OPTIMAL SPECTRAL SMOOTHING IN SHORT-TIME SPECTRAL ATTENUATION (STSA) ALGORITHMS: RESULTS OF OBJECTIVE MEASURES AND LISTENING TESTS

Matthias Brandt, Joerg Bitzer

Institute for Hearing Technology and Audiology, University of Applied Sciences Oldenburg
Ofener Str. 16/19, 26121 Oldenburg, Germany
phone: + (49) 441 7708 3732, fax: + (49) 441 7708 3777, email: brandt@fh-oow.de
web: www.hoertechnik-audiologie.de

ABSTRACT

In this paper, we investigate different types of spectral smoothing of the transfer function of single-channel noise reduction algorithms in terms of achieved audio quality. In order to determine the audio quality extensive listening tests have been conducted. Furthermore, we computed several existing objective quality measures based on technical measures or psychoacoustics. We examine whether the different forms of spectral smoothing of the weighting rule of the noise reduction algorithm are represented by the objective measures. We show that most of the known measures are insensitive to changes in the short-time spectra that are subtle in a technical way, but immediately noticeable by human listeners. The results of the listening test also indicate the optimal smoothing method for speech and audio enhancement.

1. INTRODUCTION

The design of pleasant-sounding single-channel algorithms for suppressing unwanted noise in noisy audio signals is, and always has been, a demanding task. On the one hand this is due to the challenge posed by minimizing distortions of the desired signal, and on the other hand it is hard to avoid unwanted side effects – so called artefacts –, such as unnatural sounding residual noise. Most well-sounding solutions are based on Short-Time Spectral Attenuation (STSA) [12]. A very common combination is the noise reduction rule by Ephraim and Malah [4] with noise estimation techniques based on minima tracking [14, 3]. However, the final audio signal obtained by standard algorithms can be improved by smoothing the time-varying Transfer Function (TF). This smoothing is necessary to avoid unwanted modulation of the residual noise, which is known as musical tones for poorly adjusted algorithms, but even for well-tuned algorithms some remaining artefacts are usually audible. Often recursive smoothing of adjacent blocks is used to overcome fluctuation in time. The influence of this smoothing parameter on the resulting audio quality has been examined by Rohdenburg in terms of subjective listening tests and objective measures [16].

A second solution is the smoothing of the TF in the frequency direction (see section 2). Several approaches are known, e.g. constant bandwidth or constant-Q averaging [6].

This paper concentrates on showing the effectivity of different types of spectral smoothing by evaluating the output signal quality in (subjective) listening tests and by examining several objective quality measures. In the following section the different methods for spectral smoothing will be explained briefly. In sections 3 and 4 the subjective methodology for the listening test and for the objective measures are introduced. Finally, we discuss the results and some conclusions are drawn.

2. SPECTRAL SMOOTHING

Spectral smoothing in general is the process of reducing the variance of neighboring values of the spectra of signals or the transfer functions of systems. It may be realized in a variety of ways, e.g. by computing a running average over spectral values or based on the cepstrum representation of a signal [2]. Another way is based on (linear predictive) models describing the signals [11]. In this article, we concentrate on evaluating spectral smoothing that is based on computing a running average over the frequency values. This process can be carried out directly in the frequency domain by convolving the spectrum with an appropriate (normalized) window function (see figure 1): A number of spectral values is weighted with a window function and summed up to yield one spectral value of the smoothed spectrum. The bandwidth of the smoothing window may be fixed or can be varying over frequency [6]. A common method to define frequency-dependent bandwidths is to use *fractional-octave* bandwidths.

2.1 Constant bandwidth smoothing

Keeping the number of spectral values that are incorporated into the averaging process constant over the whole frequency range corresponds to smoothing with a constant bandwidth. In this case, the latter is usually specified in $[B|_{\mathrm{Hz}}] = \mathrm{Hz}$. This method can be implemented very efficiently by multiplication with a window function in the time-domain.

