Adv. Radio Sci., 8, 95–99, 2010
www.adv-radio-sci.net/8/95/2010/
doi:10.5194/ars-8-95-2010

Advances in Radio Science

A single channel speech enhancement technique exploiting human auditory masking properties

F. X. Nsabimana, V. Subbaraman, and U. Zölzer

Department of Signal Processing and Communications, Helmut Schmidt University, Hamburg, Germany

Abstract. To enhance extreme corrupted speech signals, an Improved Psychoacoustically Motivated Spectral Weighting Rule (IPMSWR) is proposed, that controls the predefined residual noise level by a time-frequency dependent parameter. Unlike conventional Psychoacoustically Motivated Spectral Weighting Rules (PMSWR), the level of the residual noise is here varied throughout the enhanced speech based on the discrimination between the regions with speech presence and speech absence by means of segmental SNR within critical bands. Controlling in such a way the level of the residual noise in the noise only region avoids the unpleasant residual noise perceived at very low SNRs. To derive the gain coefficients, the computation of the masking curve and the estimation of the corrupting noise power are required. Since the clean speech is generally not available for a single channel speech enhancement technique, the rough clean speech components needed to compute the masking curve are here obtained using advanced spectral subtraction techniques. To estimate the corrupting noise, a new technique is employed, that relies on the noise power estimation using rapid adaptation and recursive smoothing principles. The performances of the proposed approach are objectively and subjectively compared to the conventional approaches to highlight the aforementioned improvement.

1 Introduction

The enhancement of speech degraded by environmental or background noise still remains an open topic, although many significant approaches have been presented over years. The enhancement becomes more complicated especially for single channel noise reduction techniques, where no additional information about the corrupting noise and the real clean speech are available. Since the background noise is the factor that degrades the most the quality and intelligibility of the speech, it should therefore be estimated first using adequate techniques such as (Doblinger, 1995; Martin, 2001; Cohen, 2002, 2003; Rangachar, 2004; Stouten, 2006; Nsabimana, 2009). Thereafter, the estimated noise power is used in the derivation of the gain function for a desired noise reduction technique such as (Ephraim, 1985; Gustafsson, 1998; Virag, 1999; Cohen, 2002; Nsabimana, 2009).

In (Tsoukalas, 1993; Gustafsson, 1998; Virag, 1999; Yi, 2004; Hu, 2004; Nsabimana, 2009), 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 techniques like the Log Spectral Amplitude (LSA) and the Optimally Modified Log Spectral Amplitude (OMLSA) (Ephraim, 1985; Cohen, 2002), the Psychoacoustically Motivated Spectral Weighting Rules (Gustafsson, 1998; Nsabimana, 2009) do not try for a complete noise removal, they preserve instead a predefined amount of the original corrupting noise throughout the enhanced speech to account for the loss of weak speech components. Based on the error minimization of the distortions of speech and noise power components compared to the masking curve of the rough clean speech estimate, the gain function is thereafter derived (Gustafsson, 1998; Nsabimana, 2009). While the PMSWR approach (Gustafsson, 1998) generally accounts for a predefined constant amount of the original corrupting noise throughout the enhanced speech, the IPMSWR approach (Nsabimana, 2009) controls instead the level of the residual noise based on the discrimination between the regions with speech presence and speech absence by means of segmental SNR within critical bands.

This paper presents for the noise reduction technique and speech enhancement an algorithm relying on the IPMSWR approach (Nsabimana, 2009), where the estimated corrupting noise power is obtained using rapid adaptation and recursive smoothing principles (RARS) (Nsabimana, 2009) instead of OSMS approach (Martin, 2001; Nsabimana, 2009). The investigations in (Nsabimana, 2009) revealed that the RARS

Correspondence to: F. X. Nsabimana
(nsabfran@yahoo.fr)

Published by Copernicus Publications on behalf of the URSI Landesausschuss in der Bundesrepublik Deutschland e.V.

