Combined Reduction of Time Varying Harmonic and Stationary Noise Using Frequency Warping

Thomas Esch, Matthias Rüngeler, Florian Heese, and Peter Vary
Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{ esch | ruengeler | heese | vary }@ind.rwth-aachen.de

***Abstract*—Speech enhancement under non-stationary environments is still a challenging problem. This contribution presents a noise reduction system that is capable of tracking and suppressing both time varying harmonic noise and stationary noise. In a first stage, the harmonic noise power is estimated and attenuated using a modified Minimum Statistics approach that performs frequency warping according to the harmonic's fundamental frequency. A conventional noise estimation technique is applied in a second stage in order to reduce the random components of the noise spectrum. The performance of the proposed noise suppression system is shown to be consistently better than conventional approaches.**

I. Introduction

When a speech communication device is used in environments with high levels of ambient noise, the noise picked up by the microphone significantly impairs the quality and the intelligibility of the transmitted speech signal. In order to get a reliable separation from the noise signal (e.g., engine noise, street noise), noise reduction algorithms have become part of digital speech communication systems. They are used for example in mobile communication systems, in hearing aids and in hands-free devices.

A crucial component of a practical speech enhancement system is the tracking of the instantaneous noise power spectrum. For this purpose, many approaches can be found in the literature, e.g., the application of a voice activity detector (VAD) [1], the Minimum Statistics approach [2] or the 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 under-estimation of the noise power.

In this paper, we are investigating speech enhancement in noise environments consisting of (time varying) harmonic noise and random noise 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 intercom systems for motorcycles or in the interior of motor vehicles, e.g., communication via a hands-free device inside a car where engine, wind and tyres contribute to the main noise sources. The proposed system consists of two stages. In the first stage, harmonic noise components are suppressed using a modified Minimum Statistics approach [4]. 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. In the second stage, remaining residual stationary background noise is reduced.

The remainder of this paper is organized as follows: In Sec. II, a brief overview of the proposed noise reduction system is given. Section III comprises the different noise estimation techniques in detail, experimental results are shown in Sec. IV and conclusions are drawn in Sec. V.

II. System Overview

A simplified block diagram of the proposed system is shown in Fig. 1. The speech signal $s(k)$ is assumed to be degraded by an additive harmonic noise signal $n_{\mathrm{h}}(k)$ and a stationary random noise signal $n_{\mathrm{s}}(k)$ to produce the noisy speech signal

$$y(k) = s(k) + n_{\mathrm{h}}(k) + n_{\mathrm{s}}(k), \tag{1}$$

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 spectrum of the noisy input signal is given by:

$$Y(\lambda,\mu) = S(\lambda,\mu) + N_{\mathrm{h}}(\lambda,\mu) + N_{\mathrm{s}}(\lambda,\mu), \tag{2}$$

where $S(\lambda,\mu)$, $N_{\mathrm{h}}(\lambda,\mu)$ and $N_{\mathrm{s}}(\lambda,\mu)$ represent the spectral coefficients of speech and noise signals at frequency bin μ and frame λ.

The concatenation of the two noise suppression stages relies on different noise PSD estimators. In the first stage, the harmonic noise power $\sigma^2_{\mathrm{h},N}$ is estimated using a modified Minimum Statistics approach [4] which exploits the knowledge of the instantaneous fundamental frequency f_0 (see Sec. III-A for more details). Based on the estimate $\hat{\sigma}^2_{\mathrm{h},N}$, two SNR parameters are estimated, namely the *a posteriori SNR* $\gamma_{\mathrm{h}}(\lambda,\mu)$ and the *a priori SNR* $\xi_{\mathrm{h}}(\lambda,\mu)$:

$$\gamma_{\mathrm{h}}(\lambda,\mu) = \frac{|Y(\lambda,\mu)|^2}{\sigma^2_{\mathrm{h},N}(\lambda,\mu)} \quad \text{and} \quad \xi_{\mathrm{h}}(\lambda,\mu) = \frac{\mathcal{E}\{|S(\lambda,\mu)|^2\}}{\sigma^2_{\mathrm{h},N}(\lambda,\mu)}. \tag{3}$$

