A Modified Minimum Statistics Algorithm for Reducing Time Varying Harmonic Noise

Thomas Esch, Matthias Rüngeler, Florian Heese, and Peter Vary

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, 52056 Aachen, Germany
E-Mail: {esch|ruengeler|heese|vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

In this paper, a single-channel speech enhancement system is presented that is capable of reducing (time-varying) harmonic noise. For this, the well-known Minimum Statistics approach is modified in order to track and suppress the harmonic noise based on the harmonics' fundamental frequency which is assumed to be known a priori. The performance of the proposed noise suppression system is shown to be consistently better than conventional approaches.

1 Introduction

Speech quality and intelligibility may significantly deteriorate in the presence of background noise, e.g., engine noise or street noise. The problem of enhancing speech that is degraded by additive noise has been widely studied in the past and is still an active field of research. Speech enhancement has many applications in voice communications, speech recognition and hearing aids. A generic block diagram of state-of-the-art noise reduction systems is shown in Fig. 1.

A crucial component of a practical speech enhancement system is the estimation of the noise power spectrum. For this purpose, many approaches can be found in the literature, e.g., estimating the noise power during speech pauses using a voice activity detector (VAD) [1], the Minimum Statistics approach [2] or the minimum mean square error (MMSE) based noise power spectral density (PSD) tracking algorithm [3]. Although the latter approach shows slightly better noise tracking characteristics compared to Minimum Statistics, all methods have some problems in tracking a sudden rise in noise energy leading to underestimation of the noise power.

In this paper, we are investigating speech enhancement in harmonic noise environments where strong spectral components of the noise signal are present at multiples of a fundamental frequency f_0. A possible application area can be found in the interior of motor vehicles, e.g., communication via a hands-free device inside a car where the engine is one of the main noise sources. An example can be seen in Fig. 2(a) showing the noisy spectrogram of a speech signal recorded inside a car. The original Minimum Statistics approach as well as the MMSE based noise PSD tracking algorithm for noise estimation fail in cases of sudden changes. A new modified Minimum Statistics approach is presented in the following that is capable of suppressing all harmonic oscillations quite effectively. Therefore, it is assumed that the instantaneous fundamental frequency f_0 for each frame is available to the noise reduction system, e.g., received from the vehicle's onboard computer or estimated in a separate procedure which is not the focus of this paper.

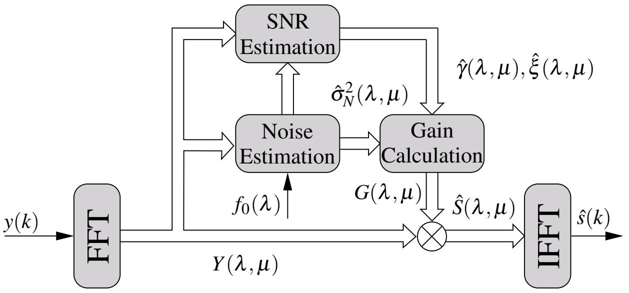


Figure 1: Proposed noise reduction system.

The remainder of this paper is organized as follows: In Sec. 2, a brief overview of the proposed noise reduction system is given including a summary of the original Minimum Statistics approach. Section 3 comprises the new noise estimation technique, experimental results are shown in Sec. 4 and conclusions are drawn in Sec. 5.

2 System Overview

According to Fig. 1, the speech signal $s(k)$ is assumed to be degraded by an additive uncorrelated noise $n(k)$ to produce the noisy speech signal

$$y(k) = s(k) + n(k), \tag{1}$$

where k is the discrete time index.
For the transformation into the frequency domain the noisy input signal $y(k)$ is first segmented into overlapping frames of length L_F. After windowing (e.g., applying a Hann-window), these frames are transformed via Fast Fourier Transform (FFT) of length M_F. The short-time spectrum of the noisy input signal is given by:

$$Y(\lambda,\mu) = S(\lambda,\mu) + N(\lambda,\mu), \tag{2}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ represent the spectral coefficients of speech and noise at frequency bin μ and frame λ. The most crucial element of the system is the estimation of the power spectral density of the noise signal which is required by most statistical estimators. A widely used approach is the well-known Minimum Statistics algorithm [2] which is briefly summarized in the next section. As this approach does not perform well in changing harmonic noise environments a modified Minimum Statistics algorithm is presented in Sec. 3 which exploits the knowledge of the instantaneous fundamental frequency f_0. Both noise estimation techniques are evaluated and compared together with the MMSE noise PSD tracking algorithm [3] in Sec. 4.

