Implementation Effective One-Channel Noise Reduction System

Jiří Tihelka, Pavel Sovka

Czech Technical University in Prague,
ČVUT FEL K331, Technická 2, Praha 6, Czech Republic
E-mail: tihelka@feld.cvut.cz, sovka@feld.cvut.cz

Abstract

This contribution addresses the problem of additive noise reduction using one-channel noise suppression system. A new implementation effective method is suggested and evaluated. The method consists of two independent parts. The noise estimation part is based on the noise matched filter producing an estimation of background noise without the need of a voice activity detector. The noise reduction part uses short-time spectral attenuation technique. The main idea reducing computational costs lies in the use of reduced number of frequency bands for computation of attenuation factors. Except of reduced number of operations this approach decreases fluctuations of estimated spectral gains, and therefore the speech distortion is low. Thus the suggested system eliminates the need of any enhanced speech postprocessing. A new effective approach is used for the inverse frequency transformation. In spite of the simplicity of the suggested method its performance is comparable with other existing one-channel noise reduction methods.

1. Introduction and problem definition

The problem of additive noise reduction has been intensively solved for last three decades. Many noise reduction systems have been developed and their main parts modified to achieve better performance and greater simplicity. This contribution is devoted to the further system simplification under constraint of preserving the system performance.

Limitations used on the system development can be summarized as follows:

(G1) one-channel method without any voice activity detector;

(G2) very low computational costs;

(G3) low speech distortion and sufficient noise suppression.

To solve these conflicting requirements the following rules in descending order of priorities were used:

(R1) the system simplicity and low computational costs;

(R2) low number of tuning parameters;

(R3) low speech distortion is more important than the level of noise reduction;

(R4) character of residual noise has greater importance than the level of noise suppression.

2. Suggested solution

The whole system consists of two parts: noise estimation and noise reduction subsystems.

1) The noise estimation subsystem with respect to (G1) may use one of the following approaches: Martin's [4], Doblinger's [5], and noise matched filter (NMF) [1]. With regards to rules (R1) and (R2), the NMF approach has been chosen despite of the fact that the quality of noise estimation is not so good as in Martin's or Doblinger's methods [3]. In order to achieve further reduction of the computational costs compared with [1], the power spectra rather than magnitude spectra are applied in the suggested approach. Unfortunately, this approach results in an underestimation of the noise spectrum and leads to the lower noise suppression. On the other hand, the residual noise level and the speech distortion are lower than for [1]. In comparison with [1], the NMF approach has been slightly modified, therefore it requires two independent tuning parameters. This does not represent any complication for tuning of the system because the value of the added parameter can be changed in a very broad interval without noticeable influence on the system performance (for details see Section 3.2).

2) The noise reduction subsystem is based on a short-time spectral attenuation (STSA) technique attenuating the complex spectrum of the input signal according to spectral gains [11]. Three methods estimating spectral gains and satisfying the rules (R1) and (R2) have been compared: the power spectral subtraction (PSS), the so-called Wiener filtering (WF) [2] and the spectral subtraction (SS) [12] (for details see Section 3.3). No tuning parameter is required for this part of the system.

Main parts of both subsystems, the noise estimation and the spectral gains estimation, use critical band analysis [9] in the frequency domain because this approach satisfies all the rules (R1–R4). Reasonable compromise between the system performance and the system complexity represents two frequency bands per one critical band. The inverse frequency transformation is realized by the linear interpolation. This approach leads to the reduced number of operations and excludes the need of any further speech post-processing used for residual noise reduction in many one-channel noise reduction methods.

3. The system description

Let us suppose that the input signal $x(n)$ consists of the clean speech $s(n)$ and an environmental noise $v(n)$ uncorrelated with $s(n)$, i.e. $x(n) = s(n) + v(n)$ and $\mathrm{E}[s(n)v(n)] = 0$. Let $X(k,l)$, $S(k,l)$, and $V(k,l)$ denote the discrete short-time spectra of the signals $x(n)$, $s(n)$, and $v(n)$, respectively. Index k denotes the frequency index, $k = 0, \ldots, N-1$, parameter N is the size of the discrete Fourier transform (DFT), and l stands for the index of frame.

3.1. General framework

The proposed noise suppression system is based on common used STSA technique (see Fig. 1). The principle consists in the attenuation of the short-time spectrum $X(k,l)$ of the input signal $x(n)$ according to a time varying spectral gain $G(k,l)$ to produce the spectrum $\hat{S}(k,l)$ of enhanced speech. In the proposed system, the short-time analysis and synthesis is performed by use of the discrete short-time Fourier transform (DSTFT).

