A SINGLE CHANNEL SPEECH ENHANCEMENT TECHNIQUE USING PSYCHOACOUSTIC PRINCIPLES

François Xavier Nsabimana, Vignesh Subbaraman and Udo Zölzer

Department of Signal Processing and Communications
Helmut Schmidt University, Hamburg, Germany
fransa@hsu-hh.de
www.hsu-hh.de/ant

ABSTRACT

This paper presents an algorithm for robust noise reduction technique and speech enhancement based on psychoacoustic principles for low SNR corrupted speech signals. Assuming the clean speech components as masker for the distortions in the enhanced speech, a good quality of the enhanced speech can be obtained, if the distortions in the enhanced speech are kept below the masking curve of the clean speech. But for a single channel noise reduction technique, the real clean speech is not available. The speech components required to compute the desired masking curve are therefore estimated from the corrupted speech using advanced spectral subtraction techniques. The computed masking curve, the estimated corrupting noise and a time-frequency dependent parameter to control the desired predefined residual noise are then used to derive the weighting function. The proposed approach performs well for low SNRs of the corrupted speech compared to other approaches. This is confirmed with the subjective listening and objective grading tests. Some audio demonstrations are available from:

http://www.hsu-hh.de/ant/index_74BUg40IeCJD04Zr.html

***Index Terms*—** Speech Enhancement, Noise Reduction.

1. INTRODUCTION

Spectral enhancement has received a lot of attention for decades, especially for single channel noise reduction techniques because of its simplicity. This attention is also increased mainly due to the necessity of integrating some noise reduction techniques in mobile phones due to the fact that transmitted speech can be corrupted by the environmental noise. In order to enhance such corrupted speech, the corrupting noise needs to be estimated first [1, 2] to finally derive a gain function for the desired noise reduction techniques [3, 4, 5, 6, 7].

In [4, 5, 6], Psychoacoustically Motivated Spectral Weighting Rules (PMSWR), which derive a gain function based on the psychoacoustical properties of the human hearing system, were proposed. Unlike the Log Spectral Amplitude (LSA) and the Optimally Modified Log Spectral Amplitude (OMLSA) estimators [3, 7], the PMSWR approach [5] does not try for a complete noise removal, it preserves instead a predefined constant amount of the original noise throughout the enhanced speech to account for the loss of weak speech components. Therefore the enhanced speech from the PMSWR approach is composed of the speech components and a predefined residual noise. Based on the error minimization of the distortions of speech and noise components compared to the masking curve of the rough clean speech estimate, a gain function was derived [5]. In [4, 6], the computed masking curve is instead used to control the adaption of the oversubtraction factor and the spectral flooring in a generalized spectral subtraction technique.

In this paper an algorithm for the noise reduction technique and speech enhancement similar to the PMSWR approaches is proposed, where the predefined residual noise level is instead controlled by a time-frequency dependent parameter. Unlike the PMSWR approach [5], the level of the residual noise is here varied throughout the enhanced speech based on discrimination between regions with speech presence and speech absence. By controlling the level of the residual noise in the noise only region and keeping it below the masking curve in the speech present region, the unpleasant modulation effect of the residual noise is avoided for very low SNRs. To determine the speech present or speech absent region, a Voice Activity Detector (VAD) as explained in [2] is applied. The clean speech components needed to compute the masking curve are obtained here using the OMLSA approach [7], while the corrupting noise is estimated using [1, 2].

The outline of the paper is as follows. Some preliminary definitions can be found in Section 2. Section 3 presents the proposed enhancement approach. Experimental results and conclusion are presented in section 4 and section 5 respectively.

2. PRELIMINARY DEFINITIONS

Consider the spectrum of a corrupted speech signal to be defined as

$$X(k,m) = S(k,m) + N(k,m), \tag{1}$$

where $S(k,m)$ and $N(k,m)$ are the short-time DFT coefficients at frequency bin k and frame number m for the clean speech and additive noise respectively. $S(k,m)$ and $N(k,m)$ are assumed to be statistically independent and zero mean. The enhancement process is done for the adjacent frames of the corrupted speech $x(n)$ overlapping by 75 % in time domain.