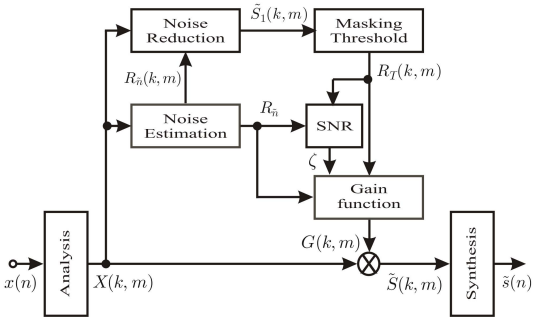

Fig. 1. Improved Psychoacoustically Motivated Spectral Weighting Rule (IPMSWR approach).

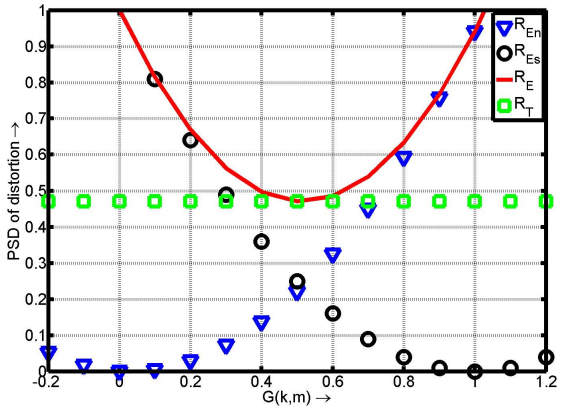

Fig. 2. Error minimization for the derivation of $G(k,m)$. PSD of residual noise distortion R_{E_n}, PSD of speech distortion R_{E_s}, PSD of estimation error R_E and masking threshold R_T.

approach adapts fast and provides accurate mean estimates than OSMS approach (Martin, 2001) especially for very low SNRs. To motivate further steps of improvements in the IPMSWR approach, the importance of the phase is also here emphasized during experimental results.

The outline of the paper is as follows: Sect. 2 presents the proposed technique, while experimental results and conclusion are presented in Sects. 3 and 4 respectively.

2 The proposed approach

Figure 1 depicts the complete system of the proposed approach. In the analysis stage, the corrupted speech is processed frame by frame with an overlapping rate of 75%. The estimated noise power $R_{\tilde{n}}(k,m)$ is computed using RARS approach (Nsabimana, 2009), while the rough clean speech estimate $\tilde{S}_1(k,m)$ needed for the computation of the masking threshold $R_T(k,m)$ is obtained using the OMLSA approach (Cohen, 2002). The masking curve $R_T(k,m)$ is computed as described in (Virag, 1999; Johnston, 1988; Zwicker, 1990; Zölzer, 2005) and summarized in (Virag, 1999). In the following, the derivation of the gain function is detailed.

Let consider the spectrum of a corrupted speech signal $X(k,m)$ 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 also assumed to be statistically independent and zero mean. As a complete noise removal is not intended for psychoacoustically motivated spectral weighting rules, the desired spectrum of the enhanced speech is therefore defined as

$$\hat{S}(k,m) = S(k,m) + \zeta(k,m)N(k,m), \tag{2}$$

where $\zeta(k,m)N(k,m)$ represents the estimated amount of the residual noise. But the estimated magnitude spectrum of the enhanced speech is given by (s. Fig. 1)

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

The difference between Eqs. (2) and (3) yields the estimation error

$$E(k,m) = S(k,m)[G-1] + N(k,m)[G-\zeta], \tag{4}$$

with the PSD of the error expressed as

$$R_E(k,m) = R_s(k,m)[G-1]^2 + R_n(k,m)[G-\zeta]^2, \tag{5}$$

where the indexes k and m are omitted for G and ζ only for the sake of simplicity. $R_s(k,m)$ and $R_n(k,m)$ represent here the PSD of the clean speech $s(n)$ and the corrupting noise $n(n)$ respectively. Equation (5) is thus composed of the speech power distortion $R_{E_s} = R_s(k,m)[G-1]^2$ and the residual noise power distortion $R_{E_n} = R_n(k,m)[G-\zeta]^2$. The optimal $G(k,m)$ can be obtained by computing the minimum of the solid red parabola (R_E) of Fig. 2, while $G(k,m)$ for the just noticeable distortion case is derived considering the crossing point between the green curve with square (R_T) and the blue curve with triangle (R_{E_n}) of Fig. 2.

