Listening Test to Determine the Mismatch Between Signal-To-Noise Ratio and Human Perception

Simon Graf[a,b], Anne Theiß[b], Tobias Herbig[a] and Gerhard Schmidt[b]

[a] *Nuance Communications Deutschland GmbH, E-mail: simon.graf@nuance.com*

[b] *Christian-Albrechts-Universität zu Kiel, Germany, E-mail: ath@tf.uni-kiel.de*

Abstract

Evaluations of speech enhancement systems are typically based on artificially generated noisy speech signals. A common approach to quantify the weighting of speech and background noise is the signal-to-noise power ratio (SNR). In contrast to the perception of human listeners the SNR is calculated based on the power of speech and noise signals separately, irrespective of their spectral distributions. For this contribution, listening tests were performed to determine the influence of the spectral distribution of noise on the audio impression of human listeners. Based on our experimental results, we evaluate objective measures and their capability to predict the subjective rating.

Introduction

Evaluations of speech enhancement systems, such as hands-free telephony or in-car communication, are typically based on artificially generated noisy speech signals. A clean speech signal $s(n)$ can be superimposed by background noise $b(n)$

$$x(n) = s(n) + \underbrace{b(n) \cdot \sigma^{-1}}_{\tilde{b}(n)}. \tag{1}$$

Thereby, the signals are assumed to be normalized to the same power, i.e. $\sum_n s^2(n) = \sum_n b^2(n)$. The ratio between speech and noise components in the resulting noisy signal is controlled by σ^2. To obtain realistic test signals, the factor has to be chosen carefully. Typically, the original speech signal is recorded under low-reverberant conditions, e.g., with a closed-talk microphone in an anechoic chamber. In this case the clean speech signal $s(n)$ results from the original speech signal filtered with a measured room impulse response [3].

A common approach to quantify the weighting of speech and background noise is the signal-to-noise power ratio (SNR)

$$\sigma^2_{\mathrm{SNR}} = \frac{\sum_n s^2(n)}{\sum_n \tilde{b}^2(n)}. \tag{2}$$

The definition directly follows from (1). The power of the clean speech signal $s(n)$ and noise signal $\tilde{b}(n)$ are calculated separately, irrespective of their spectral distributions. The subjective perception of human listeners, however, depends significantly on the spectral distributions of speech and noise signals used for signal mixing [1]. We expect that two scenarios perceived as equally weighted may result in different values of σ^2_{SNR} as illustrated in Fig. 1. Hence, the subjective impression is not reasonably predicted by the standard SNR.

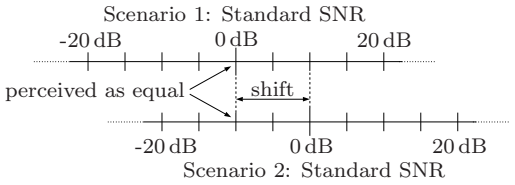

Figure 1: Illustration of the expected mismatch between SNR and human perception. Two scenarios are perceived as equally weighted, however, the SNR scales are shifted.

For this contribution, listening tests were performed to determine the influence of the spectral distribution of noise on the hearing impression of human listeners. In the following, the test setup and the experimental results are described. Based on the experimental results, we investigate the mismatch between SNR and human perception. Subsequently, alternative objective measures and their capability to predict the subjective rating are discussed.

Experimental Setup

In each test of the experiment, two examples of noisy speech signals with different spectral distributions of noise were pairwise presented to the test subjects. The reference example was mixed with a fixed σ^2 whereas the ratio of the other example was adjusted by the subjects. The subjects were instructed to adjust the noise such that the speech similarly sets apart from the noise as much as the reference example does[1]. In Fig. 2 the user interface for performing the experiment is shown.

For the experiment, six different noise signals that cover a wide range of scenarios were chosen: white noise and three bandpass filtered noises exciting different frequency ranges, noise recorded in an automotive environment at a speed of 100 kph exhibiting low frequency components, as well as babble noise with non-stationary components recorded in a cafe. Their spectral distributions are shown in Fig. 3.