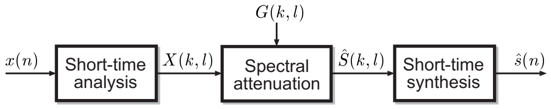

Figure 1: *Speech enhancement by STSA technique.*

The time varying spectral gain $G(k,l)$ applied to each spectral component $X(k,l)$ is determined by a noise suppression rule. This suppression rule usually relies on the "relative signal power" [11] $Q(k,l) = X_p(k,l)/\hat{P}_v(k,l)$, where $X_p(k,l) = |X(k,l)|^2$ is the power spectrum of the input signal and $\hat{P}_v(k,l)$ is an estimation of the power spectral density (PSD) of the noise.

Common used noise suppression methods evaluate the noise suppression rule for each spectral component. We have used the approach adopted from [8], which uses the fact that it is not desirable to estimate the relative signal power $Q(k,l)$ for each spectral component because in this way the estimation can be very sensitive to errors in the estimation of spectral values. The block diagram of the proposed estimation of spectral gains $G(k,l)$ is shown in Fig. 2. Firstly, the uniformly distributed spectral components $X_p(k,l)$ are mapped to the critical band scale (unit Bark) [9]. The mapping is carried out by the formula

$$X_p(b,l) = \frac{\sum_{k=k_L(b)}^{k_H(b)} X_p(k,l)}{k_H(b) - k_L(b) + 1}\,, \quad b = 0,\ldots,B-1\,, \quad (1)$$

where $k_L(b)$ and $k_H(b)$ are the lower and the upper limits of bth band, B is the total number of bands. $X_p(b,l)$ represents the power in the bth band. The result of the frequency transformation $X_p(k,l)$ to $X_p(b,l)$ for one signal frame is illustrated in Fig. 3 (the top graph).

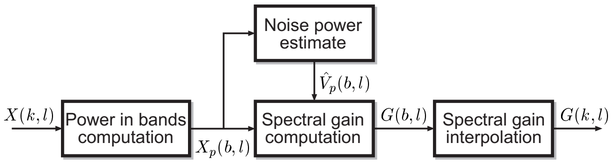
Noise power estimate
$\hat{V}_p(b,l)$
$X(k,l)$
Power in bands computation
$X_p(b,l)$
Spectral gain computation
$G(b,l)$
Spectral gain interpolation
$G(k,l)$

Figure 2: *Block diagram of the spectral gain estimation.*

The total number bands B depends on properties of the noise. There should be more bands for a narrow-band noise. We use two bands per one critical band (Bark) [9]. That means that for audible range of frequencies there are at most 52 bands. Secondly, the gain $G(b,l)$ of the bth band is computed by a common used noise suppression rule (see Section 3.3) from the power $X_p(b,l)$ of the input signal and an estimation $\hat{V}_p(b,l)$ of the power of the noise in the bth band. Finally, the spectral gains $G(k,l)$ are computed from the gains $G(b,l)$ using the linear interpolation (see Fig. 3, the bottom graph). The linear interpolation has been used to avoid a distortion caused by a considerably different attenuation of the adjacent spectral lines.

3.2. Noise estimation subsystem

The subsystem for noise estimation is a modified version of [1]. The model of the noise matched filter estimation is given in Fig. 4. The input signal $x(n)$ is composed from the clean speech $s(n)$ and an environmental noise $v(n)$. The output of the NMF $\hat{v}[n]$ should be an approximation of the noise $v(n)$. When noncausal NMF filter is used, the minimization of the cost function $J = \mathrm{E}\left[e^2(n)\right] = \mathrm{E}\left[(v(n) - \hat{v}(n))^2\right]$ leads to the equation which is similar to Wiener-Hopf equation

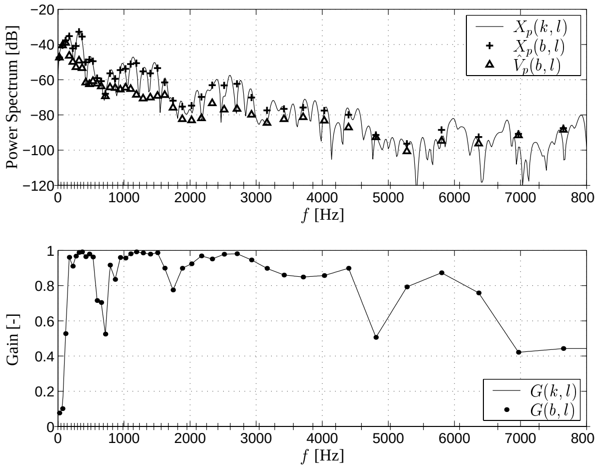

Figure 3: *Normalized power spectra $X_p(k,l)$, $X_p(b,l)$, $\hat{V}_p(b,l)$, and computed spectral gains $G(k,l)$ and $G(b,l)$.*