As a complete masking of both distortions $R_E < R_T$ is practically not possible, only the masking of the residual noise power distortions is taken into account. By masking the residual noise power distortions, the speech power distortions are also assumed to be minimized (Gustafsson, 1998). So equating noise power distortion R_{E_n} to masking curve of the rough clean speech R_T, the spectral weighting rule is derived as

$$G(k,m) = \min\left(\sqrt{\frac{R_T(k,m)}{R_n(k,m)}} + \zeta(\lambda,m), 1\right), \tag{6}$$

where λ represents herein a frequency bandwidth and $\zeta(\lambda,m)$ is chosen based on the corresponding subband segmental SNR:

$$SNR_i(\lambda,m) = 10\log_{10}\left(\frac{\sum_{k=k_s}^{k_e} R_T(k,m)}{\sum_{k=k_s}^{k_e} R_{\tilde{n}}(k,m)}\right), \tag{7}$$

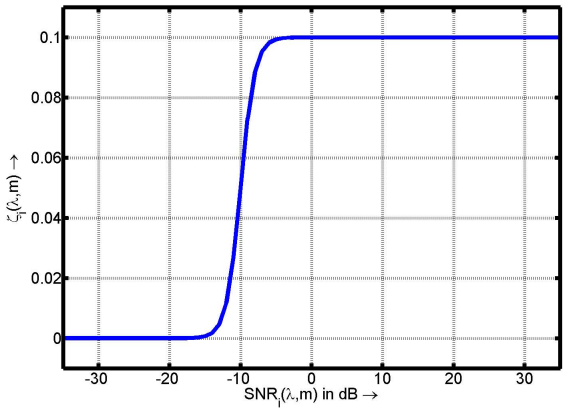

Fig. 3. $\zeta(\lambda, m)$ vs. segmental SNR in band i with bandwidth λ.

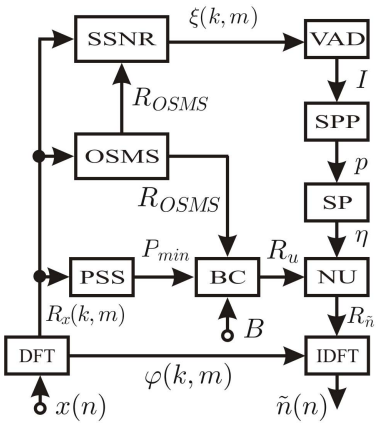

Fig. 4. RARS approach. Power Spectrum Smoothing (PSS), Bias Correction (BC), Noise Update (NU), Smoothing Parameter (SP), Speech Presence Probability (SPP), Voice Activity Detector (VAD), Smoothed SNR (SSNR).

as shown in Fig. 3, that is computed from a shifted sigmoid function. k_s and k_e represent in Eq. (7) the starting and ending bin of the i^{th} band.

To reduce the spectral outliers in specific frequency bands, the gain function is manipulated based on the energy of the coefficients within critical bands as follows:

$$\tilde{G}(k,m) = \frac{G(k,m)}{k_e - k_s + 1} \cdot \sum_{k=k_s}^{k_e} |G(k,m)|^2. \quad (8)$$

The results with this technique are shown in Figs. 5 and 6, where it is clear that the corrupting noise has been properly controlled and sibilant sounds preserved.