The a priori SNR can be determined using the decision-directed approach [5]. The actual spectral weighting of stage I is performed by multiplying the noisy spectrum $Y(\lambda,\mu)$ with weighting gains $G_{\mathrm{h}}(\lambda,\mu)$:

$$\hat{S}_{\mathrm{h}}(\lambda,\mu) = G_{\mathrm{h}}(\lambda,\mu) \cdot Y(\lambda,\mu). \tag{4}$$

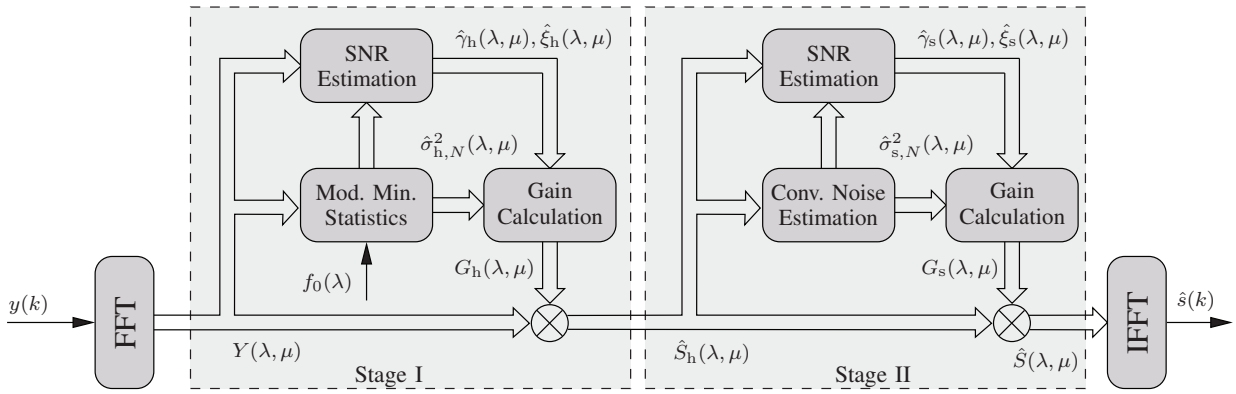


Fig. 1: Proposed noise reduction system.

The weighting gains depend on the noise reduction algorithm and are usually a function of the noise PSD estimate and the SNR estimates.

In the second stage, the residual stationary background noise is suppressed. While for the noise estimation, a conventional technique can be used, e.g., [2] or [3], the same techniques as before are applied for the subsequent a posteriori SNR γ_s and a priori SNR ξ_s estimation as well as for the gain calculation G_s. The second spectral weighting results in an estimate

$$\hat{S}(\lambda,\mu) = G_\mathrm{h}(\lambda,\mu) \cdot G_\mathrm{s}(\lambda,\mu) \cdot Y(\lambda,\mu) \tag{5}$$

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 are applied.

III. Combined Noise Suppression System

The noise estimation concepts of both stages are described in the following.

A. Harmonic Noise Estimation Using a Modified Minimum Statistics Approach

The original Minimum Statistics approach [2] 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 smoothed noisy PSD within an appropriate time window *separately* for each frequency bin. The duration of the 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. As this minimum $\hat{\sigma}^2_{Y,\min}(\lambda,\mu)$ is always smaller or equal to the mean noise power, a bias correction is necessary. The final noise PSD estimation is given by:

$$\hat{\sigma}^2_N(\lambda,\mu) = B(\lambda,\mu) \cdot \hat{\sigma}^2_{Y,\min}(\lambda,\mu), \tag{6}$$

where $B(\lambda,\mu)$ is the bias correction factor [2] that is mainly dependent on the variance of the noisy input signal.

