SUBBAND-BASED PARAMETER OPTIMIZATION IN NOISE REDUCTION SCHEMES BY MEANS OF OBJECTIVE PERCEPTUAL QUALITY MEASURES

Thomas Rohdenburg, Volker Hohmann, Birger Kollmeier

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

ABSTRACT

In general, noise reduction schemes for application in hearing-aids or car environments have parameters that are determined by technical distance measures or heuristically based on informal listening by the algorithm developers. In [1] we have shown that quality measures based on psychoacoustic models are better suited to optimize single parameters in terms of the best subjective overall quality than pure technical measures like, e.g., the signal-to-noise ratio. In other words, a test-bench based on objective quality measures and several typical noise types can support the search for the best-sounding noise reduction algorithms and their internal parameter settings. However, if the algorithms become more complex, e.g., because of frequency-dependent parameters, a single broadband measure might not be feasible to assess optimal settings because of the high dimensionality of the parameter space. In this case a subband-based perceptual quality measure might be feasible. In this study, we exemplarily apply subband-based quality prediction to parameter optimization in a noise reduction algorithm based on auditory filters.

1. INTRODUCTION

The aim of this study is to improve the applicability of perceptual objective measures to the systematic optimization of noise reduction algorithms. In particular, perceptual measures calculated in subbands are used to optimize a multidimensional parameter set band-wise. The technique is exemplarily applied to a monaural state-of-the-art noise reduction scheme, which was adopted to work with gammatone auditory filterbank signals instead of short-time fourier transformed (STFT-) signals. The parameterized noise reduction algorithm described in section 2 is then optimized with the perceptual subband measure which is defined in section 3. To assess the effects of noise reduction on the so called internal representations we take a look at processed speech signals mixed with stationary speech-shaped noise in section 5. The results are summarized in section 6.

2. ALGORITHM

The proposed noise reduction scheme (see Fig. 1) is based on the idea of Ephraim and Malah's MMSE[1] log-STSA[2] [2] algorithm. Instead of the short-time fourier transform (STFT) we use a complex-valued gammatone filterbank which is supposed to have a frequency resolution similar to that of the auditory system. The gammatone filters [3] are widely used in computational auditory models for modeling the peripheral filtering in the cochlea. [4] proposes an efficient complex-valued implementation with signal resynthesis, which is used here.
Let $s(t)$ and $n(t)$ denote the speech and the noise signals, respectively. The observed signal $x(t)$ is given by

$$x(t) \quad = \quad s(t) + n(t). \qquad (1)$$

If the noisy time-signal $x(t)$ is filtered by a 4th-order linear gammatone filterbank we get a complex time-frequency dependent signal similar to a STFT-processed signal

$$X(t,f) \quad = \quad S(t,f) + N(t,f), \qquad (2)$$

with t denoting the time-index and f the center-frequency-index of the discrete signals. We believe that $X(t,f)$ can be processed similar to an STFT-signal by multiplication with a time-varying gain $G(t,f)$ with the aim to reconstruct the desired signal's envelope. The desired signal is estimated by

$$\hat{S}(t,f) \quad = \quad G(t,f) \cdot X(t,f). \qquad (3)$$

$\hat{S}(t,f)$ can be resynthesized into a time signal $\hat{s}(t)$ with low delay by using the synthesis algorithm in [4]. $G(t,f)$ is calculated due to [2, 5] based on two SNR estimates:

$$G(t,f) \quad = \quad f\{\mathrm{SNR}_{\mathrm{post}}(t,f), \mathrm{SNR}_{\mathrm{prio}}(t,f)\} \qquad (4)$$

with

$$\mathrm{SNR}_{\mathrm{post}}(t,f) \quad = \quad P\left[\frac{\hat{\Phi}_{XX}(t,f)}{\hat{\Phi}_{NN}(t,f)} - 1\right] \qquad (5)$$

$$\text{with} \quad P[x] = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \qquad (6)$$

[1] MMSE: minimum mean squared error
[2] STSA: short-time spectral attenuation

$$\mathrm{SNR}_{\mathrm{prio}}(t,f) = \alpha\frac{\hat{\Phi}_{SS}(t,f)}{\hat{\Phi}_{NN}(t,f)} + (1-\alpha)\mathrm{SNR}_{\mathrm{post}}(t,f) \quad (7)$$