The power level of the clean speech $R_s(k,m)$, of the additive true noise $R_n(k,m)$ and of the corrupted speech $R_x(k,m)$ are obtained by squaring their respective magnitude spectrum. The estimated noise power level $R_{\tilde{n}}(k,m)$ from the corrupted speech power $R_x(k,m)$ can be computed using one of the techniques proposed in [1, 2], depending on the target true mean noise level [1] or due to the rapid adaption time to the noise change [2]. Based on the estimated noise power level $R_{\tilde{n}}(k,m)$, the gain function $G(k,m)$ can be derived as proposed in section 3 or using other desired noise reduction techniques [3, 5, 6, 7]. The spectrum of the enhanced speech is finally obtained by a simple filtering approach given by

$$\tilde{S}(k,m) = G(k,m) \cdot X(k,m). \tag{2}$$

A brief explanation for Eq. (2) is as follows. If $X(k,m)$ represents noise, $G(k,m)$ will then be a low value, which results in attenuating a particular frequency bin. Else if $X(k,m)$ represents speech, $G(k,m)$ will then be close to 1, resulting in speech preservation. To determine whether a particular frequency bin belongs to speech or noise is done with the help of the estimated noise power $R_{\tilde{n}}(k,m)$. $G(k,m)$ is therefore inversely proportional to the estimated noise power level. In the following section an explanation how to derive the expression for $G(k,m)$ in terms of the estimated noise is given.

3. THE PROPOSED APPROACH

In this work, the desired spectrum of the enhanced speech is first defined as

$$\tilde{S}(k,m) = S(k,m) + \zeta(k,m)N(k,m). \tag{3}$$

In the following, a method to control the level of the residual noise $\zeta(k,m)N(k,m)$ in such a way that it remains pleasant in the enhanced speech is proposed. Fig.1 depicts the complete system of the proposed approach. In the analysis stage, the corrupted speech is processed as explained in section 2. The rough clean speech estimate $\hat{S}(k,m)$ needed for the computation of the masking threshold $R_T(k,m)$ is obtained here using the OMLSA approach [7]. The masking curve $R_T(k,m)$ is computed as described in [6, 8, 9, 10] and summarized in [4, 6].

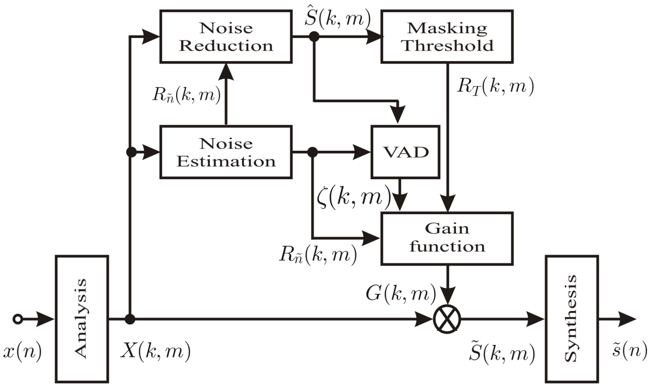

Fig. 1. Improved Psychoacoustically Motivated Spectral Weighting Rule (IPMSWR approach).