In the first stage of the proposed noise reduction system, a modified Minimum Statistics approach is applied in order to estimate the harmonic noise components [4]. Instead of tracking over time the spectral minimum which is originally associated with the stationary noise at *one* specific frequency bin (see Fig. 2, method (a)), we adaptively 'look back' inclined according to the evolution of the harmonics in the time-frequency domain (see Fig. 2, method (b)). Following one specific harmonic oscillation over time, the noise is no longer fluctuating but relatively stationary and we can apply the original Minimum Statistics concept. In order to achieve the tilted 'viewing direction', we modify the entries of the buffer out of which the minimum is tracked according to the fundamental frequency f_0. The harmonic oscillation of the current frame λ_0 at frequency $f_0(\lambda_0)$ can be found in the frame $\lambda_0 - D + 1 + i$ at frequency $f_0(\lambda_0 - D + 1 + i)$ with $0 \leq i \leq D-1$. In order to estimate the noise power at frame λ_0, we therefore compress/expand the $(\lambda_0 - D + 1 + i)$-th frame in the buffer according to the ratio:

$$r(\lambda_0, i) = \frac{f_0(\lambda_0)}{f_0(\lambda_0 - D + 1 + i)} \quad \text{with} \quad 0 \leq i \leq D-1. \tag{7}$$

This frequency warping along the frequency axis causes the respective frame to comprise the noisy signal power at the new positions $\mu' = \frac{\mu}{r(\lambda_0,i)}$. From the Minimum Statistics' point of view, the resulting noise within the warped buffer appears stationary over time and the original Minimum Statistics concept, including minimum tracking and bias correction can be applied. The final noise estimate $\hat{\sigma}^2_{\mathrm{h},N}(\lambda,\mu)$ is required for the SNR estimation and the calculation of the weighting gains $G_\mathrm{h}(\lambda,\mu)$, cf. Sec. II and Fig. 1.

B. Suppression of Stationary Background Noise

As the modified Minimum Statistics algorithm is adapted to the fundamental frequency, the stationary noise components (e.g., wind or tyre noise) are only slightly suppressed by the first stage. The second stage reduces the random parts of the noise signal. As depicted in Fig. 1, conventional noise estimation techniques can be applied for this purpose. In the following evaluation, the original Minimum Statistics

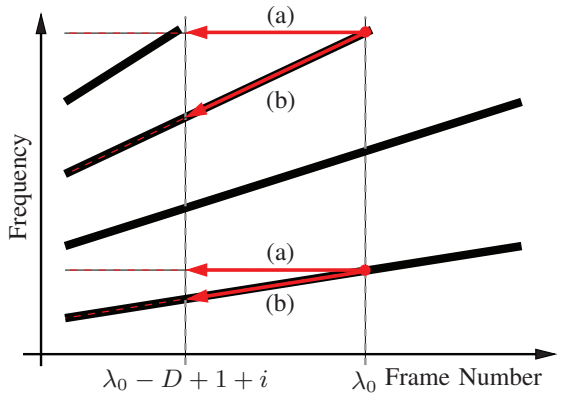


Fig. 2: 'Direction of view' of (a) original Minimum Statistics approach and (b) new method.

approach [2] and the MMSE based noise tracking algorithm [3] are investigated in the second stage of the noise reduction system.

IV. Results

The proposed noise estimation technique for harmonic and random noise environments is compared with the results of the original Minimum Statistics approach [2] and the original 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. The a priori SNR was estimated according to the decision-directed approach [5] and the well-known Wiener filter [6] is used to calculate the spectral weighting gains. Referring to Fig. 1, the following noise estimation techniques are applied in stages I and II:

Method	Stage I	Stage II
A	disabled (G_h=1)	Minimum Statistics [2]
B	disabled (G_h=1)	MMSE based noise PSD tracking [3]
C	modified Minimum Statistics (see Sec. III-A)	disabled (G_s=1)
D	modified Minimum Statistics (see Sec. III-A)	Minimum Statistics [2]
E	modified Minimum Statistics (see Sec. III-A)	MMSE based noise PSD tracking [3]

In Fig. 3, spectrograms of the processed signals are shown. In the upper plot, the spectrogram of the noisy input signal is depicted which is a recording inside a car. In addition to stationary background noise, it can be seen that the engine mainly contributes to the noise signal. The speech signal is highly disturbed by the spectral harmonics. The spectrograms of the processed signals are shown in Figs. 3a-3d for the different approaches A, B, D and E. While the conventional noise estimation (Figs. 3a and 3b) fails in this noise environment (stationary background noise slightly reduced but spectral harmonics remain almost unchanged), the new approaches (Figs. 3c and 3d) perform significantly better and are able to suppress most of the engine and stationary background noise without affecting the speech quality. As can be seen, the harmonics and the stationary background noise are effectively removed leading to a more comfortable listening condition.