2.2 Fractional-octave smoothing

From a psychoacoustic point of view it makes sense to specify the bandwidth as a *ratio* of two frequencies. In this context, the unit of one octave is commonly used, defining a doubling of frequency. The edge frequencies of a $B|_{\mathrm{oct}} = 1/x$-octave interval with center frequency f are given by [6]

$$\begin{aligned} f_{\mathrm{l}}(f) &= 0.5^{\left(\frac{1}{2}B|_{\mathrm{oct}}\right)} \cdot f \\ f_{\mathrm{u}}(f) &= 2^{\left(\frac{1}{2}B|_{\mathrm{oct}}\right)} \cdot f. \end{aligned}$$

The bandwidth in Hz is frequency-dependent in this case and results in

$$B(f)|_{\mathrm{Hz}} = f_{\mathrm{u}}(f) - f_{\mathrm{l}}(f).$$

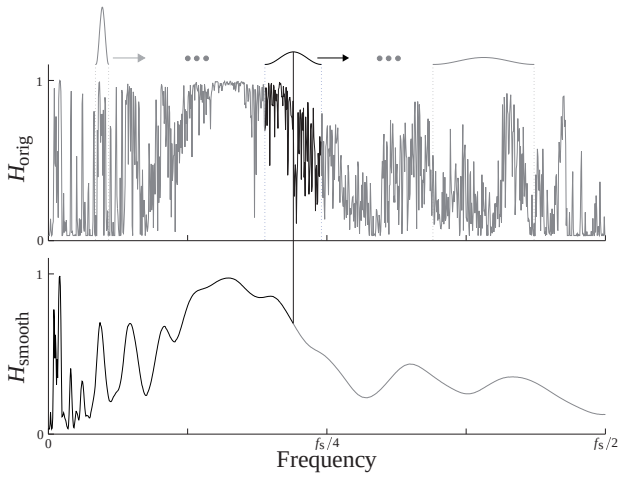

Figure 1: *Spectral smoothing by convolution in the frequency domain.* A certain, optionally frequency-dependent, number of samples of the original spectrum H_{orig} is weighted with a window function and then summed up to yield one sample of the smoothed spectrum H_{smooth}. In this example, the smoothing window gets broader for higher frequencies. The first half of the spectrum is shown.

3. TEST SIGNALS AND SUBJECTIVE QUALITY EVALUATION

To conduct the listening tests, signals from the NOIZEUS database [7] have been used, containing short sentences in English, spoken by female and male speakers. After resampling with 16kHz and adding white noise to obtain an overall SNR of 10dB, the noisy signals have been fed through a denoising algorithm which is based on short-time spectral attenuation (STSA) by Wiener filtering. The noise-floor was estimated using the minimum statistics method [14], and to reduce the musical noise effect, the decision-directed approach [17] was used and, additionally, the maximum spectral attenuation was limited to 15dB. The algorithm benefits from spectrally smoothing the transfer function of the Wiener filter to reduce the fluctuation of the residual noise which is usually perceived as very annoying. To study the effects of spectral smoothing, output signals of a denoising algorithm have been generated with different types of spectral smoothing employed: constant bandwidth (in Hz) and frequency-dependent bandwidth (in octaves).

For the subsequent listening tests the pairwise-comparison method was chosen: The ten probands were presented pairs of two randomly selected signals that had been processed with different spectral smoothing bandwidths. After listening to both signals, they were asked to choose the one containing the "most naturally sounding speech recording". Four signals for each smoothing bandwidth had to be rated this way. The ranking order of the probands' ratings was determined afterwards by applying the Bradley-Terry-Luce (BTL) model [1, 13].

4. OBJECTIVE QUALITY MEASURES

Objective measures aim at predicting the audio quality perceived by a human being. The different algorithms are categorized into so-called *intrusive* and *non-intrusive* methods. Intrusive techniques are solely able to predict the relative quality by computing some kind of distance measure to a reference signal. Non-intrusive algorithms in contrast try to predict the (absolute) quality of an audio signal without any further information.

The measures investigated in this article and their respective abbreviations are: the overall SNR (SNR), the segmental SNR (SNRseg), the log-likelihood ratio (LLR), the log-area ratio (LAR), the Itakura-Saito distance (IS), the cepstral distance (CEP), the weighted spectral slope measure (WSS) [12, 5], the ITU-T's PESQ method [10, 15] (PESQ), two composite measures presented in [8] (MARS_{sig}, MARS_{ovl}), and two measures provided by the PEMO-Q algorithm [9] (PSM, PSMt).

4.1 Description of the Measures

While the SNR and SNRseg measures directly incorporate the time domain signals, the others rely on transformations of the signal. LLR, LAR and CEP are distance measures based on the difference of the coefficients of autoregressive (AR) models of the input signals. The IS tries to predict the perceived difference of two spectra, and the WSS mainly expresses the difference in spectral peak locations [12].