In order to achieve an almost phonetically balanced speech signal, the text 'Nordwind und Sonne' was chosen

[1] German instruction: "Stellen Sie [...] das Geräusch des rechten Hörbeispiels so ein, dass sich die Sprache ähnlich von dem Geräusch abhebt wie im Referenzbeispiel."

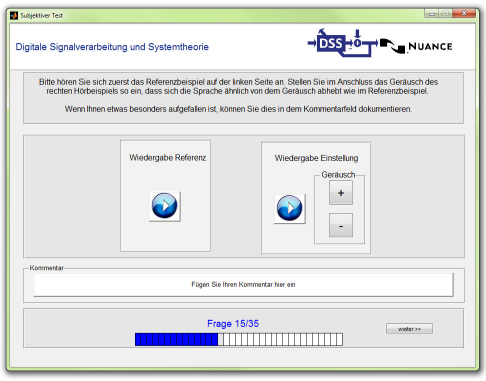

Figure 2: Screenshot of the GUI presented to the test subjects.

[5]. The text was recited by a female and male speaker without developing a Lombard effect. For each test, one of the speech signals was mixed with two different background noises to create the two examples. The spectra for both speech signals are also depicted in Fig. 3.

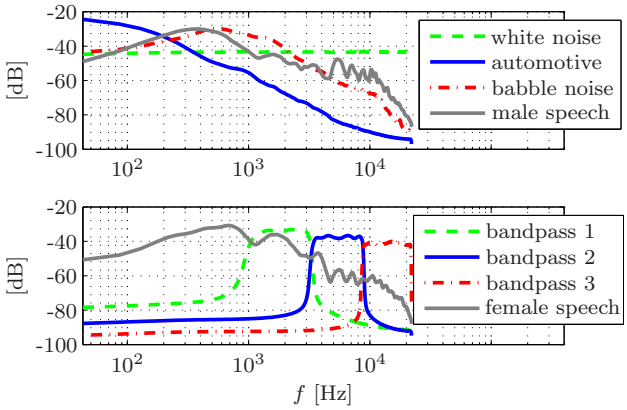

Figure 3: Spectra of the noise and speech signals used in the experiments.

The mixed signals (44.1 kHz sampling rate) were presented to the test subjects in a semi-anechoic chamber. A combination of one loudspeaker and a sub-woofer was employed for correct playback even of the low frequency automotive noise. The sound pressure level of the speech signal was calibrated to 62 dBA in 1 m distance from the loudspeaker according to [6]. A sketch of the hardware setup is shown in Fig. 4.

In total 7 female and 13 male test subjects participated the subjective experiment. All subjects were between 23 and 42 years old and had a command of German on a native speaker level. In order to achieve significant test results, a various number of noise scenario combinations were presented to the subjects, which allows to cross-check the obtained results.

For verification, the noisy signals were recorded with two binaural microphones located close to the test subject as shown in Fig. 4. The audio impression during the experiment can be reproduced based on these signals, however, they are not employed for the evaluations in this contribution.

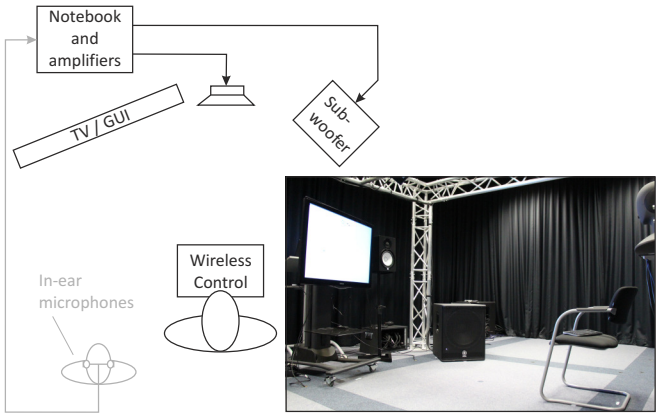

Figure 4: Hardware setup used in the experiment.

