OBJECTIVE PERCEPTUAL QUALITY MEASURES FOR THE EVALUATION OF NOISE REDUCTION SCHEMES

Thomas Rohdenburg, Volker Hohmann and Birger Kollmeier

Thomas.Rohdenburg@uni-oldenburg.de
University of Oldenburg, Medical Physics Group, D-26111 Oldenburg, Germany

ABSTRACT

In this study, different objective quality measures for the performance prediction of noise reduction schemes are compared to subjective data from psychoacoustic listening tests. It is shown that the considered perceptual measures are appropriate to define a quality test-bench which can be used for the development and optimization of noise reduction schemes.

1. INTRODUCTION

Objective quality assessment for noise reduction schemes is needed to reduce time-consuming and cost-intensive subjective listening tests. Still, for most noise reduction schemes proposed in the literature, technical measures that incorporate only little psychoacoustical knowledge have been used to evaluate the performance and to find efficient parameter settings for speech enhancement. However, there have been attempts to define standard quality evaluation measures that are better suited for this task by Hansen and Pellom in [1]. These and other psychoacoustical measures [2] have been applied and compared with subjective data in studies by Marzinzik [3] for several monaural and binaural speech enhancement algorithms.
In this study, standard noise reduction schemes are evaluated and optimized by means of perceptual quality measures and several technical measures in order to increase knowledge on quality assessment of algorithms based on recent objective measures. The algorithms considered here belong to the class of short-term spectral attenuation (STSA) algorithms that try to reconstruct the desired signal's envelope in subbands by means of a time-variant filter in the frequency domain.
The algorithms contain parameters that are affecting the amount of noise reduction. However, maximizing the attenuation of noise with these parameters in general leads to a distortion of the desired signal which can only be tolerated to a certain amount. Technical measures like, e.g., the signal-to-noise ratio enhancement (SNRE) do not incorporate the distortion of the desired signal. Linear quality measures like, e.g., the correlation between estimated and true desired signal are also inappropriate as they do not provide information whether the distortion is audible or even disturbing. Better suited for quality assessment are perceptual quality measures such as PSM from PEMO-Q [4] and PESQ [5] which have been approved for signal distortion in audio and speech codec evaluation.
In subjective quality assessment tests of noise reduction schemes subjects often have difficulties in rating the overall quality which seems to be a trade-off between the amount of background noise removal and speech distortion. Another point is that background noise can even become more annoying if processed by a suppression algorithm, e.g., by introducing musical tones or amplitude fluctuations.
The same difficulty exists for predicting the overall quality with objective measures. While it should be feasible to quantify the amount of noise reduction or to measure speech distortion separately, the prediction of the overall quality seems to be more complex. The task is to find an objective perceptual measure that has a high correlation with the results from subjective ratings. This measure or combination of various measures can then be used as a test-bench for evaluation and parameter optimization in noise reduction schemes.

2. ALGORITHMS

STSA algorithms according to Ephraim and Malah's weighting rules [6] where employed as single channel state-of-the-art algorithms. These algorithms are characterized by a strong reduction of noise while introducing only little of the well known *musical tones* or *musical noise* that result from subtracting an average noise spectrum from a non-stationary frame-based spectral estimate. A detailed description of the involved filter parameters can be found in Cappé [7]. The most important parameters are two signal-to-noise ratio (SNR) estimates: An instantaneously estimated (a posteriori) SNR and an a priori SNR estimate that is calculated by a recursive smoothing of preceding a posteriori values. The considered algorithms need a reliable noise power estimation. Here, the minimum statistics method (MinStat) by Martin [8] and a Voice Activity Detection (VAD) algorithm by Marzinzik [3] are used.