From Eq. (2), the Power Spectral Density (PSD) of the enhanced speech is given by

$$R_{\tilde{s}}(k,m) = G^2(k,m)\left[R_s(k,m) + R_n(k,m)\right]. \tag{4}$$

Eq. (4) reveals that the enhanced speech power consists of two components. One is the clean speech and the other one is the residual noise power. For the derivation of $G(k,m)$, the distortion of the speech and residual noise are first considered separately. The development of the psychoacoustical inspired weighting rule relies on the error minimization of the speech power distortion

$$R_{E_s}(k,m) = R_s(k,m)\left[G(k,m) - 1\right]^2 \tag{5}$$

and the residual noise power distortion

$$R_{E_n}(k,m) = R_n(k,m)\left[G(k,m) - \zeta(k,m)\right]^2. \tag{6}$$

Practically a complete masking of both distortions

$$R_E(k,m) = R_{E_s}(k,m) + R_{E_n}(k,m) < R_T(k,m) \tag{7}$$

is not possible. But by masking the residual noise power distortions, the speech power distortions can also be minimized [5]. So equating noise power distortion to masking curve of clean speech, the gain function is obtained from

$$R_T(k,m) = R_n(k,m)\left[G(k,m) - \zeta(k,m)\right]^2, \tag{8}$$

with $\zeta(k,m) \leq G(k,m) \leq 1$. Solving Eq. (8) using the estimated noise power $R_{\tilde{n}}(k,m)$, the spectral weighting rule is then given by

$$G(k,m) \quad = \quad \min(\sqrt{\frac{R_T(k,m)}{R_{\tilde{n}}(k,m)}} + \zeta(\lambda,m), 1), \tag{9}$$

where λ represents therefor a frequency band and

$$\zeta(\lambda,m) \quad = \quad \begin{cases} 10^{-SP/20} & , \quad \text{if } S_r(\lambda,m) > \delta(\lambda) \\ G_{\min} & , \quad \text{otherwise} \end{cases} \tag{10}$$

with

$$S_r(\lambda,m) = \frac{\sum\limits_{\lambda} \left|\hat{S}(\lambda,m)\right|^2}{\sum\limits_{\lambda} R_{\tilde{n}}(\lambda,m)}. \tag{11}$$

Instead of computing ζ in Eq. (9) for each frequency bin k, it is derived in Eq. (10) for three different frequency bands λ (0 - 1kHz, 1 - 3kHz and 3kHz - $f_S/2$) due to the energy distribution of the speech [2]. f_S represents herein the sampling rate. The three frequency bands stand for the simplified octave bands. Alternatively ζ has been also investigated using the energy within one-third octave and critical bands. Therefor the simulation did not provide reliable results especially for low frequency subbands. In Eq. (10), the parameter SP stands for Speech Preservation factor in dB and the threshold $\delta(\lambda)$ is empirically determined.

The advantage of Eq. (11) is that, both the rough enhanced speech and the noise power spectrum are not averaged over the entire frequency spectrum. Otherwise this could mask the high frequency contents since the energy of the low frequency bins is generally high compared to the energy of the high frequency bins. To overcome this, the SNR is found in three different frequency bands separately and each of them is compared to an adaptive threshold $\delta(\lambda)$ [2] to determine whether it is a speech present or speech absent region. This also helps to retain some weak consonants whose energy is concentrated in a very narrow frequency band [11].

The idea in this approach is that the residual noise is not going to be kept constant over frames. It is varied based on whether a frequency region is speech present or speech absent. The control parameter ζ is made equal to $G_{\min} < 0.1$ (-20 dB) in speech absent frequency region, yielding a gain function mostly depending on $R_T(k,m)$ and $R_{\tilde{n}}(k,m)$ (s. Eq. (9)). But in speech present region a small residual noise is allowed. There are two reasons to proceed like that for speech present region. Firstly, the residual noise will be mostly masked by the speech components and remains inaudible. Secondly, if there is a small error in speech estimation, the residual noise will then make up for the loss. For this reason the Noise Reduction factor *NR* from [5] is here replaced by the Speech Preservation factor *SP* in speech present regions. By controlling in such a way the level of the residual noise in the noise only region and keeping it below the masking curve in the speech present region, the unpleasant modulation effect of the residual noise, which is perceived in results from the PMSWR approach at very low SNRs, is avoided. To reduce spectral outliers of the final enhanced speech, a temporal smoothing of the gain function

$$\tilde{G}(k,m) = (1-\beta)\,G(k,m-1) + \beta G(k,m) \qquad (12)$$

is applied in the frequency domain. $\beta < 1$ is herein a desired constant smoothing parameter. To still control the coarseness after Eq. (9) values of $\tilde{G}(k,m) < 10^{-SP/20}$ are replaced by some constant value smaller equal to $G_{\min}$. Recall that $10^{-SP/20}$ must be chosen higher than $G_{\min}$ in Eq. (9).