Experimental Results

The experimental results are depicted in Fig. 5. For a given ratio σ^2 of the reference, the adjustable ratio σ^2_{subj}, which resulted into a similar subjective hearing impression, was determined. The logarithmic SNR shift $10\log_{10}\left(\sigma^2_{\text{subj}}/\sigma^2\right)$ can be interpreted as the mismatch between the standard SNR and human perception.

When the white noise scenario is compared to the automotive scenario a significant SNR shift can be observed. On average, the SNR for automotive noise was adjusted about 16 dB worse than in the white noise case until both examples where perceived as equally weighted. The small inter-quartile range ensures consistent results over all test subjects.

We obtained similar results when the babble noise scenario was presented as reference. The SNR was perceived similar compared to the white noise and much worse (about 16 dB) compared to the automotive noise. Again, we see a shift of 16 dB between white noise and automotive noise.

As a tendency, the SNRs of the bandpass noises were adjusted lower than the SNR of white noise for equally perceived weightings. The average shift is in the magnitude of 10 dB. However, the results are less consistent as indicated by an increased inter-quartile range. This observation confirms a frequent comment that was expressed by several subjects: the high frequency components of bandpass 3 were perceived as annoying. The latter makes it difficult to rate the different weightings of speech and noise. As no spectral overlap occurs, the speech component sets apart very well, even when the noise level starts to get painful.

To analyze the dependency between the SNR shift and the absolute SNR, the test was repeated for different values of the reference SNR. In Fig. 6, the absolute SNR of white noise is plotted against the absolute SNR for automotive noise which was perceived equal. The result suggests that the shift of 16.4 dB is independent from the absolute SNR for the considered combination.

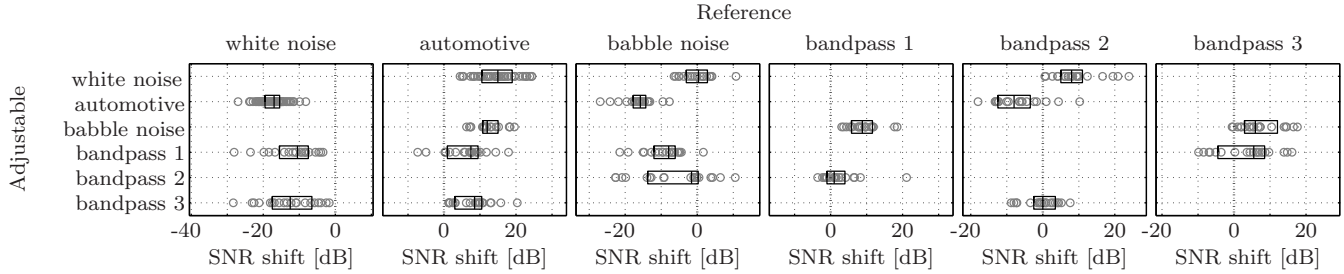

Figure 5: SNR differences between reference noise scenarios (columns) and noise scenarios with adjustable noise power (rows). The noise power was adjusted until the speech similarly sets apart from noise in both scenarios. The box plot of median and the quartiles, as well as the individual results of the test subjects indicated by circles are displayed.

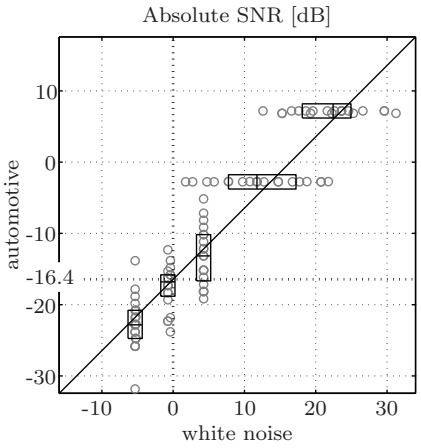

Figure 6: Dependency of absolute SNR for white and automotive noise. The circles indicate SNR combinations that were perceived as equal. The median and quartiles for different reference SNRs are displayed by the boxes. The diagonal line was fitted using the median over all data-points. The result suggests that the SNR shift of 16.4 dB between both scenarios is independent from the absolute SNR in the considered interval.