3. SIGNALS

The speech signals used here were taken from the Oldenburg Logatome Speech Corpus (OLLO) [9] and consisted of six sentences spoken by german male and female speakers. The noise signals were speech-shaped noise, cafeteria noise, icra7 noise (speech like modulated noise) and white gaussian noise. All signals had an approximate duration of 20 seconds and a sampling rate of 16 kHz. In the simulation system the signals were mixed at a SNR of 0 dB and 5 dB. The calculation of the time-variant filter was made on this mixture while the filtering process was also done on the separate speech and noise signals for subsequent quality assessment and the calculation of the SNRE and other quality measures.

4. OBJECTIVE MEASURES

The quality prediction method PEMO-Q is based on a psychoacoustically validated quantitative model of the "effective" peripheral auditory processing by Dau et al. [10].
The perceptual similarity measure (PSM), obtained from PEMO-Q, is a correlation measure between two so called *internal representations* of acoustic stimuli, i.e., the output of the modeled peripheral auditory system. PSM serves to predict the perceived similarity between two given signals, generally a reference signal and a test signal whose quality is to be measured.
The aim of the noise reduction scheme can be defined as to achieve a higher perceived similarity between the processed signal and the desired signal than between the unprocessed and the desired signal. These two perceived similarities are estimated by PSM. The difference between these two PSM measures, referred to as ΔPSM, serves to measure the performance of a noise reduction algorithm. Positive ΔPSM values predict a higher quality of the processed signal compared to the unprocessed signal, whereas negative values indicate a signal degradation.
The PSM measure was also varied with an option called Beerends-Berger-assimilation. A discussion on this assimilation step can be found in [4]. The measure with this option switched off is referred to as PSM_b.
As another psychoacoustical measure, the ITU standard measure PESQ is used to evaluate the quality of the processed data. PESQ stands for "Perceptual Evaluation of Speech Quality" and is an enhanced perceptual quality measurement for voice quality in telecommunications [5]. The difference between the unprocessed signal's quality and the processed signal's quality prediction by PESQ is referred to as ΔPESQ.
For both psychoacoustical measures, using clean speech as a reference might not be appropriate to predict the degradation of the speech signal introduced by the algorithms. Therefore it could be helpful to blind out the influence of the noise reduction on the predicted quality. As one possible solution, we suggest to use a mixture of the signal plus noise, at a SNR that corresponds to that of the processed signal, as the reference signal. In the following the measures which use a noisy reference signal are referred to as SNR_PSM and SNR_PESQ.
Besides the above perceptual measures also more "technically based" quality measures where incorporated. These were SNRE, coherence, a critical bandwidth weighted SNRE (freq. wt. SNRE), and the quality evaluation measures defined by Hansen and Pellom [1]: segmental SNR, Log-Area Ratio (LAR), Log-Likelihood Ratio (LLR), Itakura-Saito Distance (ISD) (the last three measures are based on a LPC-Model).

5. EXPERIMENTS

5.1. Signal processing and objective quality assessment

The recursive smoothing parameter τ for the a priori SNR-estimate in Ephraim and Malah's algorithms has great influence on the noise reduction strength. In order to find an optimal setting and to cover a broad range of qualities for a subsequent correlation analysis of objective and subjective measures, τ was varied in the range from $0 - 800$ms. All signals were processed with the noise estimators Minstat and VAD, respectively. For each setting, the above mentioned quality measures were calculated for a number of speech signals mixed with different types of noise (see Section 3). For subjective listening tests a subset of 7 time-constants per noise type, algorithm and input-SNR was chosen.

5.2. Subjective listening tests and quality assessment

The subjective listening tests were done according to the ITU-T Recommendation P.835 [11] which describes a methodology for evaluating the subjective quality of speech in noise and is particularly appropriate for the evaluation of noise suppression algorithms. The methodology uses separate absolute categorial rating scales (ACR) to independently estimate the subjective quality of the speech signal alone, the background noise alone and the overall quality. 16 normal hearing subjects were tested. The whole test consisted of 8 sessions with 15 trials each and took approximately 1 hour. One trial was composed of three sub-samples. Each sub-sample consisted of two sentences, male and female talkers, of 3.25 seconds duration each. In the first sub-sample the subjects were instructed to attend only to the background noise and rate it on a five category scale from "1 - sehr störend" (very disturbing) to "5 - gerade wahrnehmbar" (just noticeable). In the second sub-sample subjects were instructed to attend only to the speech signal and rate it on a scale from "1 - sehr stark verzerrt" (very much distorted) to "5 - unverzerrt" (not distorted). In the third sub-sample subjects were