There is a big advantage for the use of the OMLSA approach to compute $\hat{S}(k,m)$ instead of Spectral Subtraction as proposed in [5], despite the computation time. The OMLSA provides a better estimation of the enhanced speech and suppresses the musical noise effect better than Spectral Subtraction [12]. The OMLSA performs also well in speech absent sections, which is important for very low SNR signals. The masking threshold obtained from the OMLSA is therefore a better rough estimation than the one from Spectral Subtraction in [5].

One disadvantage observed with OMLSA approach is some speech loss in certain frequency bins. But this is controlled here by ζ governed by the Speech Preservation factor *SP* to preserve the speech components in the speech sections. In order to find whether a frequency region is speech present or speech absent, a Voice Activity Detector (VAD) as proposed in [2] and shown in Eq. (10) is applied. The proposed approach avoids musical noise efficiently as the computation of the weighting function implicitly uses the OMLSA estimator which accounts for the musical noise phenomena.

The interpretation for the derivation of $G(k,m)$ in Eq. (9) is pretty straightforward as explained in [5]. Let first consider a relatively strong estimated clean speech. a high value for the masking threshold $R_T(k,m)$ is thus obtained. Therefore $R_T(k,m)/R_{\tilde{n}}(k,m) + \zeta(k,m)$ yields $G(k,m)$ close to 1, resulting in less noise reduction. This is expected because speech will mask the noise and noise reduction is thus not required. If, on the contrary the estimated clean speech is relatively weak, $R_T(k,m)$ will consequently be a low value. The ratio $R_T(k,m)/R_{\tilde{n}}(k,m)$ will be close to zero and $G(k,m) \rightarrow \zeta$, yielding a strong reduction of the corrupting noise.

4. EXPERIMENTAL RESULTS

This section presents the performance evaluation of the proposed enhancement technique in comparison with the PMSWR [5] and the OMLSA [7] approaches. To have a fair comparison, tests were carried out for different noise characteristics (s. Fig. 6). Some parameters for the simulation in the proposed approach are $SP \leq 20$ dB, $G_{\min} \leq 0.1$ ($\leq$ -20 dB), $\delta(\lambda) = 1.3$ and $\beta = 0.95$. A window length of 512 samples with a hopsize of 75 % for analysis and synthesis is applied. The noise power $R_{\tilde{n}}(k,m)$ used in the three approaches is estimated using Optimal Smoothing and Minimum Statistics algorithm [1]. The enhanced speech from all three approaches is subjected to the tests listed below. Fig. 2 depicts the results with the PMSWR and IPMSWR approach for a speech signal corrupted with car noise at 5 dB SNR (s. Fig. 2(b)). Only for the sake of clarity the result with OMLSA is not reported for the following two figures.

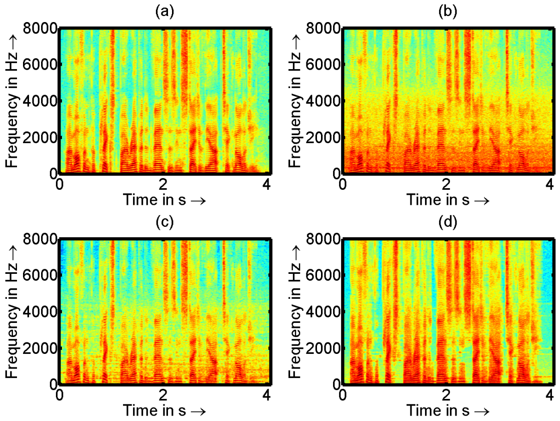

Fig. 2. Results from investigated speech enhancement techniques. Clean speech (a), corrupted speech with care noise (b), PMSWR approach (c) and IPMSWR approach (d).