Objective Measures

Based on our experimental results, we evaluate several dedicated objective measures and their capability to predict the subjective rating. In addition to the standard SNR as defined in Eq. (2), we examine alternative definitions that consider the spectral distribution of speech and noise.

As a first modification, A-weighting is applied to the noise and the speech signal before the ratio $\sigma^2_{\text{A-weighted}}$ is calculated analogous to Eq. (2).

For the ratio $\sigma^2_{\text{modified SNR}}$ based on the modified definition of SNR in [1], the A-weighting is applied only to the speech signal, the noise signal is filtered with the frequency characteristic according to ITU-R 468 instead.

To explicitly address the spectral overlap between speech and noise, we employed a frequency-weighted segmental SNR [2, pp. 509][4]. Based on the estimated power spectral densities of speech $\hat{\Phi}_{SS}(n,\omega)$ and noise $\hat{\Phi}_{BB}(n,\omega)$, a local SNR

$$\text{SNR}_{\text{dB}}(n,\omega) = 10\log_{10}\frac{\hat{\Phi}_{SS}(n,\omega)}{\hat{\Phi}_{BB}(n,\omega)} \tag{3}$$

as well as a weighting function

$$w(n,\omega) = \hat{\Phi}_{SS}(n,\omega)^{0.2} \tag{4}$$

are determined. To estimate the PSDs at frame index n and frequency index ω, we employ a FFT of length $N_{\text{FFT}} = 4096$ at a sampling rate of 44.1 kHz. To reduce the influence of outliers, $\text{SNR}_{\text{dB}}(n,\omega)$ is limited to values between -10 dB and 35 dB. The frequency-weighted segmental SNR then is calculated by

$$10\log_{10}\sigma^2_{\text{fwSNRseg}} = \frac{\sum_{n,\omega} w(n,\omega)\cdot\text{SNR}_{\text{dB}}(n,\omega)}{\sum_{n,\omega} w(n,\omega)}. \tag{5}$$

In contrast to the standard SNR definition, averaging thereby is performed over logarithmic values.

We use another alternative measure that adopts averaging over logarithmic values but avoids calculation of a weighting function. Based on $\hat{\Phi}_{SS}(n,\omega)$ and $\hat{\Phi}_{BB}(n,\omega)$, the average spectra (as depicted in Fig. 3) are estimated and the ratio in dB is averaged over frequency

$$10\log_{10}\sigma^2_{\text{logMean}} = \frac{1}{N_{\text{FFT}}}\sum_{\omega}\underbrace{10\log_{10}\frac{\sum_n\hat{\Phi}_{SS}(n,\omega)}{\sum_n\hat{\Phi}_{BB}(n,\omega)}}_{\text{SNR}_{\text{dB}}(\omega)}. \tag{6}$$

The averaging operation can be interpreted as a geometric mean over frequency which implies that the result may be dominated by small values ≈ 0 ($\hat{=}-\infty$ dB). Therefore, we limit $\text{SNR}_{\text{dB}}(\omega)$ again to values between -10 dB and 35 dB.

By applying the objective measures to the data collected in the subjective experiment, we determine their capability to predict the subjective impression. Analogous to the results for the standard SNR depicted in Fig. 5, we calculate the shift between two scenarios that were perceived as equally weighted. Thereby, we only consider the median over the subjects answers. The results for all objective measures are summarized in Fig. 7.

As already stated, the standard SNR fails to predict the subjective rating especially when comparing white noise with the automotive scenario. The significant shift of about 16 dB suggests to employ a different objective measure.