In this equations $\hat{\Phi}_{NN}, \hat{\Phi}_{XX}$ and $\hat{\Phi}_{SS}$ denote power estimates of the signals N, X and $\hat{S}$, respectively. In practice, Eq. (4) is precalculated and stored in a two-dimensional gain table spanned by the two SNR estimates. Eq. (7) is known as the "decision directed approach". The a priori SNR, $\mathrm{SNR}_{\mathrm{prio}}$, is a weighted sum of the previously estimated SNR and the instantaneous a posteriori SNR. The weighting factor α has the character of a smoothing constant with the equivalent low-pass time-constant $\tau(f) = \frac{-T_a}{\ln(\alpha(f))}$, T_a : sampling period (block period). $\hat{\Phi}_{NN}(t,f)$ is estimated using a modified version of the minimum statistics method by Martin [6]. $\hat{\Phi}_{XX}$ and $\hat{\Phi}_{SS}$ are calculated as follows:

$$\hat{\Phi}_{SS}(t,f) = \alpha_s(f)\hat{\Phi}_{SS}(t-1,f) + (1-\alpha_s(f))|\hat{S}(t-1,f)|^2 \quad (8)$$

$$\hat{\Phi}_{XX}(t,f) = \alpha_x(f)\hat{\Phi}_{XX}(t-1,f) + (1-\alpha_x(f))|X(t,f)|^2 \quad (9)$$

It has been found experimentally that frequency dependent smoothing of the power estimates for X and $\hat{S}$ with the smoothing parameters α_x, α_s (lowpass time constants τ_x, τ_s) is useful when processing gammatone filterbank signals. In STFT-based algorithms these smoothing parameters are 0, accordingly

$$\hat{\Phi}_{XX}(t,f) = |X(t,f)|^2 \quad (10)$$

$$\hat{\Phi}_{SS}(t,f) = |\hat{S}(t-1,f)|^2. \quad (11)$$

The amount of smoothing reduces amplitude modulations and has to be selected carefully to not destroy important speech information. On the other hand, the choice of the time constants has an influence on distortions in the filtered output signal. Therefore constants will be evaluated experimentally with a perceptual quality measure that is discussed in the following section.

3. PERCEPTUAL QUALITY MEASURES

The perceptual similarity measure (PSM) obtained from PEMO-Q [7] is a broadband measure which is suitable for measuring the overall quality of an audio signal. PEMO-Q is based on a quantitative model of the "effective" auditory signal processing by Dau et al. [8]. The audio signal is transformed in several stages into an internal representation corresponding to physiological findings about the human hearing system. The first stage is a linear 4th order gammatone filterbank [3] accounting for the basilar membrane's bandpass characteristic. The following stages are

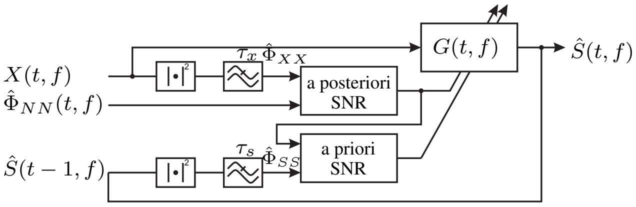


Figure 1: *Noise reduction scheme based on gammatone auditory filterbank*

- a halfwave rectification and lowpass-filtering which are roughly simulating the transformation of mechanical oscillations into neural firing rates,
- an absolute threshold accounting for the hearing threshold,
- five cascaded feedback loops that model temporal masking and adaptation effects of the hearing system and
- a modulation lowpass filter with subsequent resampling to 100 Hz sampling rate.

In [7] the last lowpass filter is replaced by a modulation filterbank which is better suited for detecting small signal degradations introduced by, e.g., audio codecs, for which these quality measures were originally designed. Here, a lowpass filter appears to be sufficient as in case of noise reduction the signal degradation produced by the lossy system (i.e. additive noise + noise reduction) is much higher than that of audio codecs.

The broadband quality measure PSM is the mean of the subband correlations between the internal representations of the processed test signal and a reference. Generally, the desired signal for the analyzed algorithm is taken as the reference for the objective quality measure. In terms of noise reduction schemes, the desired signal can be clean speech or a noisy signal which has a higher SNR than the input signal.

For frequency dependent quality measurement we use the subband correlations and the averaging across all frequency bands is omitted. Let I_{tf} denote the time-frequency dependent internal representation of the estimated speech signal $\hat{s}(t)$. D_{tf} is the internal representation of the desired signal, the reference. μ_I, μ_D denote the temporal mean of the internal representations I_{tf} and D_{tf}. The subband similarity measure is then given by

$$\mathrm{PSM}(f) = \frac{\sum\limits_t (I_{tf} - \mu_I(f))(D_{tf} - \mu_D(f))}{\sqrt{\sum\limits_t (I_{tf} - \mu_I)^2 (D_{tf} - \mu_D(f))^2}} \quad (12)$$

4. PARAMETER OPTIMIZATION

In [1] we showed that the perceptual quality measure PSM from PEMO-Q has a high correlation with the subjective ratings of the overall quality. By varying the smoothing parameter of the STFT-based Ephraim-Malah algorithm (according to α in eq. 7) we could predict the optimal smoothing in terms of subjective overall quality. As a consequence, the parameter τ could be optimized by maximizing PSM.