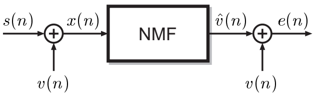

Figure 4: *Noise matched filter estimation.*

$$\sum_{i=-\infty}^{\infty} h(i) R_{xx}(m-i) = R_{vx}(m+1),\ -\infty < m < \infty,\ (2)$$

where the $h(i)$ is the least mean squares estimation of the impulse response of the matched filter, $R_{xx}(m-i)$ denotes the autocorrelation coefficients of $x(n)$, and $R_{vx}(m+1)$ is the cross-correlation between $x(n)$ and $v(n)$. When $s(n)$ is uncorrelated with $v(n)$, $R_{vx}(m) = R_{vv}(m)$ and (2) can be reformulated to give the frequency response of the noncausal NMF

$$H(e^{j\theta}) = \frac{\Gamma_{vv}(e^{j\theta})}{\Gamma_{xx}(e^{j\theta})}, \quad (3)$$

where $\Gamma_{vv}(e^{j\theta})$ and $\Gamma_{xx}(e^{j\theta})$ denote the theoretical PSDs of signals $v(n)$, $x(n)$, respectively. The denominator of (3) can be expressed in the form

$$\Gamma_{xx}(e^{j\theta}) = \Gamma_{ss}(e^{j\theta}) + \Gamma_{vv}(e^{j\theta}), \quad (4)$$

where $\Gamma_{ss}(e^{j\theta})$ is the PSD of the unknown clean speech $s(n)$.

The implementation of (3) requires to replace the time invariant NMF by a time variant approximation of NMF [1] as follows:

1) The theoretical PSDs $\Gamma_{vv}(e^{j\theta})$ and $\Gamma_{xx}(e^{j\theta})$ have to be replaced by their time dependent estimations $\hat{P}_v(b,l)$ and $\hat{P}_x(b,l)$, respectively.

2) The evaluation of $\hat{P}_x(b,l)$ requires a spectral smoothing that causes an unacceptable speech distortion. Therefore $\hat{P}_x(b,l)$ is evaluated using (4) which leads to the form $\hat{P}_x(b,l) = \hat{P}_s(b,l) + \hat{P}_v(b,l)$.

3) Directly unmeasurable $\hat{P}_s(b,l)$ is estimated according to $\hat{P}_s(b,l) = \max\{X_p(b,l) - \hat{P}_v(b,l), 0\}$, where $\hat{P}_v(b,l)$ is the exponentially averaged NMF output $\hat{V}_p(b,l)$. The $\hat{P}_s(b,l)$ is then further smoothed by exponential averaging.
4) The final frequency response of the NMF is given by

$$\hat{H}(b,l) = \frac{\hat{P}_v(b,l-1)}{\hat{P}_s(b,l-1) + \hat{P}_v(b,l-1)} \,. \tag{5}$$

The time delay between $\hat{H}(b,l)$ and PSDs $\hat{P}_v(b,l-1)$ and $\hat{P}_s(b,l-1)$ is required for the NMF to be causal.
The whole noise estimation subsystem is shown in Fig. 5.

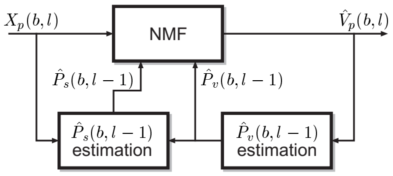

Figure 5: *Noise matched filter estimation.*

3.3. Noise suppression subsystem

The noise reduction subsystem exploits very simple suppression rules like PSS [2], WF [2] and SS [12]. Spectral gains $G(b,l)$ of these methods are summarized in Table 1 as functions of the relative signal level $Q(b,l) = X_p(b,l)/\hat{V}_p(b,l)$, where $\hat{V}_p(b,l)$ is used as an approximation of $\hat{P}_v(b,l)$. Values of $Q(b,l)$ lower than 1 are artificially set to 1 by use of the expression $\max\{Q(b,l), 1\}$, and therefore spectral gains $G(b,l)$ for $Q(b,l) < 0$ is set to 0. Fig. 6 displays dependences of $G(b,l)$ on the relative signal power. It can be seen that SS attenuates also the spectral components with high relative signal level $Q(b,l)$, and therefore it is not suitable for signals with high signal-to-noise ratios (SNRs). It is more convenient to use WF or PSS (see Section 5).