ISBN 978-3-8007-3300-2 © VDE VERLAG GMBH · Berlin · Offenbach

Based on the estimate $\hat{\sigma}_N^2$ of the noise PSD, two SNR parameters are estimated, namely the *a posteriori SNR* $\gamma(\lambda,\mu)$ and the *a priori SNR* $\xi(\lambda,\mu)$:

$$\gamma(\lambda,\mu) = \frac{|Y(\lambda,\mu)|^2}{\sigma_N^2(\lambda,\mu)} \quad \text{and} \quad \xi(\lambda,\mu) = \frac{\mathscr{E}\{|S(\lambda,\mu)|^2\}}{\sigma_N^2(\lambda,\mu)}. \tag{3}$$

The a priori SNR can be estimated using the decision-directed approach [4]. The actual spectral weighting is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ with the weighting gains $G(\lambda,\mu)$:

$$\hat{S}(\lambda,\mu) = G(\lambda,\mu)\cdot Y(\lambda,\mu). \tag{4}$$

The weighting gains depend on the noise reduction algorithm and are usually a function of the noise PSD estimate $\hat{\sigma}_N^2(\lambda,\mu)$ and the SNR estimates $\hat{\gamma}(\lambda,\mu)$ and $\hat{\xi}(\lambda,\mu)$, as stated before. The spectral weighting results in an estimate $\hat{S}(\lambda,\mu)$ of the clean speech coefficient $S(\lambda,\mu)$. In order to obtain the enhanced signal in the time domain, an Inverse Fast Fourier Transform (IFFT) and overlap-add is applied.

2.1 Minimum Statistics Review

The Minimum Statistics [2] approach is an efficient procedure to estimate the PSD without requiring a voice activity detector. The basic principles of this method are briefly outlined in the following (see [2] for more detailed information). Minimum Statistics relies on two basic assumptions:

- speech and noise are statistically independent and
- the power of the noisy signal often decays to the power level of the noise signal (e.g., in speech pauses).

Based on these assumptions it is possible to track the minimum of the noisy PSD. As this minimum is always smaller or equal to the mean noise power, a bias correction is necessary.

At first, the noisy periodogram $|Y(\lambda,\mu)|^2$ is recursively smoothed over time. The smoothed signal power $\hat{\sigma}_Y^2(\lambda,\mu)$ is thereby given by

$$\hat{\sigma}_Y^2(\lambda,\mu) = \alpha_{\mathrm{MS}}\cdot\hat{\sigma}_Y^2(\lambda-1,\mu) + (1-\alpha_{\mathrm{MS}})\cdot|Y(\lambda,\mu)|^2, \tag{5}$$

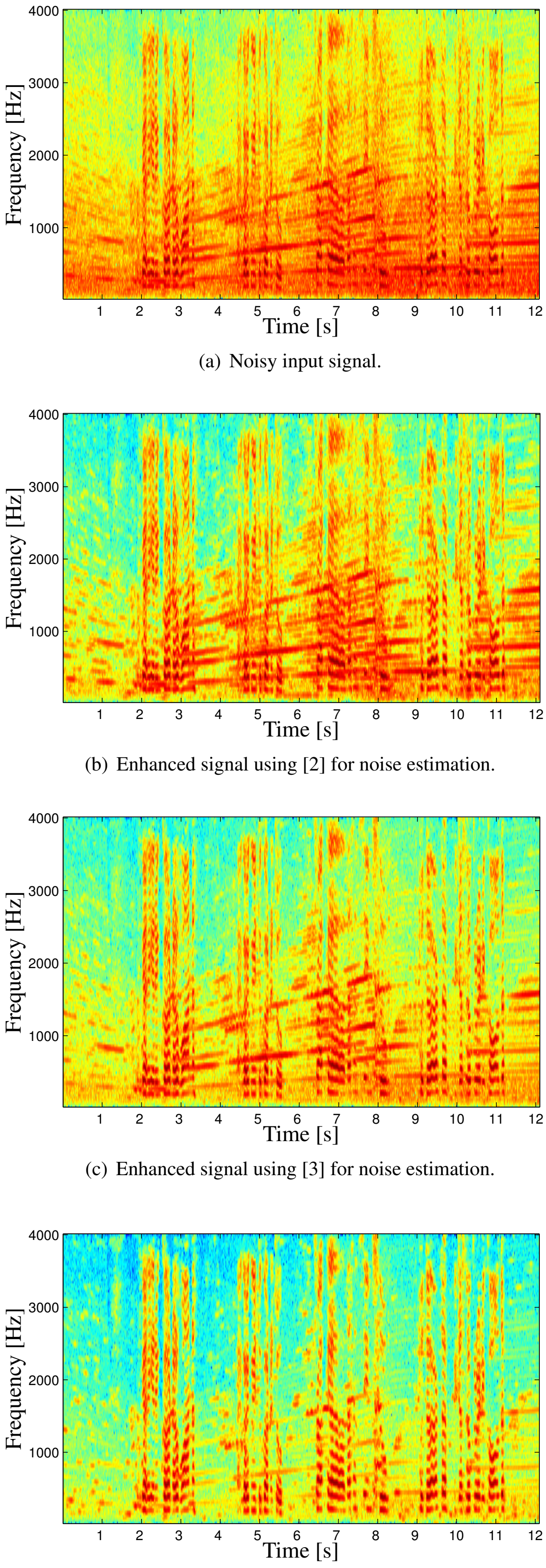
where $\alpha_{\mathrm{MS}} \in [0,1]$ is the adaptive smoothing factor. For each frequency bin μ, the signal power $\hat{\sigma}_Y^2$ of the previous D frames is buffered in the matrix