2.1 Noise estimation

As the computation of Eq. (6) requires the knowledge of the corrupting noise power, only an estimated noise power can be used for the single channel case. Therefore, the Rapid Adaptation and Recursive Smoothing (RARS approach) (Nsabimana, 2009), that is depicted in Fig. 4, is applied here.

In the RARS approach, first the noise power is estimated first using Optimal Smoothing and Minimum Statistics (OSMS) approach (Martin, 2001) with a very short window. This yields an overestimation of the estimated noise power. Based on the smoothed posteriori SNR from the OSMS noise power a VAD index I is derived to compute the speech presence probability P and a smoothing parameter η. This smoothing parameter is finally applied to the unbiased estimated noise power R_u from OSMS approach to account for the overestimation. In order to improve the adaptation time for the estimated noise power, a condition BC is used to track quickly the fast changes in the noise power. Results from (Nsabimana, 2009) reveal that the RARS approach adapts fast and provides accurate mean estimates than OSMS approach (Martin, 2001) especially for very low SNRs.

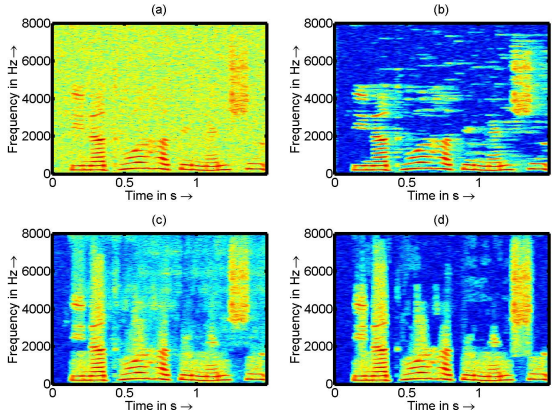

Fig. 5. Results from investigated speech enhancement techniques. Corrupted speech at 10 dB **(a)**, OMLSA aprroach **(b)**, PMSWR approach **(c)** and IPMSWR approach **(d)**.

3 Experimental results

This section presents the performance evaluation of the proposed enhancement technique using the phase of the corrupted speech on one hand (s. Figs. 5 and 6) and the phase of the clean speech on the other hand (s. Fig. 8). To get a fair comparison, tests were carried out for different SNRs using additive white gaussian noise. A window length of 512 samples with a hop size of 25% for analysis and synthesis is applied for all approaches. Figures 5 and 6 present a subjective comparison in terms of spectrogram. These results show that the IPMSWR approach preserves sibilants (s-like sounds) even for very low SNRs (5–10 dB).

Figure 7 presents again the results obtained during listening test with headphones (Nsabimana, 2009). The fifteen subjects recruited for this test were 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

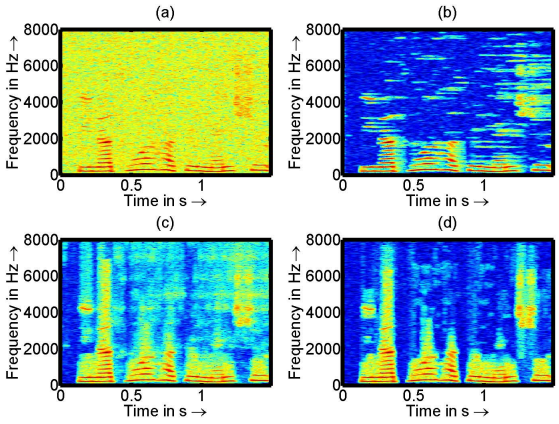

Fig. 6. Results from investigated speech enhancement techniques. Corrupted speech at 5 dB **(a)**, OMLSA aprroach **(b)**, PMSWR approach **(c)** and IPMSWR approach **(d)**.

simulated algorithms are then compared to the reference signal grade. The Mean Opinion Score (ITU-T P.862) represents the grades of the three enhancement techniques for three different kinds of noise. Figure 7 reveal that the IPMSWR approach was graded best.