Considering the definition of the overall SNR – incorporating the whole signal at once –, a small correlation to the perceived quality is to be expected as human beings continuously observe the audio signal to make their decisions concerning quality. The segmental SNR takes this fact into account by averaging the SNRs of *short blocks* of audio. However, the spectral distribution of the energy is disregarded in both cases.

AR model based measures are capable of effectively indicating differences of speech spectra. These models reproduce the spectral shaping of the vocal tract. Depending on the model order, the spectral properties are captured rather roughly which makes those LPC based measures insensitive to minor changes in the signals.

The PESQ, PSM and PSMt measures aim at simulating the processing performed inside the human auditory system. These methods use some kind of auditory transform of both reference signal and signal under test. The measure of quality – or better: similarity – is then computed in the "auditory domain". The PESQ measure has been developed to assess the quality of speech transmission systems. (Although an extension to the original ITU-T Recommendation describes the application of the PESQ method for wideband audio signals, in this article the basic implementation, assuming low-bandwidth speech signals, is used.) Coarsly, the PESQ algorithm consists of

1. filtering both reference and test signal with a telephone handset filter
2. piecewise time alignment and equalisation
3. auditory transform
4. extraction of distortion parameters between the transforms of both signals
5. mapping to a prediction of a mean opinion score (MOS) rating

By incorporating Bark spectra, sone-loudness mapping and (simultaneous) masking effects, a subset of the mechanisms in the human auditory system is effectively reproduced.

The PEMO-Q algorithm consists of

1. time delay and level matching
2. shortening silent intervals to 200ms

3. auditory transform (employing basilar-membrane filtering, envelope extraction, adaptation and filtering by a modulation filterbank)

The auditory transform of the PEMO-Q method is able to simulate effects of the absolute hearing threshold, temporal masking and adaptation.

The composite measures presented in [8], MARS_{sig} and MARS_{ovl} combine the IS and PESQ measure to attain a high correlation to subjective ratings, concerning quality of the desired signal and overall signal, respectively.

4.2 Range of Values

To gain a grasp of which values are possible for the different objective measures, we mixed speech signals from the NOIZEUS database with white Gaussian noise to obtain different SNRs. Afterwards, all objective measures have been computed using noise-free signals for reference. The results are depicted as a reference for the final results in figure 2. Some objective measures show a limit for low SNR values which is caused by the underlying models that are unsuitable for highly noisy signals.

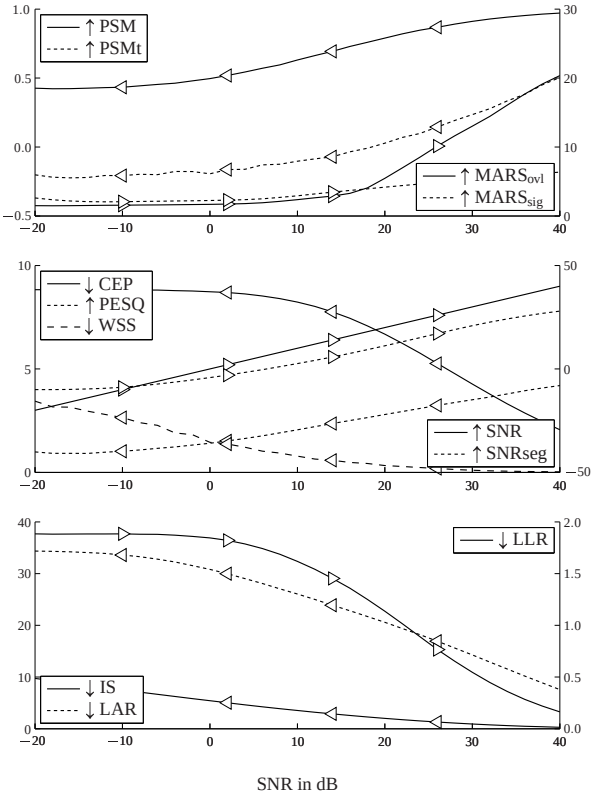

Figure 2: *Objective measures for varying input SNR.* This plot indicates how the objective measures react to different input SNRs. All diagrams in this article that contain two y-axes use triangles to assign curves. Triangles pointing to the left (◁) correspond to the left axis and legend, triangles pointing to the right (▷) correspond to the right axis and legend. Generally, up or down pointing arrows (↑, ↓) inside legend boxes indicate the direction of smaller distance to the reference signal – which is equivalent to higher audio quality of the wanted signal in the context of this article.