$$\hat{\Sigma}^2(\lambda) = \begin{pmatrix} \hat{\sigma}_Y^2(\lambda-D+1,0) & \dots & \hat{\sigma}_Y^2(\lambda,0) \\ \hat{\sigma}_Y^2(\lambda-D+1,1) & \dots & \hat{\sigma}_Y^2(\lambda,1) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda-D+1,\mu) & \dots & \hat{\sigma}_Y^2(\lambda,\mu) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_Y^2(\lambda-D+1,M_{\mathrm{F}}-1) & \dots & \hat{\sigma}_Y^2(\lambda,M_{\mathrm{F}}-1) \end{pmatrix}. \tag{6}$$

Afterwards, the minimum is tracked for each row *separately* for each frequency bin according to $\hat{\Sigma}^2(\lambda)$:

$$\hat{\sigma}_{Y,\min}^2(\lambda,\mu) = \min\left(\hat{\Sigma}^2(\lambda,\mu)\right), \tag{7}$$

where $\hat{\Sigma}^2(\lambda,\mu) = \left(\hat{\sigma}_Y^2(\lambda-D+1,\mu) \quad \dots \quad \hat{\sigma}_Y^2(\lambda,\mu)\right)$ represents the μ-th row of $\hat{\Sigma}^2(\lambda)$. The duration of the



(a) Noisy input signal.

(b) Enhanced signal using [2] for noise estimation.

(c) Enhanced signal using [3] for noise estimation.

(d) Enhanced signal using new approach for noise estimation.

Figure 2: Spectrograms of (a) noisy input signal, (b) enhanced signal using original Minimum Statistics, (c) enhanced signal using MMSE based noise PSD tracking algorithm and (d) enhanced signal using the new approach.

time window D for the minimum search should be approximately 1.5 seconds and states a trade-off between fast noise tracking and speech distortions.
Finally the minimum value is multiplied by a bias correction factor $B(\lambda,\mu)$, which is mainly dependent on the variance of the noisy signal. The final noise PSD estimation is given by:

$$\hat{\sigma}_N^2(\lambda,\mu) = B(\lambda,\mu)\hat{\sigma}_{Y,\min}^2(\lambda,\mu). \qquad (8)$$

3 Noise Estimation Based on a Modified Minimum Statistics Approach

The original Minimum Statistics approach performs well in stationary and slowly changing noise conditions as the minimum at each frequency bin within the search time window provides a good estimate of the actual noise power. However, when it comes to a sudden rise in the noise power, Minimum Statistics is not able to track this rise in the specific frequency bin due to the large window length D.

In this paper, we are investigating speech signals disturbed by harmonic noise characterized by strong spectral components at multiples of the fundamental frequency f_0. As the fundamental frequency might change over time very fast (e.g., when the engine accelerates or when a gear change occurs), the original Minimum Statistics approach fails in tracking the spectral harmonics. This can be seen, e.g., in Fig. 2(b), where the noise power is underestimated and the harmonic noise remains almost unchanged compared to the noisy input signal in Fig. 2(a).