instructed to listen to the speech + background and rate it on a five category overall quality scale from ”1 - schlecht” (bad) to ”5 - ausgezeichnet” (excellent). The ratings were done with an ACR - software using sliders that allowed a sub-categorial rating in 0.1 steps.

6. RESULTS

Fig. 2 shows the subjective data in the left panels for both noise estimators and two of the four noise types. Initially, it can be stated that all subjective tests - independent of noise type, input-SNR or noise estimators - show consistent behavior in the way that the perceived speech-signal qualities (red curves) decrease and the amount of perceived noise reduction (green curves) increase monotonically by increasing the smoothing constant τ. As stated before, the overall quality ratings (black curves) seem to be a trade-off between both rating tasks. Obviously the subjects prefer in virtually all cases a certain amount of smoothing. Another point is that the two noise estimators, MinStat and VAD, show different performance for fluctuating noise, e.g., speech-modulated icra7-noise, but similar behavior for stationary noise while the mean opinion score (MOS) for the overall quality is almost the same. The subjects reported that - especially in cases of fluctuating noise - they were uncertain what to prefer - reducing noise and accepting signal-distortion or the opposite. This may be the reason why there was no preference for one of the noise estimators observable although the outputs were very different.

To find out which objective measure describes the respective quality rating tasks best, the correlation between different objective measures and the subjective data were evaluated (see Tab. 1). The highest correlations of all objective measures are indicated with bold black numbers, the highest negative correlations are printed in red. The first four columns show the correlation for each noise type separately. The last column contains the overall correlation for all signal types, algorithms, and input-SNR’s. Rows 1-7 show more technically measures, i.e. these measures are not based on a complex psychoacoustic model. Rows 8-12 contain the perceptual measures and their relative enhancement representations (ΔPSM, ΔPESQ) all with clean speech reference. The last rows contain the perceptual measures but with an output-SNR-aligned noisy reference signal, indicated by the prefix ”SNR_”.

As for the background noise rating, the highest correlations are gained by the SNRE. This means that SNRE is a good measure to rate the amount of noise reduction by an algorithm, independent of the speech signal quality. Also, high correlation values are gained by the ΔPSM measure if different types of background noise are considered separately. The correlation for ΔPSM with the subjective data can be seen in Fig. 1. The functional relationship between subjective and objective measures varies across different types of background noise (color-coded), hence the overall correlation is less. As a consequence the objective measure should incorporate some noise dependent scaling to better model the subjective data.

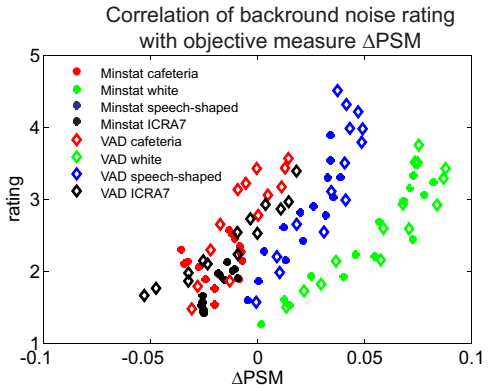

Figure 1: Noise dependent correlation between objective and subjective data

In terms of speech-signal rating most of the correlations are negative. The strongest anticorrelated measure is the frequency weighted SNRE, which may result from the fact that noise reduction and speech distortion are competing processes in the considered algorithms. The strongest correlations show the perceptual measures with the noisy reference, especially SNR_PESQ, as expected.