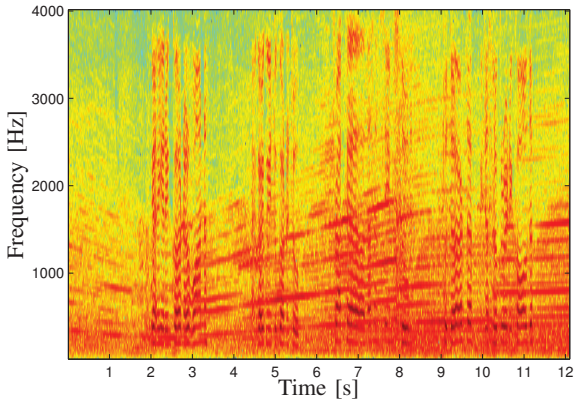
For the objective measurements, five different (real) noise recordings were each added to three male and two female speech sequences (each with a length of 8 s taken from the NTT speech database) at an input SNR varying between -10 dB and 15 dB with an increment of 5 dB. The parameters that are used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

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 [7]).

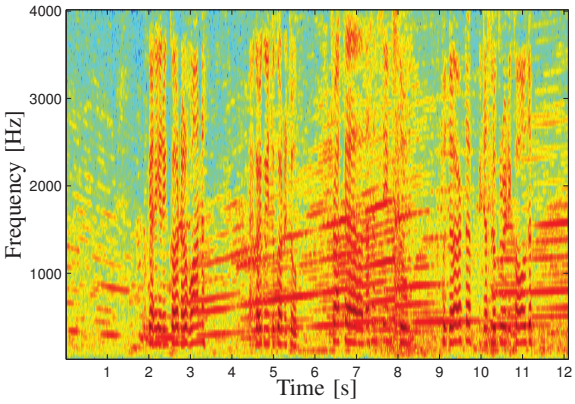
The averaged results are depicted in Figs. 4 and 5. Figure 4 shows the difference between noise and speech attenuation where higher scores indicate a better performance of the respective approach. It can be seen that the proposed 2-stage system (methods D and E) consistently improves the results of the conventional noise estimation techniques (methods A and B) as well as the results of the modified Minimum Statistics approach (method C). Moreover, the combined system consisting of the modified Minimum Statistics and the MMSE based noise PSD tracking algorithm yields the best performance with respect to noise attenuation and speech attenuation and outperforms all other approaches. In Fig. 5, the segmental SNR is plotted over the input SNR. Although the results of methods C, D and E are quite similar in this measurement, the improvements compared to the conventional techniques (methods A and B) can clearly be seen.

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)

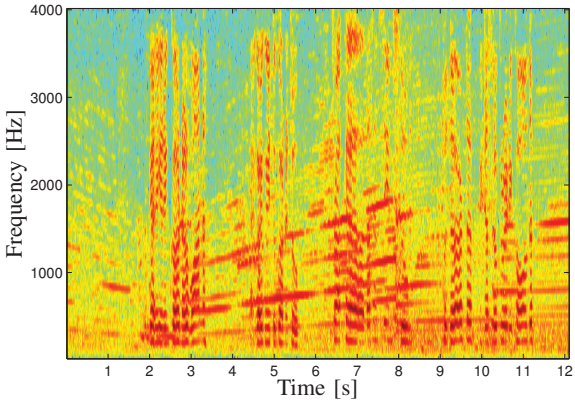
TABLE I: System settings.