Fig. 3 reports here the results with the PMSWR and IPMSWR approach for a speech signal corrupted with white noise at 9 dB SNR (s. Fig. 3(b)).

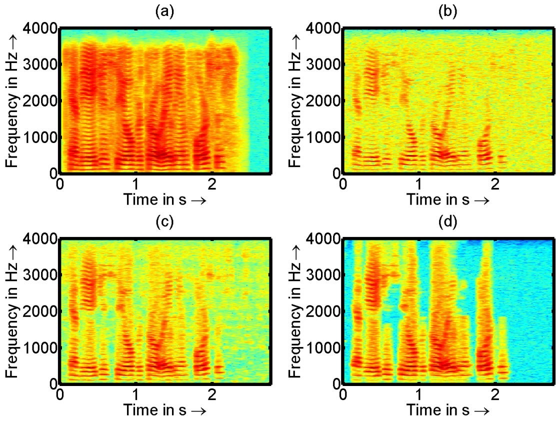

Fig. 3. Results from investigated speech enhancement techniques. Clean speech (a), corrupted speech with white noise (b), PMSWR approach (c) and IPMSWR approach (d).

As the real clean speech is generally not available for a single channel noise reduction technique, Fig. 4 and 5 present the simulation results with the OMLSA, PMSWR and IPMSWR approach for a speech signal corrupted with cockpit noise at 0 dB SNR. For this case, the clean speech was not available. Fig. 4(b) presents result with the OMLSA as implemented in [11]. The result in Fig. 4(b) reveals that the noise only regions are properly attenuated. Although the speech regions are preserved, some speech components are removed at this low SNR signal. This drawback of decision directed approaches [3, 7] at low SNRs has been already observed and investigated in [13]. The result in Fig. 5(b) is obtained applying advanced spectral subtraction techniques (OMLSA) to estimate the rough clean speech $\hat{S}(k,m)$ needed for the computation of the masking threshold $R_T(k,m)$ as stated in [5]. The parameter ζ is set to 0.1 ($NR = 20$ dB) and can be varied.

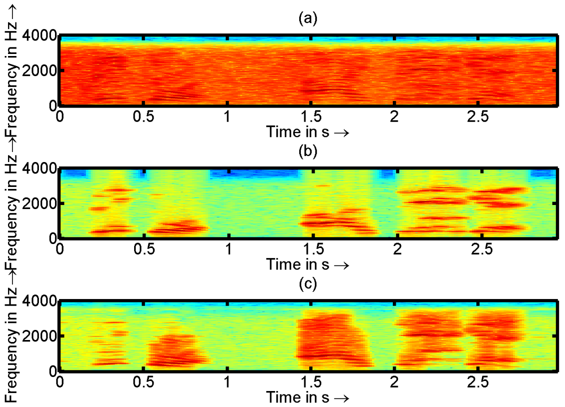

Fig. 4. Results from investigated speech enhancement techniques. Corrupted speech (a), OMSLA approach (b) and IPMSWR approach (c).

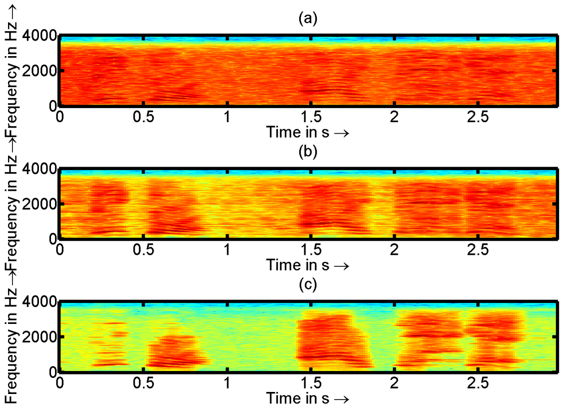

Fig. 5. Results from investigated speech enhancement techniques. Corrupted speech (a), PMSWR approach (b) and IPMSWR approach (c).