In this new approach, the original Minimum Statistics procedure is modified. Instead of tracking over time the spectral minimum which is associated with the (stationary) noise at *one* specific frequency bin (see Fig. 3, method (a)), we adaptively 'look back' inclined according to the evolution of the harmonics in the time-frequency domain (see Fig. 3, method (b)). Following one specific harmonic oscillation over time, the noise is no longer fluctuating but relatively stationary and we can apply the Minimum Statistics concept. In order to achieve this tilted 'viewing direction', we need to modify the entries of the matrix $\hat{\Sigma}^2(\lambda)$ according to the fundamental frequency f_0 that has to be known a priori for each frame. The harmonic oscillation of the current frame λ at frequency $f_0(\lambda)$ can be found in the frame $\lambda - D + 1 + i$ at frequency $f_0(\lambda - D + 1 + i)$ with $0 \le i \le D-1$. In order to estimate the noise power at frame λ, we therefore compress/expand the i-th column of the matrix $\hat{\Sigma}^2(\lambda)$ according to the ratio:

$$r(\lambda) = \frac{f_0(\lambda)}{f_0(\lambda - D + 1 + i)} \quad \text{with} \quad 0 \le i \le D-1. \qquad (9)$$

After transformation, the i-th column of the modified matrix $\hat{\Sigma}^2_{\text{mod}}(\lambda)$ comprises the noisy signal power at the new positions $\mu' = r(\lambda - D + 1 + i) \cdot \mu$. For this curve fitting problem, linear interpolation is used. In the case of compression ($r < 1$), the elements of the i-th column in the interval $M_F \cdot r(\lambda) \le \mu' \le M_F$ are missing and replaced by $\hat{\sigma}_Y^2(\lambda - D + 1 + i, M_F)$. The adaptive smoothing factors α_{MS} (see Eq. 5) are warped similar to the columns of $\hat{\Sigma}^2_{\text{mod}}(\lambda)$ with the same ratio $r(\lambda)$.

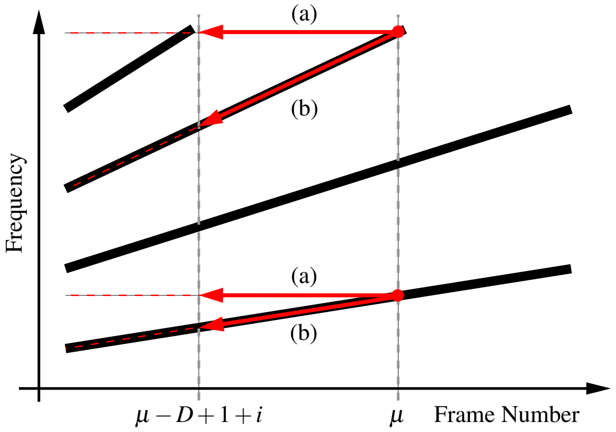


Figure 3: 'Direction of view' of (a) original Minimum Statistics approach and (b) new method. The black lines show the time-frequency evolution of the harmonic oscillations.

Based on the modified matrix $\hat{\Sigma}^2_{\text{mod}}(\lambda)$, the minimum is again tracked for each row similar as in the original approach:

$$\hat{\sigma}_{Y,\min}^2(\lambda,\mu) = \min\left(\hat{\Sigma}^2_{\text{mod}}(\lambda,\mu)\right), \qquad (10)$$

where $\hat{\Sigma}^2_{\text{mod}}(\lambda,\mu)$ represents the μ-th row of $\hat{\Sigma}^2_{\text{mod}}(\lambda)$. From the Minimum Statistics' point of view, the harmonics in the time-frequency domain of $\hat{\Sigma}^2_{\text{mod}}(\lambda)$ seem to be warped and the resulting noise appears stationary over time. Finally, the bias is calculated according to Sec. 2.1 and Eq. 8 is applied.

4 Results

The proposed noise estimation technique for harmonic noise environments has been compared with the original Minimum Statistics approach [2] and the MMSE based noise PSD tracking algorithm [3]. Therefore, the speech enhancement system depicted in Fig. 1 was used incorporating f_0 which was provided by the vehicle's onboard computer. Please note that f_0 is only required for the modified Minimum Statistics approach and not for the other two noise estimation techniques. The a priori SNR was estimated according to the decision-directed approach [4] and the well-known Wiener filter [5] is used to calculate the spectral weighting gains.

For the evaluation, five different (real) noise recordings were each added to three male and two female speech sequences (each with a length of 8 s taken from the NTT speech database) at an input SNR varying between -10 dB and 15 dB with an increment of 5 dB. The parameters that are used in the simulations are listed in Tab. 1.

In the simulation, the speech and noise signal can be filtered separately with weighting gains adapted for the noisy signal. Hence, the output signal can additionally be stated as $\hat{s}(k) = \tilde{s}(k) + \tilde{n}(k)$, where $\tilde{s}(k)$ is merely the filtered speech signal and $\tilde{n}(k)$ the filtered noise signal. Based on these quantities, the segmental speech and noise attenuation (SA and NA) and the segmental SNR (SegSNR) were calculated (e.g., Chap. 4 in [6]).