The best correlation in terms of overall quality rating show the perceptual measures with clean speech reference, especially PESQ and PSM_b.

The right panels in Fig. 2 show the prediction of the subjective data on the left panels by the objective measures that had the highest correlations for each rating task, i.e., SNRE for the prediction of perceived noise reduction, SNR_PESQ for the speech-signal degradation and PSM_b for the prediction of the overall quality. The curves have been linearly fitted to match the scaling of the MOS.

7. SUMMARY AND OUTLOOK

The results show that objective measures are able to predict subjective ratings in noise reduction schemes. In terms of noise reduction alone the SNRE measure is appropriate, but for objective assessment of perceived speech signal distortion or overall quality, perceptual measures such as PESQ and PSM (PEMO-Q) are better suited. Whereas PESQ was optimized for speech quality, PSM is a global audio quality measure that is also applicable to, e.g., processed music and transients. A selection of the best objective measures for each rating task together with representative noise types will be used as a test-bench for the development of novel noise reduction schemes.

Correlation with background noise rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	**0.93**	0.91	0.88	**0.90**	**0.75**
Coherence	0.50	0.67	0.58	0.68	0.53
seg. SNRE	0.71	0.62	0.63	0.84	0.54
freq. wt. SNRE	0.70	0.79	0.54	0.66	0.49
mean LAR	0.33	-0.89	0.18	0.35	-0.06
mean LLR	-0.08	-0.73	0.06	0.05	-0.14
mean ISD	0.55	-0.51	0.54	0.67	0.20
PSM	0.57	0.89	0.84	0.70	0.69
PSM_b	0.52	0.66	0.70	0.69	0.60
PESQ	0.37	0.66	0.64	0.62	0.63
ΔPSM	0.76	**0.92**	**0.89**	0.83	0.62
ΔPESQ	0.42	0.81	0.64	0.84	0.56
SNR_PSM	-0.56	-0.49	-0.39	-0.58	-0.28
SNR_PESQ	-0.60	-0.81	-0.58	-0.53	-0.41

Correlation with speech signal rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	-0.67	-0.77	-0.94	-0.87	-0.67
Coherence	0.27	0.02	-0.21	-0.04	-0.05
seg. SNRE	-0.06	0.09	-0.32	-0.46	-0.17
freq. wt. SNRE	-0.90	-0.89	-0.79	-0.93	-0.70
mean LAR	-0.88	0.33	-0.46	-0.67	-0.06
mean LLR	-0.66	0.01	-0.41	-0.72	-0.22
mean ISD	-0.79	-0.13	-0.63	-0.89	-0.62
PSM	0.22	-0.31	-0.58	-0.07	-0.15
PSM_b	0.25	0.07	-0.38	-0.06	-0.02
PESQ	0.41	0.06	-0.33	0.05	-0.01
ΔPSM	-0.05	-0.75	-0.90	-0.49	-0.39
ΔPESQ	0.34	-0.27	-0.73	-0.52	-0.23
SNR_PSM	0.84	0.76	0.67	**0.87**	0.61
SNR_PESQ	**0.87**	**0.92**	**0.86**	**0.87**	**0.74**

Correlation with overall quality rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	0.35	0.66	0.41	0.29	0.35
Coherence	0.83	0.88	0.93	0.93	0.65
seg. SNRE	0.74	0.89	0.89	0.71	0.53
freq. wt. SNRE	-0.17	0.40	-0.05	-0.11	0.00
mean LAR	-0.46	-0.90	-0.45	-0.28	-0.07
mean LLR	-0.75	-0.94	-0.61	-0.68	-0.43
mean ISD	-0.24	-0.79	-0.04	-0.16	-0.34
PSM	0.82	**0.92**	0.93	0.87	0.70
PSM_b	0.85	0.91	**0.94**	0.93	0.76
PESQ	0.85	**0.92**	**0.94**	**0.94**	**0.81**
ΔPSM	0.71	0.69	0.58	0.54	0.39
ΔPESQ	**0.86**	**0.92**	0.48	0.65	0.47
SNR_PSM	0.12	-0.04	0.09	0.01	0.04
SNR_PESQ	0.09	-0.36	0.05	0.07	0.00

Table 1: Correlation between objective and subjective measures for the three rating tasks and the types of background noises.