5. RESULTS

5.1 Subjective Ratings

In a first run, two listening tests were carried out to determine the preferred bandwidth of constant-bandwidth spectral smoothing and fractional-octave smoothing independently. The results of these listening tests are given in table 1. Additionally, the BTL model distances have been plotted versus the relative frequencies of the probands' ratings in figure 3a) to visualize the correlation between both values: A fair correlation can be observed, justifying the use of the BTL model – however, because the latter disregards test results with low consistency values, the highest relative frequency does not necessarily lead to the best ranking (see $1/3$oct vs. $1/6$oct).

For constant-bandwidth smoothing, the preferred bandwidth is $B|_{\text{Hz}} = 200\text{Hz}$, for fractional-octave smoothing, the preferred bandwidth is $B|_{\text{oct}} = 1/6\,\text{oct}$, closely followed by $B|_{\text{oct}} = 1/3\,\text{oct}$ with a very small BTL model distance, which means both methods are rated more or less the same. The consistency is not as high as we wished, showing that the test persons were not able to judge the different methods without contradiction in their rating. The level of significance for the accordance test is 99%, indicating a high agreement between the different test persons.

The results indicate that spectral smoothing with medium bandwidths has a positive influence on the perceived quality. Furthermore, too much smoothing will jeopardize the quality. This indicates that the very broad filters introduce some unwanted artefacts to the desired signal and we believe that the broader smoothing introduces unnatural sound when the filter opens, similar to breathing or sibilance sounds at higher frequencies which causes the poor rating.

Ranking	Fixed Bandwidth	Frequency-Dependent Bandwidth
1	200Hz (0.00)	$1/6\,\text{oct}$ (0.00)
2	500Hz (0.24)	$1/3\,\text{oct}$ (0.03)
3	100Hz (0.35)	$1\,\text{oct}$ (0.78)
4	50Hz (0.91)	$1/12\,\text{oct}$ (1.14)
5	2000Hz (0.96)	$1/24\,\text{oct}$ (2.08)
Consistency $\varnothing$	0.72	0.62
Level of significance	0.99	0.99

Table 1: *Results of listening tests to determine the preferred spectral smoothing bandwidths.* The listeners' task was to choose the signal with the higher naturalness of the speech sound. The number of participants was ten.

In a subsequent listening test, the first two preferred bandwidths of each type of smoothing had to be rated by the listeners to determine the overall preferred type of spectral smoothing. The results are shown in table 2 and figure 3b). The overall preferred type of spectral smoothing is fractional-octave smoothing with $B|_{\text{oct}} = 1/3\,\text{oct}$. The consistency is even lower compared to the preceding tests, which is reasonable since the test signals were much more similar concerning sound quality. The low consistency shows that for a broad range of the smoothing parameter the perceived quality is close. However, it can be seen that appropriate smoothing is a necessary step for high sound quality by the distance to the hardly-smoothed and heavy-smoothed results.

5.2 Objective Measures

The curves of all objective measures in dependence on the smoothing bandwidth are presented in figure 4 for constant-

Ranking	Bandwidth
1	$1/3$ oct (0.00)
2	500Hz (1.03)
3	$1/6$ oct (1.20)
4	200Hz (1.22)
Consistency $\varnothing$	0.60
Level of significance	0.99

Table 2: *Results of listening tests to determine the overall preferred spectral smoothing bandwidth.* The listeners' task was to choose the signal with the higher naturalness of the speech sound. The number of participants was ten.

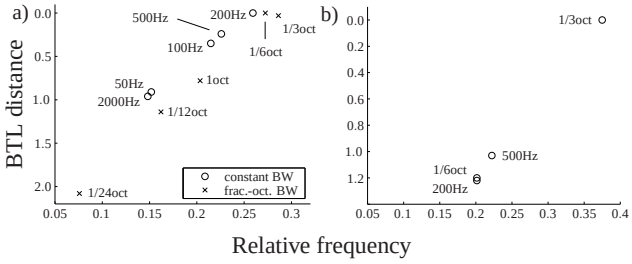

Figure 3: *Computed BTL model distances plotted versus the relative frequencies of the probands' ratings.* The results of the first two runs to determine the preferred constant bandwidth and fractional-octave bandwidth smoothing individually are shown in diagram a), the results of the second run to determine the overall preferred bandwidth are shown in diagram b).

bandwidth smoothing and figure 5 for fractional-octave smoothing. The values of the objective measures are the average results for 30 test signals (15 male, 15 female speakers) per smoothing bandwidth.