Clearly the Psychoacoustically Motivated Spectral Weighting Rules preserve the speech regions and speech components properly compared to the OMLSA from Fig. 4(b). Combining the advantages of the OMLSA approach in speech absent regions (s. Fig. 4(b)) and the advantages of the PMSWR approach for speech preservation (s. Fig. 5(b)), the IPMSWR approach yields good results at low SNRs as shown in Fig. 4(c)-5(c) and confirmed with subjective and objective validations. Obviously the IPMSWR approach enhances better the corrupted speech than the compared approaches (s. Fig. 4 - 6).

4.1 Informal listening tests

In Fig. 6 results obtained during listening test with headphones are presented. The fifteen subjects recruited for this test are all working in our lab. For this test, subjects had first to find the hidden reference signal and assign it 100%. The results from the simulated algorithms are then compared to the reference signal grade. The Mean Opinion Score (MOS) represents the grades of the three enhancement techniques for three different kinds of noise (Fig. 6). Table 1 and Fig. 6 reveal that the IPMSWR approach was graded best.

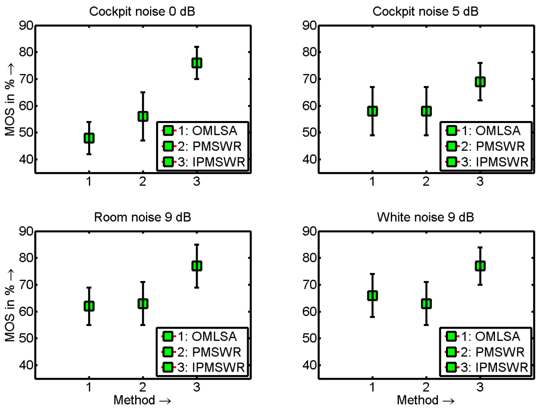

Fig. 6. Results from listening test using headphones. Bars denote 95 % confidence interval.

Table 1. Comparison of the methods using preference in %. Cockpit noise at 0dB (A), cockpit noise at 5dB (B), room noise at 9dB (C) and white noise at 9dB (D).

Method	A	B	C	D
OMLSA	48.43	58.14	62.0	65.9
PMSWR	55.57	58.07	63.1	63.4
IPMSWR	75.64	69.14	77.4	77.4

Table 1 presents the averaged grades of the three compared techniques (OMLSA, PMSWR, IPMSWR) for each of the four simulated signals (A,B,C,D) respectively. The results in Table 1 are derived from the informal listening tests and should therefore reflect the results in Fig. 6.

4.2 Objective quality measures

The objective evaluation is performed as proposed in [14] using the codes available in [11]. The clean speech and the enhanced speech are compared based on Log Likelihood ratio (LLR), Segmental SNR (SSNR), Weighted Spectral Slope (WSS), Perceptual Evaluation of Speech Quality

(P_Q), Speech quality ratio (S_Q) and Overall quality ratio (O_Q) using the composite tool from [11]. The results obtained for a simulation of a clean speech signal corrupted with car noise at 5 dB SNR are shown in Table 2. In the second row of Table 2, the clean speech is compared to itself yielding the upper limits of the reference grades. In the third row results obtained by comparing clean speech and corrupted speech are shown. These scores represent the lower limits of the reference grades. The LLR and WSS scores indicate the speech loss and therefore should be minimum. Whereas the remaining parameters should be maximum. The results in Table 2 show that the IPMSWR approach is graded best for LLR, SSNR and WSS parameters, but remains close to the PMSWR approach for the P_Q and O_Q parameters. Results in Table 2 clearly reveal that the IPMSWR approach combines the advantages of the OMLSA approach in terms of LLR and SSNR improvement and the advantages of the PMSWR approach in terms of WSS, P_Q, S_Q and O_Q improvements.