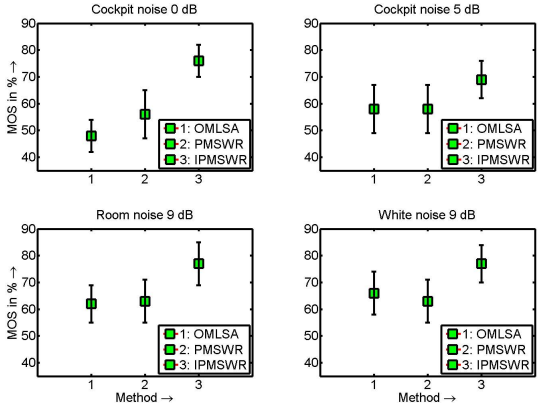

Fig. 7. Results from listening test using headphones. Bars denote 95% confidence interval.

3.1 Usefulness of phase information

The importance of the phase information in speech enhancement is currently being investigated (Shannon, 2006; Shi, 2006; Aarabi, 2006; shi, 2007). To motivate further steps of improvements in the IPMSWR approach, the role of the phase is here emphasized using clean speech degraded with artificial additive white gaussian noise at different SNRs from 0 to 35 dB (s. Fig. 8).

Figure 8 depicts the segmental SNR improvement with IPMSWR approach using the phase of the disturbed speech for the resynthesis on one hand and the phase of the clean speech for the resynthesis on the other hand. The black curve with square represents here the segmental SNR of the

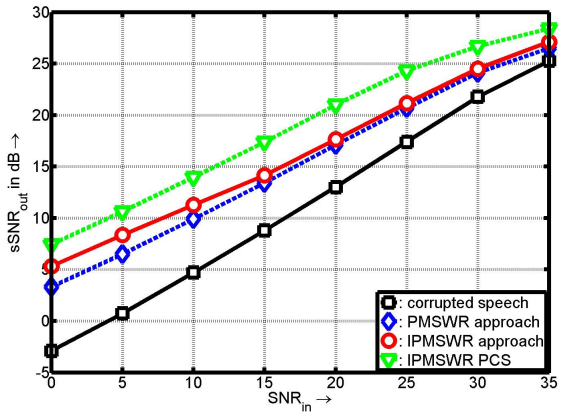

Fig. 8. Segmental SNR improvement with PMSWR and IPMSWR approaches. PCS means Phase of Clean Speech.

disturbed speech, which stands for the reference segmental SNR. The red curve with circle, which depicts the results with the IPMSWR using the phase of the disturbed speech, clearly reveals an overall segmental SNR gain of $\cong 5$ dB. The dashed blue curve with diamond, which depicts the results with the PMSWR using the phase of the disturbed speech, remains close to the results of the IPMSWR approach only for SNRs higher than 15 dB as expected. The dashed green curve with triangle, which depicts the results with the IPMSWR using the phase of the clean speech, reveals instead an overall segmental SNR gain of $\cong 8$ dB. This clearly outlines the usefulness of the phase information.

4 Conclusions

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 within critical bands and a better estimate of the rough clean speech. As additional information on the corrupting noise is not available, a technique to estimate the corrupting noise power has been presented. Simulations results at different SNRs reveal that the proposed technique performs best and preserves sibilant sounds even at very low SNRs. To motivate further steps of improvements in the IPMSWR approach, the importance of the phase information has been emphasized here during experimental results. The obtained results show that an increase of SNR gain is achieved when the phase of the clean speech is used.

Future works should thus address the estimation of the phase to increase the speech intelligibility. To avoid abrupt jumps, the gain coefficients should also be properly controlled. The Parameter optimization remains necessary.

References

Doblinger, G.:Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands, in: Eurospeech 1995, vol. 2, pp. 1513–1516, September 1995.

Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on Speech and Audio Processing, 9(5), pp. 504–512, July 2001.