Most of the measures indicate a gain in quality of the denoised signal compared to the unprocessed one. For example the LAR is on average at a value of eight for all linear smoothing methods. This low value corresponds to an unprocessed quality at an SNR of 40dB (see figure 2), which means a quality gain equivalent of 30 dB SNR enhancement was achieved (input SNR was 10dB). The values for other measures (e.g. PESQ) are much smaller but mostly above a corresponding SNR of 20dB, which still means an enhancement of 10dB compared to the unprocessed signal.

The range of all computed measures is very small (note the different scaling of the y-axes) compared to the overall range given in figure 2. However, the differences between smoothing with narrow and broad bandwidth and the corresponding perceived signal quality rated by human subjects is much higher.

If we compare the objective results with the results of the listening tests it can be seen that none of the objective measures has a clear maximum like the results in the listening tests. Most show a monotonic relationship between quality measure and bandwidth of the smoothing. Measures with slight maxima like IS (figure 5) and LLR predicted the worst quality at regions were the subjective tests indicates best quality. For the psychoacoustically motivated measures (PESQ and PSM) the results are not very encouraging. They indicate that more smoothing is better.

The only measure that follows the subjective listening results in an overall trend is MARS_{sig}. This measure predicted the best quality at $1/3$ octave, which is the best result given in the listening test.

For comparison purposes, the bandwidths corresponding to the best sound quality are summarized in table 3: results of the listening tests are juxtaposed with the bandwidths that the objective measures indicate to be optimum.

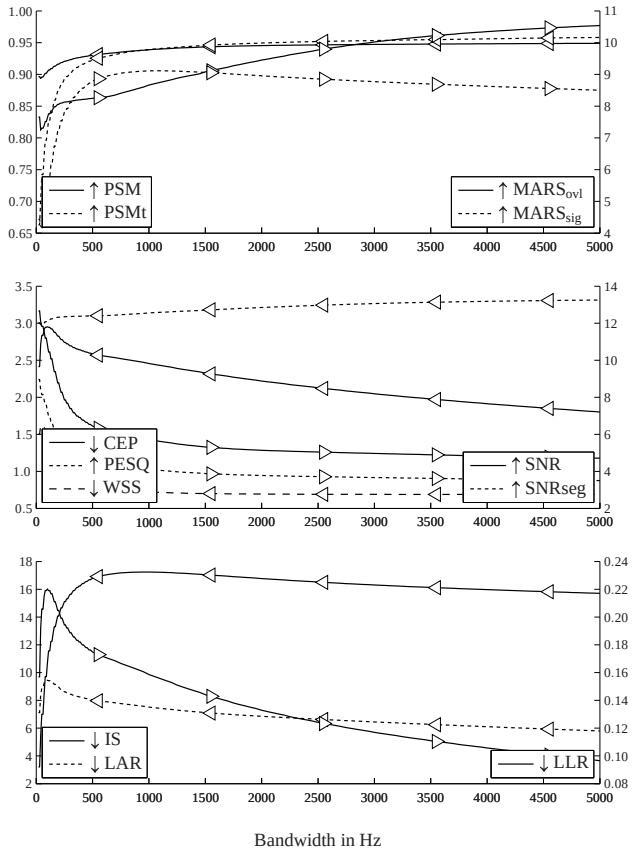

Figure 4: *Objective measures for constant-bandwidth smoothing.*

6. CONCLUSIONS

In this paper we have shown by analyzing listening test results that spectral smoothing is a necessary step for high-quality noise reduction systems. The results clearly indicate that the smoothing should not be too broad because of unwanted side effects like "noise breathing" and not too narrow since the desired reduction of noise modulation is not successful in this case. The choice of the optimal solution is not that obvious, it seems as if it is a matter of taste and sound material. However, smoothing is a vital component for noise reduction. Furthermore, the results of the objective measures show a relatively small dependency on the employed smoothing method, even though the subjective impact of smoothing is large for the noise and the desired signal quality.