Table 2. Objective quatily measurement.

	LLR	SSNR	WSS	P_Q	S_Q	O_Q
Reference	0	35	0	4.5	5	5
Corrupted	0.68	3.09	30.66	2.27	3.49	2.86
OMLSA	0.6	7.11	60.18	2.51	3.45	2.89
PMSWR	0.70	6.33	38.4	2.7	3.68	3.2
IPMSWR	0.51	7.44	28.65	2.75	3.96	3.34

The results in Table 2 clearly reveal that P_Q measure does not properly correlate with the distortions introduced by the noise reduction techniques, as it is generally the case with distortions introduced with speech transmitted via communication networks [11]. Although a numerical minor correlation with results in Fig. 6 can be observed, the results in Table 2 are still to handle with care as the composite tool does not yet provide 100 % reliable results.

5. CONCLUSION

A speech enhancement technique based on psychoacoustics principles is proposed here. The key components of this approach are a time-frequency dependent control parameter for the residual noise and a better estimate of the rough clean speech. Since the residual noise is more reduced in the speech absent sections, this algorithm is very well suited for low SNR signals. The estimate of the rough enhanced speech by OMLSA approach also supports this activity as its masking curve is less distorted. Moreover a small correction factor to control the estimation of the enhanced speech is proposed. Subjective and objective quality measurements reveal that the proposed approach performs better than the PMSWR and OMLSA approaches for very low SNR signals. Although the proposed approach currently provides good results, its parameter optimization remains necessary.

Future works will thus concern the increase of the speech intelligibility by appropriately smoothing the gain coefficients and addressing the phase information to account for some speech loss. Moreover, an investigation of the performance of the VAD for a large number of speech segments at various SNRs is required. Finally a big campaign for the comparative study over a very large number of corrupted speech at different SNRs should be conducted.

REFERENCES

[1] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, July 2001.

[2] S. Rangachari, P.C. Loizou, and Yi Hu, “A noise estimation algorithm with rapid adaptation for highly non-stationary environments,” *IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings of ICASSP’04*, vol. 1, pp. I–305–8 vol.1, 17-21 May 2004.

[3] 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*, vol. 33, no. 2, pp. 443–445, April 1985.

[4] N. Virag, “Speech enhancement based on masking properties of the auditory system,” *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95*, vol. 1, pp. 796–799 vol.1, May 1995.

[5] S. Gustafsson, P. Jax, and P. Vary, “A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics,” *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98*, vol. 1, pp. 397–400 vol.1, 12-15 May 1998.

[6] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, March 1999.

[7] I. Cohen, “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *Signal Processing Letters, IEEE*, vol. 9, no. 4, pp. 113–116, Apr 2002.

[8] J.D. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb 1988.

[9] E. Zwicker and H. Fastl., *Psychoacoustics facts and models*, Springer Verlag, Berlin, 1990.

[10] U. Zölzer, *Digitale Audiosignalverabeitung*, B.G. Teubner, 2005.

[11] P. Loizou, *Speech Enhancement Theory and Practice*, Taylor and Francis, 2007.

[12] M. Schwab, H-G. Kim, Wiryadi, and P. Noll, “Robust noise estimation applied to different speech estimators,” in *Asilomar Conference on Signals, Systems, and Computers*, 9-12 November 2003.

[13] J. Jensen J. Erkelens and R. Heusdens, “A general optimization procedure for spectral enhancement methods,” in *Proc. of EUSIPCO’06, Florence, Italy*, 4-8 September 2006.

[14] Yi Hu and P.C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.