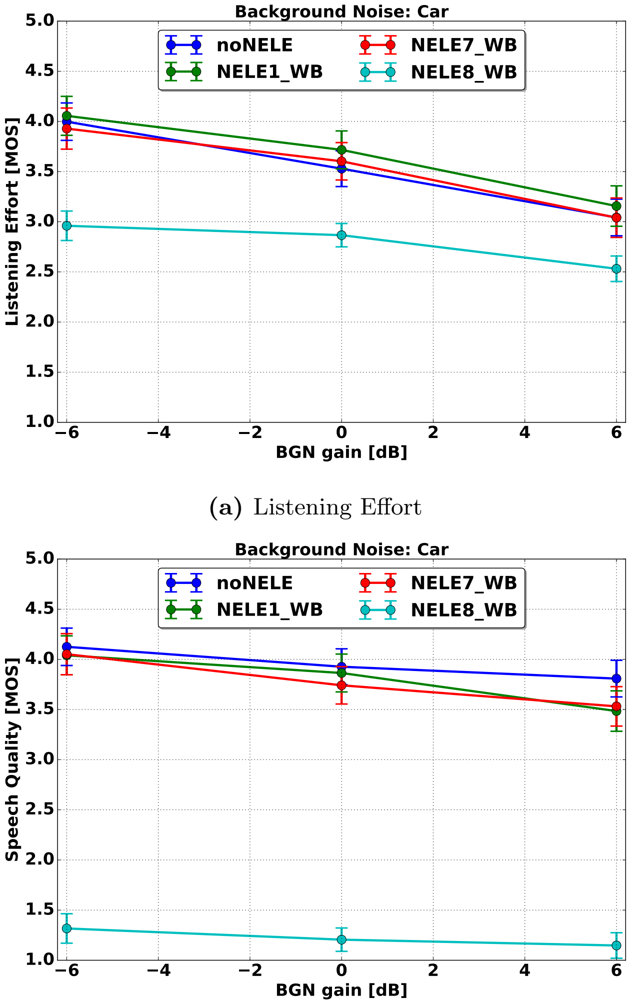
For BWE algorithms, no improvement regarding listening effort nor speech quality can be observed. Since the auditory results are similar for each background noise, figure 5a and 5b show the averaged results over all noises (excluding silence conditions). It is visible that concerning speech quality, AMR-WB is clearly preferred by the



(a) Listening Effort

(b) Speech quality

Figure 3: Auditory results for train station noise



(a) Listening Effort

(b) Speech quality

Figure 4: Auditory results for car noise 130 km/h

test subjects compared to any BWE/NELE combination. Concerning listening effort, also the combined methods of NELE and BWE do not exceed the unprocessed AMR-NB or AMR-WB test cases.

Auditory Results - Silence conditions

In order to check consistency and the borderline case of silence condition, figure 6 shows the auditory results of this special listening situation. It is obvious that listening effort reaches maximum scores $MOS_{LE} > 4.5$ in all cases, independent of the bandwidth. For speech quality, pure WB case obtains $MOS_{SQ} \approx 4.5$ in contrast to AMR-WB with $MOS_{SQ} \approx 4.2$. These are typical scores which can also be found in literature. Similar, the speech quality rating of pure NB ($MOS_{SQ} \approx 3.4$) and AMR-NB ($MOS_{SQ} \approx 3.2$) are consistent compared to literature when dealing with mixed-mode auditory experiments.

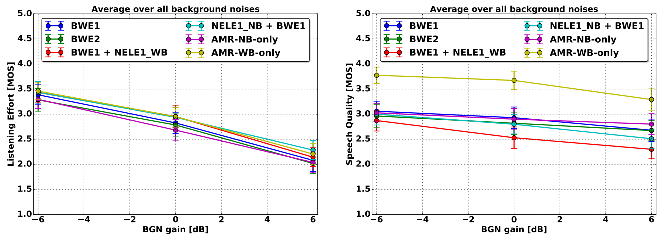
Also in silence conditions, both BWE algorithms clearly fail to reach WB scores ($MOS_{SQ} \approx 3.2$) and even do not exceed AMR-NB significantly in this case. At least for the implementations used here, the well-known audible artifacts of BWE were weighted stronger by the subjects than the increased bandwidth.

Auditory Results - Orthogonality

To check the orthogonality of both attributes (speech quality vs. listening effort), figure 7 provides comparison for all 197 condition ratings by means of a scatter plot. The correlation coefficient according to Pearson is determined to $r_{Pearson} = 0.52$, which indicates at least a minor correlation. This can be explained by the fact that good speech quality ratings (i.e. $MOS_{SQ} > 4.5$) cannot be expected for very low listening effort scores (i.e. $MOS_{LE} < 1.5$). On the other hand, even in silent or noise-free situations (i.e. $MOS_{LE} > 4.5$), a bad speech quality (i.e. $MOS_{SQ} < 1.5$) affects also the perceived listening effort.