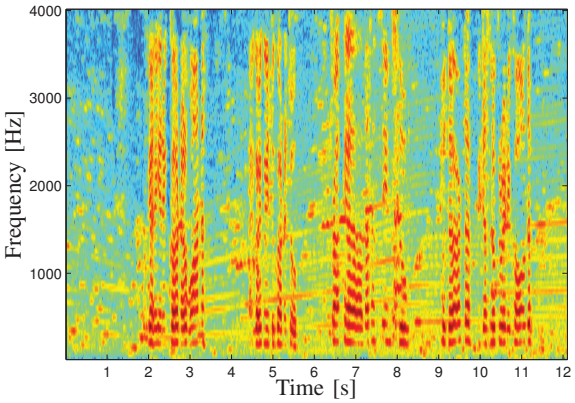
Noisy input signal



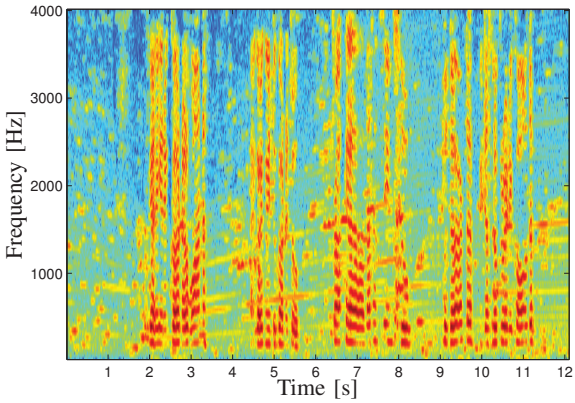
(a) Enhanced signal processed by method A



(b) Enhanced signal processed by method B



(c) Enhanced signal processed by method D



(d) Enhanced signal processed by method E

Fig. 3: Spectrograms of noisy and processed signals: (a) enhanced signal using original Minimum Statistics approach [2], (b) enhanced signal using original MMSE based noise PSD tracking algorithm [3], (c) enhanced signal using new approach by combining modified and original Minimum Statistics approach and (d) enhanced signal using new approach by combining modified Minimum Statistics and MMSE based noise PSD tracking algorithm.

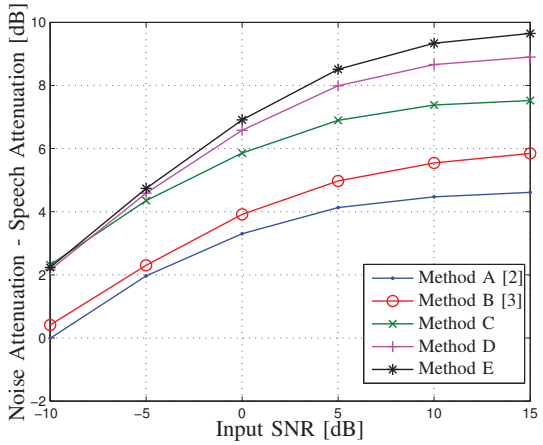


Fig. 4: Difference between noise and speech attenuation plotted over input SNR.

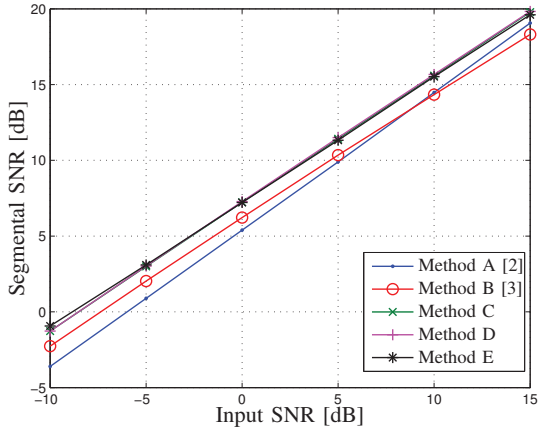


Fig. 5: Segmental SNR plotted over input SNR.

V. Conclusions

This paper deals with single channel speech enhancement in noisy environments consisting of (time varying) harmonic and stationary random noise. Conventional noise estimation techniques usually have problems in this specific environment as the tracking of rapidly varying noise often leads to an underestimation of the noise power. Therefore, the harmonic noise components are reduced in this contribution in a first stage by using a modified Minimum Statistics approach which performs frequency warping according to the harmonic's fundamental frequency in order to track and suppress the harmonic noise quite effectively. The remaining random noise components in the signal are estimated and reduced in a second stage using conventional noise estimation techniques. Instrumental measurements show a consistent improvement in terms of noise/speech attenuation and segmental SNR compared to the original Minimum Statistics approach and an MMSE based noise tracking algorithm. The objective measurements were confirmed by informal listening tests.

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] T. Esch, M. Rüngeler, F. Heese, and P. Vary, "A Modified Minimum Statistics Algorithm for Reducing Time Varying Harmonic Noise," in *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.

[5] 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.

[6] 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.

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