Cohen, I. and Berdugo, B.:Noise estimation by minima controlled recursive averaging for robust speech enhancement, IEEE Signal Processing Letters, vol. 9(1), pp. 12–15, Januar 2002.

Cohen, I.:Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, IEEE Transactions on Speech and Audio Processing, vol. 11(5), pp. 466–475, September 2003.

Rangachari, S., Loizou, P. C., and Hu, Y.: A noise estimation algorithm with rapid adaptation for highly nonstationary environments, IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings of ICASSP'04, vol. 1, pp. I–305–308, 17–21 May 2004.

Stouten, V., Van Hamme, H. and Wambacq, P.: Application of Minimum Statistics and Minima Controlled Recursive Averaging Methods to Estimate a Cepstral Noise Model for Robust ASR, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006, 1, pp. 765–768, May 2006.

Nsabimana, F. X.,Subbaraman, V., and Zölzer, U.: Noise power estimation using rapid adaptation and recursive smoothing principles, in: Proc. of the International Conference on Signal Processing and Multimedia Applications (SIGMAP 2009), pp. 13–18, Milan, Italy, 7–10 July 2009.

Ephraim, Y. and Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 33(2), pp. 443–445, April 1985.

Tsoukalas, D., Paraskevas, M. and Mourjopoulos, J.: Speech enhancement using psychoacoustic criteria, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93, vol. 2, pp. 359–362, April 1993.

Gustafsson, S., Jax, P., and Vary, P.: 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, 12–15 May 1998.

Virag, N.: Single channel speech enhancement based on masking properties of the human auditory system, IEEE Transactions on Speech and Audio Processing, vol. 7(2), pp. 126–137, March 1999.

Cohen, I.: Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator, Signal Processing Letters, IEEE, vol. 9(4), pp. 113–116, April 2002.

Yi, H. and Loizou, P.C.: Incorporating a psychoacoustical model in frequency domain speech enhancement, Signal Processing Letters, IEEE, vol. 11(2), pp. 270–273, ISSN 1558-2361, Februar 2004 April 2002.

Hu, R. and Anderson, D.V.: Improved perceptually inspired speech enhancement using a psychoacoustic model, IEEE Journal on Signals, Systems and Computers, Record of the Thirty-Eighth Asilomar Conference, vol. 1, pp. 440–444, November 2004 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-93, vol. 2, pp. 359-362, April 1993.

Nsabimana, F. X., Subbaraman, V., and Zölzer, U.: A single channel speech enhancement technique using psychoacoustic principles, in: Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009), pp. 170–174, Glasgow, Scotland, 24–28 August 2009.

Johnston, J. D.: Transform coding of audio signals using perceptual noise criteria, IEEE Journal on Selected Areas in Communications, vol. 6(2), pp. 314–323, February 1988.

Zwicker, E. and Fastl, H.: Psychoacoustics facts and models, Springer Verlag, Berlin, 1990.

Zölzer, U.: Digitale Audiosignalverabeitung, B. G. Teubner, Wiesbaden 2005, ISBN 3-519-26180-4.

ITU-T 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.

Shannon, B. J. and Paliwal, K. K.: Role of phase estimation in speech enhancement, in: Interspeech 2006 – ICSLP, pp. 1423–1426, Pittsburgh, Pennsylvania, 17–21 September 2006.

Shi, G., Shanechi, M., Aarabi, P., On the importance of phase in human speech recognition, IEEE Transactions on Audio, Speech, and Language Processing, vol. 14(5), pp. 1867–1874, September 2006.

Aarabi, P., Shi, G., Shanechi, M. M and Rabi, S.A: Phase-Based Speech Processing, Published by, World Scientific Publishing Co. Pte Ltd, ISBN 981-256-612-0.

Shi, G., Aarabi, P., and Jiang, H.: Phase-Based Dual-Microphone Speech Enhancement Using A Prior Speech Model, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15(1), pp. 109–118, Januar 2007.