In the case of the gammatone-filterbank based algorithm we have multiple parameters that are frequency dependent because of variable filter bandwidths and time resolution. The filterbank in the noise reduction system is similar to that used in PEMO-Q. This allows us to see the effects of frequency dependent algorithm parameters on each subband of the internal representation. If we combine the smoothing parameters $\tau_X(f)$ and $\tau_S(f)$ (eqns. 8,9) to a parameter vector

$$\mathrm{params}(f) \quad = \quad \{\tau_X(f), \tau_S(f), \ldots\} \qquad (13)$$

then

$$\mathrm{params}^{\mathrm{opt}}(f) \quad = \quad \arg \max_{\mathrm{params}(f)} \mathrm{PSM}(f) \qquad (14)$$

describes the optimal frequency dependent parameter vector. This can be used for automatic subband quality optimization of the proposed algorithm. However, unconstrained independent subband optimization can lead to a large variation of the optimal parameters across frequency bands. This happens if the variance of subband PSM-values for different settings is small, then, slight numerical changes of the input signal can cause a great change of ”optimal” values. To overcome this problem we suggest to add the constraint that only small parameter changes between adjacent frequency bands are allowed and that the parameter values change monotonically with frequency.

The following section discusses the effects of the noise reduction system on the internal representation.

5. EFFECTS OF THE NOISE REDUCTION ON THE INTERNAL REPRESENTATIONS

Figure 2 (a) shows a temporal section of the power envelope and (b) the related internal representation (IR) in subband 10 (569 Hz) for the clean speech signal (red) and the noisy (speech-shaped noise, 5dB SNR) input signal (black). It can be seen in (b) that the most striking differences between clean speech and noisy signal are the stronger overshoots and undershoots in the clean speech signal IR whereas the behavior of the subband power envelope (a) is different: Here, the additive noise only influences the envelope in speech pauses and does not raise the peaks significantly. For stationary inputs, the (adaptation) feedback loops of the auditory model have a compressive effect (see [7]). Therefore, the noisy signal IR has less peaks than the clean speech signal IR, assuming that the noise is stationary compared to the speech signal. The task of the noise reduction scheme can be interpreted as reconstruction of the peaks of the speech signal IR. One drawback of spectral subtraction based noise reduction schemes is the occurrence of musical tones that can be identified in the IR as erroneous peaks (see Fig. 2 (c)). Here, the parameters of the noise reduction algorithm have been optimized to generate a processed signal IR (green) that has the highest possible correlation for the given parameter space. The correlation between the reference IR (red) and the processed signal IR (green) is only slightly higher than the correlations of the IRs in (b), while the difference between the related audio signals is clearly audible. This shows that the results in single channel noise reduction systems are always a suboptimal tradeoff between noise reduction and speech distortion. Even if the SNR is enhanced, the perceptual quality, predicted by the measure PSM, can hardly be improved. This implies that with the given parameter space of the algorithm it is impossible to get closer to the desired clean speech reference. Note that a perfect match (correlation = 1) between the IRs would predict that the processed signal is indiscriminable from the clean speech.

In summary, the clean speech signal IR cannot be reconstructed by single channel noise reduction schemes. This

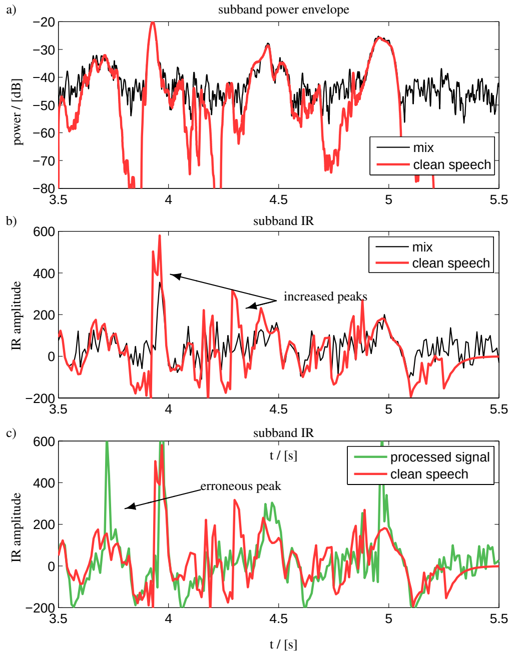
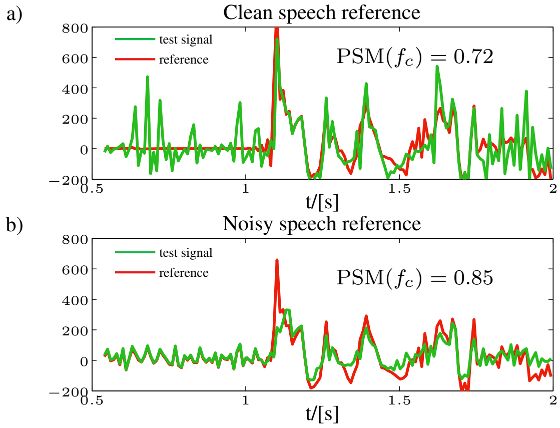