Table 1: *Used spectral gains.*

PSS	$G(b,l) = \sqrt{1 - \frac{1}{\max\{Q(b,l),1\}}}$
WF	$G(b,l) = 1 - \frac{1}{\max\{Q(b,l),1\}}$
SS	$G(b,l) = 1 - \sqrt{\frac{1}{\max\{Q(b,l),1\}}}$

As we can see in Fig. 6, all introduced spectral gains $G(b,l)$ (as well as most of others) are monotonically increasing with the relative signal level $Q(b,l)$, and therefore the lower level of the noise spectrum leads to the lower noise suppression.

4. Notes on system implementation

4.1. Signal analysis and synthesis

The parameters of DSTFT were set as follows: The input frame length M corresponds to 30 ms. The size of the DFT $N = 2^{\lceil \log_2(2M) \rceil}$. The analysis window Blackman-Harris (-92 dB) (BHW) [13], the overlap 75 % (it should be emphasized that exactly for BHW the overlap should be 87.5%). The synthesis window rectangular. Experiments revealed that the BHW choice reduces the level of crackling (in comparison with frequently used Hann or Hamming windows) in the synthesized speech.

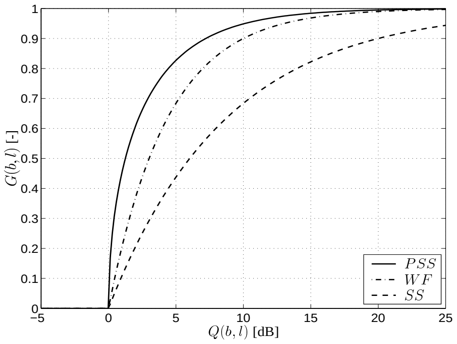

Figure 6: *Gain $G(b,l)$ versus relative signal level $Q(b,l)$. PSS - power spectral subtraction, WF – Wiener filtering, SS – amplitude spectral subtraction.*

As mentioned before, the main reduction of computational costs is achieved by the evaluation of the suppression rule in B (instead of $K = N/2+1$) bands. The implemented system uses $B = 43$ and $K = 513$ for the sampling rate 16000 Hz. The computational cost of the linear interpolation used in the inverse transformation is very low because the indices of centre frequencies of bands are in ascending order.

4.2. Noise estimation subsystem

$\hat{H}(b,l)$ should not be zero for the proper behaviour of the MNF. This is ensured by $\hat{P}_v = \max\{\hat{P}_v, \varepsilon\}$, where ε is a properly chosen small constant. The time constant used in $\hat{P}_v(b,l)$ estimation varies in the interval 50–150 ms. This parameter has the strong influence on the performance of the system. The possible range of values for the time constant used in $\hat{P}_s(b,l)$ estimation is 8–15 ms (11 ms was used). This value has a limited influence on the performance of the system. Smoothing of $X_p(b,l)$ prevents the spectral gains $G(b,l)$ to be zero, thus it acts similarly as the noise floor.

4.3. Noise suppression subsystem

The noise suppression part was implemented according to the rules given in Table 1.

5. Experiments and results

5.1. Evaluation criteria

The proposed speech enhancement technique was evaluated using segmental signal-to-noise ratio (SNR) enhancement and informal listening tests. The local SNR value of frame l with length M can be expressed as

$$\mathrm{SNR}(l) = 10 \log \frac{\sum_{n=Ml}^{Ml+M-1} s^2(n)}{\sum_{n=Ml}^{Ml+M-1} \big(y(n) - s(n)\big)^2} \,, \tag{6}$$

where $y(n)$ stands either for $x(n)$ (for SNR of the input signal $\mathrm{SNR}_{\mathrm{IN}}$) or for $\hat{s}(n)$ (for SNR of the enhanced signal $\mathrm{SNR}_{\mathrm{OUT}}$). Since frames with SNRs above 35 dB and bellow -20 dB do not reflect the perceptual contributions of the signal [14], the local SNRs were limited to be in range $\langle -20, 35 \rangle$ dB. The segmental SNR is formed by averaging the local SNRs as follows

$$\mathrm{SSNR} = \frac{1}{L} \sum_{l=0}^{L-1} \mathrm{SNR}(l) \,, \tag{7}$$

where L is the total number of frames. The segmental SNR enhancement (SSNRE) is evaluated as the diference of input and output SSNRs

$$\mathrm{SSNRE} = \mathrm{SSNR}_{\mathrm{IN}} - \mathrm{SSNR}_{\mathrm{OUT}} \ . \tag{8}$$