7. ACKNOWLEDGEMENT

This research was (partly) funded by grant 17N3008 of the German Federal Ministry of Education and Research (BMBF). The views and conclusions contained in this document, however, are those of the authors.

	subjective	SNR	SNRseg	LLR	LAR	CEP	IS	WSS	PESQ	PSM	PSMt	$MARS_{sig}$	$MARS_{ovl}$
constant	200Hz	20Hz	20Hz	5kHz	5kHz	5kHz	20Hz	4.5kHz	5kHz	5kHz	5kHz	1.1kHz	5kHz
fract.-oct.	1/6 oct	1/36oct	1/36oct	4oct	4oct	4oct	1/36oct	4oct	4oct	4oct	4oct	1oct	4oct

Table 3: *Values resulting in best sound quality for constant-bandwidth smoothing and fractional-octave smoothing.* The results of the listening tests and the bandwidths indicated by the objective measures are listed.

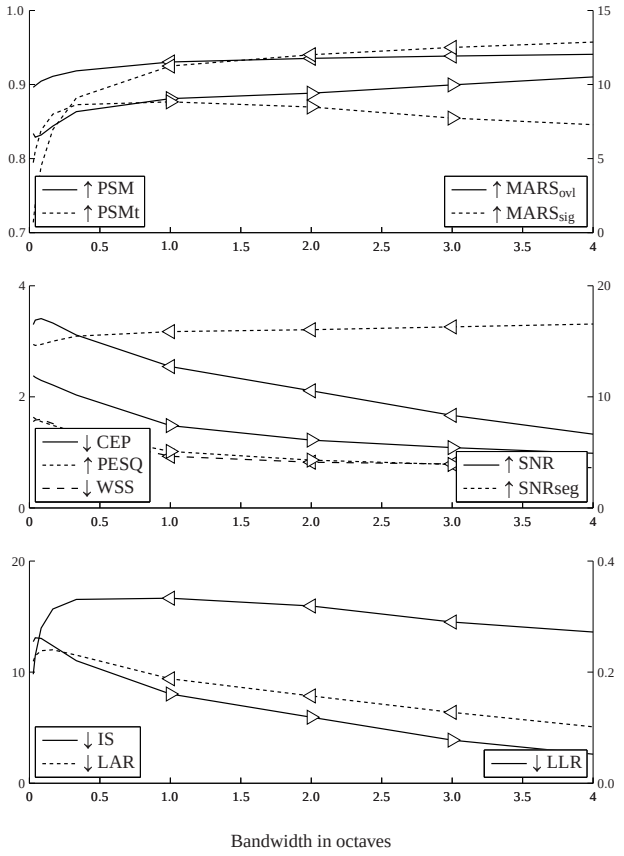

Figure 5: *Objective measures for fractional-octave smoothing.*

We would like to thank the reviewers for their helpful comments, not all of which could be considered, unfortunately, due to the limitation of space.

REFERENCES

[1] R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs – I. The Method of Paired Comparisons. *Biometrika*, 39(3–4):324–345, 1952.

[2] D. C. Childers, D. P. Skinner, and R. C. Kemerait. The Cepstrum: A Guide to Processing. *Proceedings of the IEEE*, 65(10), 1977.

[3] I. Cohen. Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475, 2003.

[4] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(2):443–445, 1985.

[5] J. H. L. Hansen and B. L. Pellom. An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms. In *Proceedings of the International Conference on Speech and Language Processing*, pages 2819–2822, 1998.

[6] P. D. Hatziantoniou and J. N. Mourjopoulos. Generalized Fractional-Octave Smoothing of Audio and Acoustic Responses. *The Journal of the Acoustical Society of America*, 48(4):259–280, 2000.

[7] Y. Hu and P. C. Loizou. Subjective Comparison and Evaluation of Speech Enhancement Algorithms. *Speech Communication*, 49:588–601, 2006.

[8] Y. Hu and P. C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:229–238, 2008.

[9] R. Huber and B. Kollmeier. PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1902–1911, 2006.

[10] ITU-T. *Recommendation P.862 – Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[11] J. Makhoul. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, 63(4), 1975.

[12] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press LLC, 1st edition, June 2007.

[13] R. D. Luce. *Individual Choice Behaviour: A Theoretical Analysis*. Wiley, 1959.

[14] R. Martin. Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, 2001.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

[16] T. Rohdenburg. *Development and Objective Perceptual Quality Assessment of Monaural and Binaural Noise Reduction Schemes for Hearing Aids*. PhD thesis, University of Oldenburg, Oldenburg, Germany, 2008.

[17] P. Scalart, J. V. Filho, and J. G. Chiquito. On Speech Enhancement Algorithms Based on MMSE Estimation. *8th European Signal Processing Conference*, 1996.