By using the A-weighted SNR, the shift for automotive noise can be corrected. The A-weighting considers the

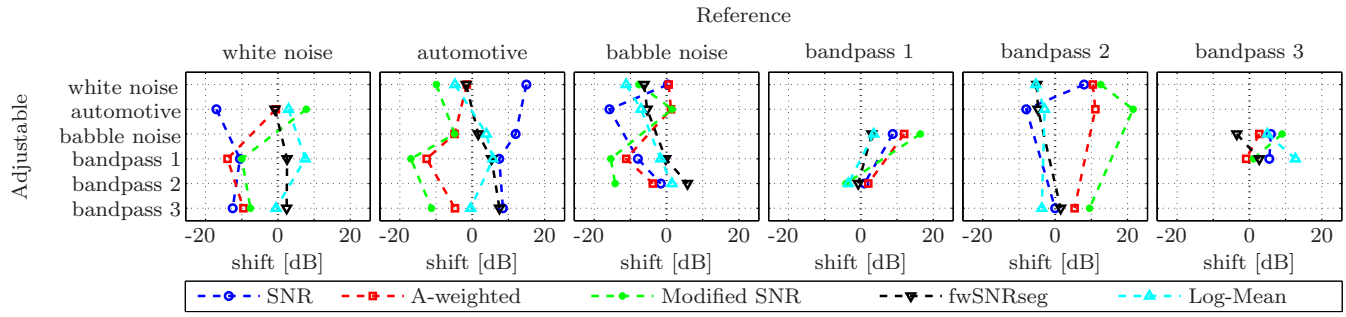

Figure 7: Median shifts calculated with different objective measures. Optimally, an objective measure predicts the subjective impression. In this case, the SNR shift is next to 0 dB for equally perceived weightings.

reduced sensitivity of the human ear for low frequencies which results in a reasonable match of objective and subjective rating. When comparing white, automotive, and babble noise, the shift is almost completely removed. In contrast, between white noise and bandpass noise still a shift of about 10 dB occurs.

In our experiments, no improvements are achieved by employing different weightings for speech and noise.

The measure according to Eq. (5) (fwSNRseg) in our experiment outperforms the other measures. For white and automotive noise, the results are similar to A-weighting but also for the bandpass scenarios a reasonable prediction of the subjective impression is achieved.

The last measure according to Eq. (6), based on averaging over logarithmic values, generates slightly worse results compared to fwSNRseg. Nevertheless, because of its simplicity it could be preferred.

The averaged results for selected groups of scenarios are summarized in Table 1. For this, the mean over the absolute values of the median in Fig. 7 was calculated. The good prediction of the A-weighted SNR for the first three scenarios, as well as the improved results of fwSNRseg are noticeable.

Table 1: Averaged absolute SNR shift for different objective measures

Measure	Mean absolute shift [dB]		
	white, auto, babble	**bandpass 1,2,3**	**all**
SNR (2)	12.0	2.2	8.1
A-weighted	**1.7**	2.8	6.1
Modified SNR [1]	6.3	5.0	10.1
fwSNRseg (5)	3.1	**1.6**	**3.4**
Log-mean (6)	6.0	6.6	4.6

Conclusions

For this contribution, we investigated the effect of the spectral distribution of noise on the perception of human listeners. Specifically, we determined the SNR shift between two scenarios where the weighting between speech and noise was perceived as being equal. The subjective results of our experiments was then compared to the outcome of different objective measures. As expected, the standard SNR fails to predict the subjective impression, especially for low-frequency automotive noise. This effect can be avoided by applying A-weighting to speech and noise signals before calculating the SNR. Other measures that average the local SNR in dB reasonably predict the human perception.

References

[1] K. Linhard, H. Schnepp *Modified SNR for Evaluation of Speech Quality*, ITG-Fachtagung Sprachkommunikation, Aachen, 2008.

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, CRC Press, 2013.

[3] H. G. Hirsch, H. Finster *The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems*, European Conference on Speech Communication and Technology, Lisboa, Portugal, 2005.

[4] J.M. Tribolet, P. Noll, B. McDermott, R.E. Crochiere *A study of complexity and quality of speech waveform coders*, ICASSP, Tulsa, Oklahoma, USA, 1978

[5] Handbook of the International Phonetic Association: *A Guide to the Use of the International Phonetic Alphabet*, International P. Association, C.A.I. Corporate, June 1999

[6] ANSI S3.5, 1997 *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute; New York: Reaffirmed, 2007.