The values of SSNRE for three input signal-to-noise ratios $\mathrm{SSNR}_{\mathrm{IN}}$ for the proposed method are given in Table 2. NMF in the abbreviation of method means that the noise matched filter is used for the noise power estimation, Martin means Martin's method [4]. CB means a method where suppression rule is evaluated in reduced number of frequency bands. PSS, WF, SS, and MMSElog mean power spectral subtraction [2], Wiener filter [2], spectral subtraction [12], and minimum mean-square error short-time log-spectral amplitude estimator [7], respectively. The NMF method of noise estimation combined with three noise suppression rules (SS, WF, PSS) is compared with NMF & MMSElog and Martin & MMSElog

Table 2: *Values of SSNRE [dB] for different input SSNR [dB].*

Method	$\mathrm{SSNR}_{\mathrm{IN}}$ [dB]		
abbreviation	-6	0	6
NMF-PSS	3.64	3.40	2.41
NMF-PSS CB	3.88	3.52	2.48
NMF-SS	5.74	3.20	-0.46
NMF-SS CB	5.62	2.97	-0.60
NMF-WF	5.39	4.50	2.48
NMF-WF CB	5.49	4.37	2.35
NMF-MMSElog	6.24	4.08	0.92
NMF-MMSElog CB	5.69	3.66	0.73
Martin-MMSElog	7.09	6.30	4.80
Martin-MMSElog CB	5.20	5.06	4.59

Table 2 shows that the methods with CB and without CB have approximately the same SSNRE level, even though CB methods have lower computational demands. The NMF has lower SSNRE in comparison with Martin's noise estimation for high input SSNR. It is the consequence of an inaccurate estimation of the noise PSD by the NMF. As mentioned before, the SS rule has high attenuation of spectral components for high input SSNR, and therefore it is not suitable in this case.

Informal listening tests confirmed that the quality of enhanced speech of the proposed method is similar to (or slightly worse than) the quality of the method composed of Martin's method with MMSElog. PSS provides lower noise suppression and speech distrotion than WF. SS produces high speech distortion for high input SSNRs.

6. Conclusions

The implementation effective noise suppression system which does not need a voice active detector has been presented. This system belongs to the class of short-time spectral attenuation techniques. The main reduction of the computational cost has been achieved by the reduction of the number of frequency bands where the suppression rule is computed. To obtain the spectral gains of all spectral components the linear interpolation has been suggested. The noise estimation subsystem is based on the noise matched filter which implementation is very effective. The noise suppression has been achieved by a simple suppression rule like the power spectral subtraction or the Wiener filter. Experiments confirmed that this method, in spite of its computational cost, is comparable with MMSElog method combined with Martin's noise estimation approach.

7. Acknowledgements

This project was supported by the grant GACR 102/96/k087 "The Theory and Application of Speech Communication in Czech" and Transdisciplinary Research in Biomedical Engineering J04/98:210000012.

8. References

[1] P. Sovka, P. Pollak, and J. Kybic, "Extended spectral subtraction," in *EUSIPCO'96*, Trieste, Italy, Sep. 1996.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[3] J. Meyer, K. U. Simmer, and K. D. Kammeyer, "Comparison of one- and two-channel noise-estimation techniques," in *IWAENC'97*, Imperial College London, UK, 1997, pp. 17–20.

[4] R. Martin, "Spectral subtraction based on minimum statistics," in *EUSIPCO'94*, Edinburgh, Scotland, U.K., Sep. 1994, pp. 1182–1185.

[5] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *EuroSpeech'95*, Madrid, Spain, Sep. 1995, pp. 1513–1516.

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

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[8] E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech, Audio Processing*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[9] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer-Verlag, Berlin, Germany, 1999.

[10] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.

[11] O. Cappé, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 1, pp. 84–93, Jan. 1995.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[13] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[14] J.H.L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *ICSLP'98*, Sydney, Australia, Dec. 1998, pp. 2819–2822.