Figure 2: *Subband power envelope (a) and internal representations (b,c) for subband 10 (center frequency* $f_c =$ 569 *Hz)*

leads us to the assumption that a noisy signal at a higher SNR is better suited than a clean speech reference for perceptual optimization. The effect of the optimization with different reference signals on the IR is depicted in Fig. 3. It shows the IRs of the test signals (green) and the references (red) at higher frequency bands ($f_c = 2119$ Hz). The noise reduction seems to be better suited for high frequency bands and therefore also the correlation between test signal and reference IR is higher compared to low frequency bands. We found that the optimization with the noisy reference signal sometimes leads to less artifacts (erroneous peaks in the test signal, see Fig. 3 (a) 1.6 sec.), because the noisy reference allows for residual noise after noise reduction (0.5-1 sec.). Using a noisy signal with an SNR of 25 dB (20 dB above the input signal) as a reference for the quality measure and incorporating the optimization constraints mentioned above, the perceptually optimal time constants of the proposed noise reduction algorithm were:


a) Clean speech reference

b) Noisy speech reference

Figure 3: Parameter optimization using a clean speech reference (a) and a noisy (25 dB SNR) speech reference (b)

Band	1	2	3	4	5	6	…	26	27
τ_S / [ms]	2.0	2.0	1.9	1.9	1.8	1.8	…	1.0	1.0
τ_x / [ms]	100.0	96.5	93.1	89.6	86.2	82.7	…	13.5	10.0

The subjective quality of the output signals was significantly better compared to the direct implementation with time constants $\tau_x, \tau_s = 0$. This approves the assumption that automatic parameter optimization with perceptual quality measures leads to a higher quality of the processed audio signal. However, compared to the STFT-based noise reduction scheme the quality could not be improved, yet. Two reasons can be given for that: First, the bandwidths of the auditory filterbank for low frequencies are smaller than typical FFT-bandwidths which results in stronger envelope fluctuations in these bands. This means that the discrimination between speech and noise based on statistical properties of the envelope is more difficult and leads to more errors. Second, the proposed gain-table by [2] was optimized on the statistical properties of STFT-signals and does not hold for filterbank-signals with variable bandwidths.

6. SUMMARY

In this paper we proposed a new method for subband based quality prediction and parameter optimization which was tested and analyzed on a monaural noise reduction algorithm. The direct conversion of the STFT-based algorithm due to [2] led to strong interferences and artifacts in the audio signal. These artifacts could be reduced by the proposed perceptual quality optimization scheme. With these settings the processed audio signal had a quality which was comparable to STFT-processed signals. However, we could not yet improve the noise reduction scheme by using an auditory filterbank instead of a constant bandwidth STFT-method. Looking at details of the internal representations in subbands, we were able to interpret the principle limitations of the monaural noise reduction scheme.

7. REFERENCES

[1] T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Objective perceptual quality measures for the evaluation of noise reduction schemes,” in *9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 169–172. 1, 3

[2] 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. 1, 4

[3] R. D. Patterson, J. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” pres. at meeting of the IOC Speech Group on Auditory Modelling at RSRE, Dec. 14–15 1987. 1, 2

[4] V. Hohmann, “Frequency analysis and synthesis using a gammatone filterbank,” *Acta acustica / Acustica*, vol. 88, no. 3, pp. 433–442, 2002. 1

[5] 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. 1

[6] R. Martin, “Spectral subtraction based on minimum statistics,” in *Proc. EURASIP European Signal Processing Conference (EUSIPCO)*, Edingburgh, Sept. 1994, pp. 1182–1185. 2

[7] R. Huber and B. Kollmeier, “PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception.,” *IEEE Trans. on Audio, Speech and Language Processing*, Special Issue on Objective Quality Assessment of Speech and Audio, to be published 2, 3

[8] 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. 2