Acknowledgment

The offline processing of NELE and BWE algorithms was kindly provided by the Institute of Communication Systems of RWTH Aachen.



(a) Listening Effort for BWE algorithms (b) Speech quality for BWE algorithms

Figure 5: Auditory results averaged across noises

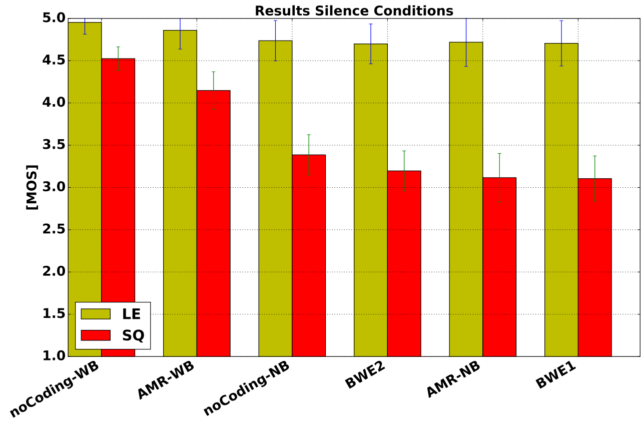


Figure 6: Listening effort and speech quality for silence conditions

Conclusions

In this contribution, a large test corpus was designed and auditorily evaluated. A new auditory test method was proposed and successfully introduced for the presented material.

The auditory results of the evaluation clearly show the impact of receive-side signal processing. NELE algorithms help to improve listening effort, but may slightly reduce speech quality. At least in this work, BWE algorithms (and possible in combination with NELE), cannot help to significantly improve listening effort. The performance of the included BWE algorithms regarding speech quality is similar to NB.

In a next step, the impact of varying active speech level (here: fixed at 79 dB SPL) should be analyzed since not all devices target at a constant listening level. Additionally, common NELE algorithms utilize the maximum possible and allowed speech level.

The proposed auditory testing can be applied to scenarios, especially where listening effort is crucial. Possible applications are in-car communication, hands-free and hand-held devices as well. In particular, eCall modules can benefit a lot from NELE algorithms.

Finally, since no practical evaluation procedures are available neither for listening effort nor speech quality, an instrumental method for the combined assessment of listening effort and speech quality is strongly desirable.

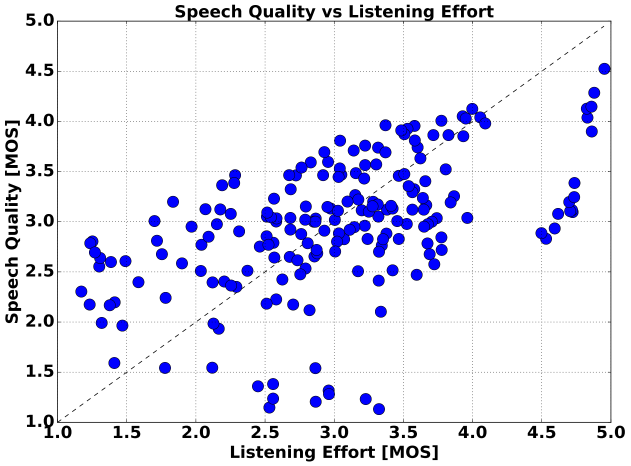


Figure 7: Speech Quality vs. Listening Effort

References

[1] European Telecommunications Standards Institute. *Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database*, August 2014.

[2] ITU-T Recommendation P.501. *Test signals for use in telephonometry*, Jan. 2012.

[3] 3GPP TS 26.071. *Mandatory speech CODEC speech processing functions; AMR speech Codec; General description*, Dec. 2009.

[4] ITU-T Recommendation G.722.2. *Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*, Jul. 2003.

[5] ITU-T Recommendation P.56. *Objective measurement of active speech level*, Dec. 2011.

[6] Jan Reimes. Listening effort vs. speech intelligibility in car environments. In *Fortschritte der Akustik - DAGA 2015*. DEGA e.V., Berlin, 2015.

[7] ITU-T Recommendation P.800. *Methods for subjective determination of transmission quality*, Aug. 1996.

[8] ITU-T Recommendation P.835. *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, Nov. 2003.

[9] Russell J Niederjohn and J Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):277–282, 1976.