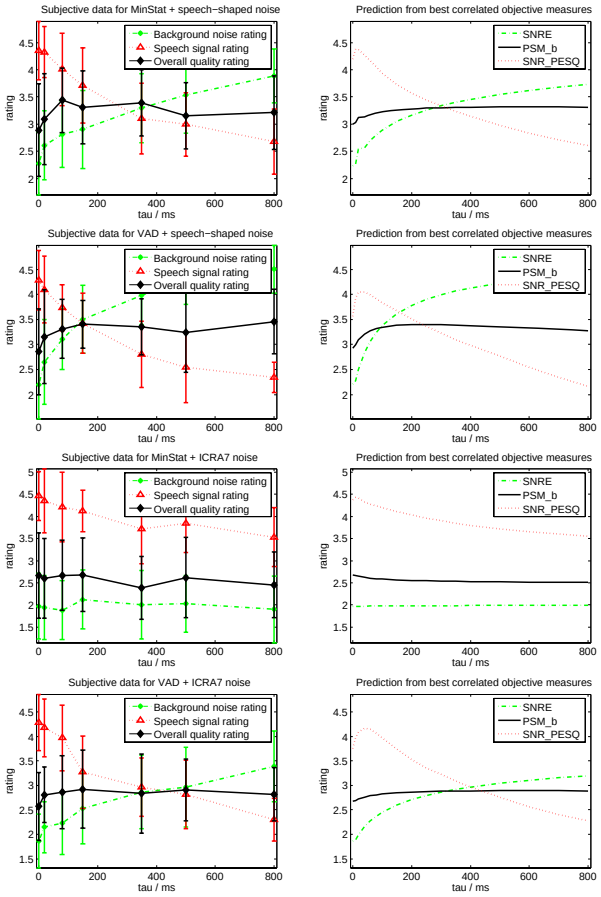

Figure 2: Subjective (left panel) and objective (right panel) data for different noise types and algorithms

7.1. Acknowledgements

We like to thank OPTICOM GmbH for kindly providing us the perceptual speech quality measurement PESQ for scientific purposes and Dr. Rainer Huber from HörTech gGmbH for his intense support and collaboration in this study.

8. REFERENCES

[1] J.H.L Hansen and Bryan Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, Dec. 1998. 1, 4

[2] M. Hansen and B. Kollmeier, "Objective modelling of speech quality with a psychoacoustically validated auditory model," *Journal of the Audio Engeneering Society (JAES)*, vol. 48, pp. 395–409, 2000. 1

[3] M. Marzinzik, *Noise Reduction Schemes for Digital Hearing Aids and their Use for Hearing Impaired*, Ph.D. thesis, University of Oldenburg, Nov. 2000. 1, 2

[4] R. Huber, *Objective assessment of audio quality using an auditory processing model*, Ph.D. thesis, University of Oldenburg, 2003, "http://docserver.bis.uni-oldenburg.de/publikationen/dissertation/2004/hubobj03/hubobj03.html". 1, 4

[5] ITU-T, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Series P: Telephone Transmission Quality Recommendation P.862, International Telecommunications Union, Feb 2001. 1, 4

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error short–time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984. 2

[7] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994. 2

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001. 2

[9] T. Wesker, B. Meyer, K. Wagener, B. Kollmeier, and A. Mertins, "Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines," in *Interspeech*, 2005. 3

[10] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system i," *Journal of the Acoustical Society of America (JASA)*, vol. 99, no. 6, 1996. 4

[11] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Series P: Telephone Transmission Quality Recommendation P.835, ITU, Nov. 2003. 5.2