The averaged results are depicted in Figs. 4 and 5. Fig. 4 shows the difference between noise and speech at-

Parameter	Setting
Sampling frequency	8 kHz
Frame length L_F	160 (20 ms)
FFT length M_F	256 (including zero-padding)
Frame overlap	50% (Hann window)

Table 1: System settings.

tenuation and Fig. 5 the segmental SNR plotted over the input SNR. In both figures, higher scores indicate a better performance of the respective approach. The objective measurements show that the noise estimation performed with the new modified Minimum Statistics approach consistently improves the results of the other two conventional noise estimation techniques. If the speech attenuation is kept constant, a gain of up to 3.5 dB/2 dB in noise attenuation is possible, the segmental speech SNR can be increase by up to 3 dB/1.5 dB depending on which technique is used as reference: the original Minimum Statistics approach or the MMSE based noise PSD tracking algorithm. Comparing all three noise estimation techniques, the new approach achieves the best performance whereas the original Minimum Statistics approach shows the worst noise tracking characteristics.

In addition to these quantitative results, spectrograms of the processed signals are shown in Fig. 2(b) for the original Minimum Statistics approach, in Fig. 2(c) for the MMSE based noise PSD tracking algorithm and in Fig. 2(d) for the new approach. Compared to the noisy input signal in 2(a), the approaches in Figs. 2(b) and 2(c) fail in this noisy environment (the spectral harmonics remain almost unchanged). The new approach performs significantly better and is able to suppress most of the engine noise. As can be seen, the harmonics are effectively removed leading to a more comfortable listening condition. Even if the harmonics' fundamental frequency coincides with the human's pitch frequency, there is usually no huge performance loss as the window length D is large enough to cope with these periods of speech activity.

5 Conclusions

This paper deals with single channel speech enhancement in (time-varying) harmonic noisy environments. As conventional noise estimation techniques usually have problems in tracking a sudden rise in noise energy leading to under-estimation of the noise power the well-known Minimum Statistics approach was modified in order to overcome this problem in this specific environment. The new approach performs frequency warping according to the harmonics' fundamental frequency and is able to track and suppress the harmonic noise quite effectively. For this, it is assumed that the fundamental frequency is known a priori. Instrumental measurements show a consistent improvement in terms of noise/speech attenuation and segmental SNR compared to the original Minimum Statistics approach and another recently published noise estimation technique. The objective measurements were confirmed by informal listening tests.

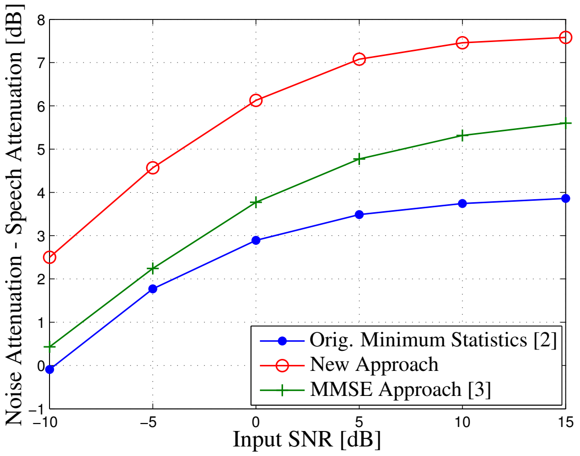


Figure 4: Difference between noise attenuation and speech attenuation plotted over input SNR.

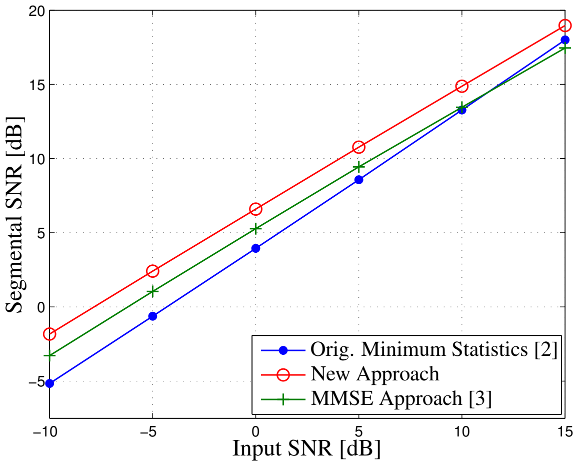


Figure 5: Segmental SNR plotted over input SNR.

References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," vol. 6, no. 1, pp. 1–3, 1999.

[2] 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, pp. 501–512, 2001.

[3] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE Based Noise PSD Tracking with Low Complexity," in *Proc. of ICASSP*, Dallas, USA, 2010.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[6] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer, Berlin, 2005.